Dev Genes Evol (2016) 226:139–158
DOI 10.1007/s00427-016-0537-4

REVIEW

Measurement error in geometric morphometrics

Carmelo Fruciano [1,2]

Received: 13 September 2015 / Accepted: 28 December 2015 / Published online: 1 April 2016

Abstract Geometric morphometrics—a set of methods for the statistical analysis of shape once saluted as a revolutionary advancement in the analysis of morphology —is now mature and routinely used in ecology and evolution. However, a factor often disregarded in empirical studies is the presence and the extent of measurement error. This is potentially a very serious issue because random measurement error can inflate the amount of variance and, since many statistical analyses are based on the amount of "explained" relative to "residual" variance, can result in loss of statistical power. On the other hand, systematic bias can affect statistical analyses by biasing the results (i.e. variation due to bias is incorporated in the analysis and treated as biologically-meaningful variation). Here, I briefly review common sources of error in geometric morphometrics. I then review the most commonly used methods to measure and account for both random and non-random measurement error, providing a worked example using a real dataset.

Communicated by Nico Posnien and Nikola-Michael Prpic

This article is part of the Special Issue "Size and Shape: Integration of morphometrics, mathematical modelling, developmental and evolutionary biology", Guest Editors: Nico Posnien—Nikola-Michael Prpic.

✉ Carmelo Fruciano
c.fruciano@unict.it

[1] Department of Biological, Geological and Environmental Sciences, University of Catania, via Androne 81, 95124 Catania, Italy

[2] Present address: School of Earth, Environmental and Biological Sciences, Queensland University of Technology, Brisbane, Queensland 4000, Australia

Keywords Geometric morphometrics · Measurement error · Multivariate analysis · Bias

Introduction

Measurement error is ubiquitous in empirical scientific work and morphometrics is no exception. However, its presence and impact has often been overlooked in geometric morphometric studies. Morphometrics—the statistical analysis of morphological traits and their covariation with other variables (Bookstein 1991; Dryden and Mardia 1998)—has undergone a "revolution" (Rohlf and Marcus 1993) when methods retaining the geometrical properties of objects—collectively dubbed "geometric morphometrics"—were developed (for historical perspectives see Bookstein 1993; Adams et al. 2004; Adams et al. 2013). These methods have gained, over the course of the last 20 years, extreme popularity and can nowadays be considered standard tools in biological research. One of the reasons for their popularity is that they combine statistical rigour and ease of interpretation. Indeed, geometric morphometric analyses preserve the statistical rigour that was typical of traditional morphometric methods—mainly concerned with the analysis of linear distances. At the same time, however, geometric morphometrics allows the interpretation of results in terms of shape changes. This was earlier a prerogative of descriptive approaches to the study of morphology, where one would describe with words the shape of biological structures. Landmark-based geometric morphometrics is the most commonly used geometric morphometric approach and, as the name implies, it is based on sets of homologous points (landmarks). This approach has then be extended to include points along curves and surfaces—called "semilandmarks"—which are not strictly homologous (i.e. they are not type I landmarks sensu Bookstein 1991) but retain positional correspondence (Gunz and Mitteroecker 2013). In a

typical geometric morphometric analysis, data is acquired in the form of x,y(,z) coordinates of landmarks (and semilandmarks), subjected to generalized Procrustes analysis (Rohlf and Slice 1990; Bookstein 1997) and then analyzed with multivariate statistical methods. Most of the statistical analyses performed in geometric morphometric studies are the same analyses used in traditional morphometrics, such as general linear models or principal component analysis (Hotelling 1933). However, an ever-growing array of methods are being developed for (or adapted to) geometric morphometric data, with examples comprising approaches to study the covariation between parts (Rohlf and Corti 2000; Klingenberg 2009; Fruciano et al. 2013; Bookstein 2015), analyses of variation in geographic space (Cardini et al. 2007; Cardini and Elton 2009; Fruciano et al. 2011a, b), phylogenetic comparative analyses (Rüber and Adams 2001; Rohlf 2002; Sidlauskas 2008; Klingenberg and Gidaszewski 2010; Adams and Felice 2014) and quantitative genetic analyses (Klingenberg et al. 2001; Posnien et al. 2012; Franchini et al. 2014; Maga et al. 2015). Geometric morphometric methods are now routinely used across the tree of life and represent standard tools in biological research.

Despite being largely used in all sorts of biological studies, a factor that is often disregarded in empirical studies is the presence and the extent of random measurement error and bias. Here, I will use the term "random measurement error" referring to any random variation in a sample that has no biological origin (i.e. it is artifactual). I will, instead, use the term "bias" to indicate non-random error (or systematic, i.e. error that is systematically distributed with respect to the true value). An increase in random measurement error results in an increase in variance (as the total variance will comprise the variance due to true biological variation plus the variance due to measurement error) but no variation in mean. On the other hand, in the case of systematic error the estimated quantity is systematically different from the true quantity (i.e. it has a different mean). The presence of random and/or systematic error in geometric morphometric datasets can have profound consequences on the results of these analyses. Random measurement error can inflate the amount of variance and, since many statistical analyses are based on the amount of "explained" relative to "residual" variance, can result in loss of statistical power (Yezerinac et al. 1992; Arnqvist and Mårtensson 1998). In other words, in presence of high measurement error, the level of "noise" can obscure the "biological signal". For instance, when comparing two groups and testing the null hypothesis of no difference in means (whether with parametric or non-parametric approaches), one might end up not rejecting the null hypothesis because of the high amount of variation within each group due to measurement error (Fig. 1). The effect of measurement error can, however, be more subtle, especially when there is a non-negligible proportion of systematic error. For instance, when combining data from two different operators, a simple difference in the amount of measurement error introduced by each operator could result in apparent patterns of variation in disparity/morphospace occupation and, at the same time, obscure true differences in means (Fig. 1). Obviously, different operators could introduce not only different levels of measurement error but also differences in its direction (inter-operator difference in means). Such a concern is clearly growing in importance as researchers increasingly share morphometric data and some even "crowdsource" the process of data acquisition (Chang and Alfaro 2015). Similarly, the use of automated systems for the acquisition of morphometric data (Loy et al. 1996; Bromiley et al. 2014) obviously relies on their accuracy. Measurement error from various sources can also interact with other kinds of error such as the incorrect estimation of the population multivariate mean when using small sample sizes (Cardini and Elton 2007). When feasible, the quantification of measurement error is, therefore, of paramount importance. Arnqvist and Mårtensson (1998) have written the most detailed treatment of the subject to date. Although their study is certainly still important and relevant, new empirical data on the subject has been accumulating over the last 18 years and a review covering such advances seems appropriate. Therefore, here, I briefly review common sources of measurement error in geometric morphometric studies and current methods used to measure and reduce it.

Sources of measurement error

Measurement error is ubiquitous and can be introduced at any phase of a morphometric analysis. Considering that any experimental procedure normally introduces some level of error and that the acquisition of geometric morphometric data is composed of different phases, we can expect some error to be introduced during all these phases. The question is, then, which phases introduce a non-negligible amount of error and what can be done to limit its effects. What has been described as "error due to specimen preparation" (Arnqvist and Mårtensson 1998) is produced by a heterogeneous array of causes. Specimens subjected to morphometric analyses have, in fact, a history that often includes collection, preservation and presentation of the specimen to the device that acquires morphometric data (for instance a camera or a laser scanner). The effect of preservation on geometric morphometric data has been already assessed in a number of studies. The commonly used fixation of fish in formalin (whether or not preceded by freezing) followed by long-term storage in ethanol produces significant differences in body shape relative to fresh specimens (Martinez et al. 2013; Vergara-Solana et al. 2014). Similarly, differences in fish body shape have been described—relative to the shape of fresh specimens—for both frozen specimens and specimens kept in 95 % ethanol (Valentin et al. 2008; Berbel-Filho et al. 2013). In general, these studies show that combining fish preserved in different ways can have important effects on morphometric analyses (for instance, differences due to different preservation methods could be mistaken for biologically-relevant differences). It is

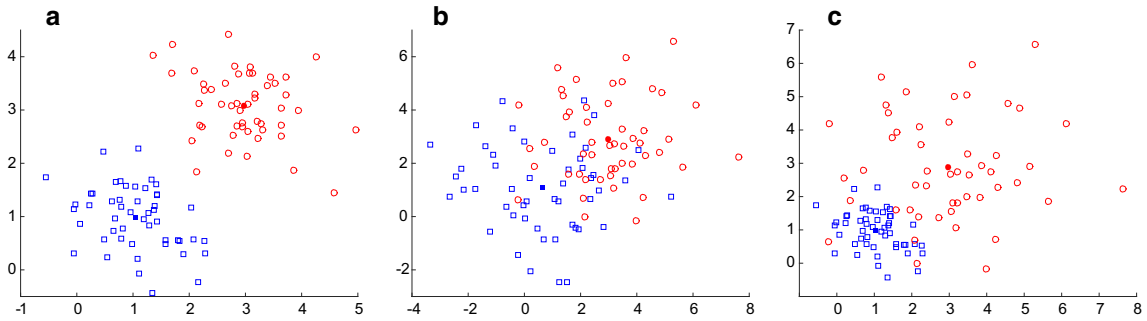

Fig. 1 Schematic representation of some problems induced by measurement error. The graphs have been produced by drawing 50 random samples from two multivariate normal distributions. The mean of each multivariate distribution [(1,1) and (3,3), respectively] from which the samples are drawn is the same across the three panels, only the covariance matrix differs. In all cases, *empty symbols* represent the randomly drawn observations, filled symbols the averages of each sample. **a**. Random samples with limited variance, here used to represent the case of no or very low measurement error and shows a good separation of the two groups. **b**. The variance is increased due to high measurement error and the groups are not as clearly distinct as in panel a. **c**. The variance is increased due to high measurement error just in one of the two groups, this could happen, for instance, when the data from a more reliable operator and a less reliable operator are combined

also entirely possible that the extent of measurement error induced by preservation varies from species to species. These empirical results also suggest that different preservation methods can have different effects on adult fish body shape. However, to test for this—and possibly for the variation of error among species—studies investigating multiple preservation techniques on the same fish (and possibly on multiple species) should be undertaken. Interestingly, however, a study of this kind (Nikolakakis et al. 2014) has been recently performed on larvae of a marine fish species, the European seabass *Dicentrarchus labrax*. The Authors studied the effect of four different preservation approaches, the use of anaesthesia and multiple microscope slide mounting techniques. Apart from the main specific results of this study (i.e. that fixing the larvae in glutaraldehyde after anaesthesia did not induce significant differences in body size and shape), it is important to notice that this study confirms the perhaps intuitive notion that different preservation procedures induce different levels of morphometric variation. On the other hand, another study failed to find differences in adult body shape between fish preserved in ethanol relative to fish preserved in formalin (Hood et al. 2000). The effect of preservatives has also been documented to alter the craniofacial morphology of mouse embryos (Schmidt et al. 2010). Interestingly, an analysis of the effect of formalin on mouse embryonic brains (Weisbecker 2012) has shown temporal patterns: their size first increased and then slowly decreased while their shape changed abruptly in the first 24 h and then remained relatively stable. This shows that the temporal component of preservation should be considered (i.e. mixing specimens which have been preserved for a different time—especially if some of them are "almost fresh" while others are "long preserved"—can induce artifactual variation in a sample). Differences in the way specimens are positioned in front of the data-acquisition device (i.e. photocamera, laser scanner, CT scanner) have been long recognized as another source of measurement error. For instance, Bonneau et al. (2012) found that the shape of dry human pelvic bones after reassembly was different relative to the shape of the complete fresh pelvis. Another analysis on a sample of carabid beetles (Alibert et al. 2001) showed a relatively large amount of measurement error due to presentation associated with the second principal component. Clearly, using a 2D projection for 3D structures has also the potential of introducing non-trivial amounts of measurement error (Cardini 2014). The extent of measurement error induced by acquiring 2D data from 3D objects has been found—using multiple approaches—to be relatively small (Cardini 2014). However, this error was sensibly higher in more spherical structures (crania) compared to relatively flat objects (hemimandibles). Photo cameras also produced more accurate representations of 3D structures relative to flat-bed scanners (Cardini 2014). Incidentally, while one can worry about the error introduced by using 2D projections for 3D objects, measurement error can certainly be introduced also when generating 3D data. In particular, a potential concern comes from the use of multiple pictures to generate 3D data (Fadda et al. 1997; Olsen and Westneat 2015). Fadda et al. (1997), when evaluating the measurement error in repeated measurements of their photo camera-based device to acquire 3D morphometric data found, however, levels of error lower than digitization error. An issue related to the projection of a three-dimensional object in two dimensions (such as when taking a picture) is the effect of parallax, which occurs when the photo camera is placed too close to the object (Mullin and Taylor 2002). Apart from placing the camera farther from the object, this source of error is not expected to affect the analysis as long as all the specimens are affected in the same way (Mullin and Taylor 2002). This, in turn, can be obtained by standardizing the data acquisition procedure. An often overlooked source of artifactual variation, which can be considered as a presentation error, is the one arising by positioning different subjects in different parts of a photo camera field. A study (Riaño et al. 2009) has quantified measurement error due to optical distortion by taking pictures in different positions of the focal field of a camera and

found that the first few principal components were not affected much by measurement error but there were increasing levels of measurement error in higher-order principal components. Unfortunately, due to the fact that measurement error was measured for each principal component separately and was reported only for the first three principal components, it remains hard to evaluate the level of measurement error relative to total variation in multivariate space. The next typical step in geometric morphometric analyses is to acquire data in the form of x,y(,z) (semi)landmark coordinates. This is often termed "digitization", in particular when referring to 2D landmarks from photographs. Given that recording landmark coordinates multiple times is obviously easier than acquiring data from multiple presentations, digitization error is the source of measurement error most commonly quantified in empirical studies. The digitization error reported in empirical studies is typically low relative to the variation among individuals (Debat et al. 2003; Breuker et al. 2006; Kitthawee and Dujardin 2009; Klingenberg et al. 2010; Takahashi et al. 2010; Ayala et al. 2011; Takahashi 2013). Interestingly, digitization error has been found to be low even in a dataset of primate three-dimensional landmarks where part of the sample was re-digitized two years later (Singh et al. 2012). The quantification of measurement error, and in particular of digitization error, is often performed in studies of asymmetry (Breuker et al. 2006; Klingenberg et al. 2010). The rationale for this is that the statistical effect of asymmetry is normally small relative to other potential sources of variation so it is important to check that levels of measurement error are not too high as this might potentially obscure the biological effect, if any (Leamy and Klingenberg 2005; Klingenberg et al. 2010). Intuitively, one might expect digitization error to be lower than presentation error. This could often be the case but studies showing similar levels of measurement errors (Verhaegen et al. 2007; Barrow and Macleod 2008) or even a digitization error higher than presentation error (Verhaegen et al. 2007) suggest that such a natural expectation is not always met. The involvement of multiple operators in any of the various phases detailed above can be an additional source of random and/or systematic measurement error. The logical expectation that experience and learning have an effect on the levels of measurement error has been recently documented (Osis et al. 2015). In fact, Osis et al. (2015) by comparing a novice and an expert operator and by tracking the novice operator over time show that the novice operator learns receiving suggestions from more experienced researchers. It should be noticed, however, that it is not clear how independent from each other and from the expert operator were the suggestions that the two experienced researchers gave to the novice. Clearly, different levels of experience are not the only factor affecting the inter-operator measurement error. Fagertun et al. (2014) show that inter-operator variation was associated to particular landmarks, perhaps reflecting the intuitive notion that certain landmarks are easier to visualize/digitize and, therefore, contain less measurement error (Campomanes-Álvarez et al. 2015). Inter-operator variation has the potential of being a very important source of measurement error and it has been previously found that inter-operator digitization error was higher than intra-operator digitization error (Wilson et al. 2011).

In general, measurement error has a very high number of potential sources and it is each researcher's responsibility to identify the most likely sources of artifactual variation in a certain study and the feasibility of measuring and testing for them. It should also be noticed that most probably there is a reporting bias in the levels of measurement error reported in empirical studies. That is, one can safely assume that datasets where high measurement error is found during the analysis are then improved before being used for the final analysis and publication. For this reason, the typical levels of measurement error in real-life datasets might be higher than the empirical results reported above suggest.

Assessing measurement error

Essentially, two different conceptual approaches have been used to date to assess measurement error: repeated measures and comparisons to a "gold standard". These two approaches clearly serve two different purposes and the latter is often not practical due to the absence of such "gold standard" in the first place. As suggested by a reviewer of this paper, a "gold standard" could be constructed ad hoc by producing unwarped images or 3D models starting from known average shapes and distributions. This would allow the study of measurement errors already knowing the true underlying shapes and quantifying measurement error precisely. It has to be noticed that sometimes a comparison to a "gold standard" is implicit. For instance, the underlying assumption of studies comparing bi-dimensional configurations of landmarks with their three-dimensional counterparts (Cardini 2014) is that the three-dimensional version of the dataset represents a "gold standard". While this assumption is often justified—like in the case of the example—there is a risk of neglecting sources of error in the "gold standard". Repeated measures designs—whether implicitly referring to a gold standard or not—are by far the most commonly used way of quantifying measurement error. In this approach, one generates data multiple times on the same subjects and then measures how much (and possibly how) different measures differ. For instance, when assessing the extent of digitizing error, one repeats the digitization step multiple times on the same pictures (in the case of most 2D morphometric studies) and then quantifies how similar different digitizations are to each other. The typical way of using this approach is to use a statistic to quantify the amount of random measurement error in terms of agreement among different measurements. As an alternative to using a statistic measuring the level of agreement among multiple measurements, plots of the scores along the first few principal components are also often used (O'Higgins and Jones 1998; Viðarsdóttir et al. 2002; Franklin et al. 2007) to

check for clustering of repeated measurements of the same individual. That is, if repeated measurements of the same individual tend cluster together, measurement error is deemed to be low. While these approaches are most commonly used to assess the extent of measurement error, nothing speaks against using a repeated measure design to identify the directions of measurement error (i.e. how multiple measurements differ from each other). In the example above, one could ask him/herself whether over time digitizations of the same specimens tend to produce a different shape. A perhaps more useful case would be when the agreement between multiple operators is of interest. In that case, after multiple operators have digitized the same specimens, one might be interested in knowing not only in how much the different digitizations of different operators agree but also if there is some specific, non-random, pattern of disagreement (i.e. for instance, one operator could consistently digitize more elongated shapes relative to another operator).

Intraclass correlation coefficient

A relatively large number of geometric morphometric studies have used the intraclass correlation coefficient (Fisher 1958)—often termed also "repeatability"—to quantify measurement error in repeated measures designs. The intraclass correlation coefficient is a useful measure of the extent of random measurement error as the variation among repeated measures is compared to the variation among individuals. It can be computed performing a one-way ANOVA on repeated measurements using the individual as categorical variable and then applying the formula (Fleiss 1977; Sokal and Rohlf 1995; Arnqvist and Mårtensson 1998):

$$R = {S^2}_A/\left({S^2}_W + {S^2}_A\right) \quad (1)$$

Where ${S^2}_A$ is the among-individuals variance component and S^2_W is the within-individuals variance component. ${S^2}_A$ and ${S^2}_W$ can, in turn, be computed as:

$${S^2}_A = \left(MS_{\text{among}} - MS_{\text{within}}\right)/n \quad (2)$$

$${S^2}_w = MS_{\text{within}} \quad (3)$$

where n is the number of repeated measurements while MS_{among} and MS_{within} are, respectively, the among-groups and within-groups ANOVA sum of squares.

The repeatability value that would be obtained averaging repeated measurements can be estimated using the formula (Fleiss 1977; Arnqvist and Mårtensson 1998):

$$R_n = nR/[1 + (n-1)R] \quad (4)$$

where R is the value of repeatability computed with Eq. 1 and n is the number of repeated measures.

From Eq. 1, it is, then, clear how the intraclass correlation coefficient reflects the relative amount of measurement error: the among-individuals variance component (${S^2}_A$, that in our biological application we can assume to be reflective of biological variation) is divided by the total variation in the sample (i.e. the among-individuals component summed to the within individuals component, the latter reflecting measurement error). Clearly, the closer this ratio gets to one, the smaller (in a relative sense) is the within individual variation (i.e. the smaller the variation among repeated measures of the same subjects, reflective of measurement error). The intraclass correlation coefficient has been used in empirical geometric morphometric studies to assess the relative measurement error due to digitization (Kitthawee and Dujardin 2009; Henry et al. 2010; Takahashi et al. 2010), presentation (Riaño et al. 2009), variation among operators (Dujardin et al. 2010; Gonzalez et al. 2011). One of the main limitations to the use of the intraclass correlation coefficient is the absence of reference values based on which one can decide whether the amount of measurement error present in a sample is reasonably small. This can be considered a subjective decision and often there is a disagreement among practitioners on what constitutes a "good" repeatability (Costa-Santos et al. 2011). A number of other problems with the intraclass correlation coefficient have been recognized based on statistical grounds. For instance, the value of the intraclass correlation coefficient is dependent on the range of the measuring scale and on the number of repeated measurements (Müller and Büttner 1994). While the formulation given above is the most frequently encountered, other intraclass correlation coefficients have been described (Shrout and Fleiss 1979) and they produce different values when applied on the same data. The most important limitation to the use of the intraclass correlation coefficient in geometric morphometrics is, however, its univariate definition [although multivariate extensions have been proposed; see Ahrens (1976)]. This is obviously a serious limitation in the case of inherently multidimensional data, such as shape data. Despite these limitations, the intraclass correlation coefficient can still be considered a useful exploratory tool to gauge the extent of random measurement error in a sample, especially in pilot studies. It can, for instance, be used to measure the level of repeatability of centroid size, which is a univariate measure. It should be noticed, however, that the repeatability of centroid size in empirical studies has been normally found to be quite high and higher than the repeatability of shape variables across a range of organisms and traits (Arnqvist and Mårtensson 1998; Sinclair and Hoffmann 2003; Langerhans et al. 2007; Simmons and Kotiaho 2007; Simmons and Garcia-Gonzalez 2011; Van Heerwaarden and Sgrò 2011). Multiple studies have also computed the intraclass correlation coefficient on scores along each principal component separately (Arnqvist and Mårtensson 1998; Simmons and Kotiaho 2007). Although this would imply that each principal component can be interpreted separately—which is something problematic and normally not suggested—this approach could still be useful as an

exploratory tool prior to dimension reduction. That is, if the principal components with the lowest repeatability are the ones of highest order one can assume that the "biological signal" is concentrated in the first few principal components while the following principal components are more affected by measurement error. This information can then be used to decide—possibly assisted by other approaches (Anderson 1963; Mitteroecker and Bookstein 2011)—whether dimensionality reduction would be an appropriate choice. Although no specialized morphometric software computes the intraclass correlation coefficient, its computation is extremely easy if one obtains the mean square estimates from a general statistical package and applies the formulas above. Considering that the intraclass correlation coefficient equals the Pearson correlation coefficient in the case of two repeated measurements, a logical extension to multivariate data in the case of two measurements would seem the Escoufier RV coefficient (Escoufier 1973). This coefficient is being increasingly used after being proposed (Klingenberg 2009) for geometric morphometric studies of modularity and implemented in the free and easy-to-use software MorphoJ (Klingenberg 2011). The reason why the Escoufier RV coefficient would seem at first an ideal choice is that this coefficient has been described as a "multivariate analogue" of the correlation coefficient. However, the value of the RV coefficient has been shown to depend on sample size (Smilde et al. 2009; Fruciano et al. 2013) and this would make its interpretation even more problematic than the interpretation of the intraclass correlation coefficient. Perhaps the use of resampling-based approaches to obtain sample size-corrected values of the RV coefficient (Fruciano et al. 2013; El Ghaziri and Qannari 2015) could produce results easier to interpret. Ideally, such approaches should be complemented by tests of the null hypothesis of perfect association between the two datasets. This is possibly an area of future research.

Procrustes ANOVA

Another popular approach for quantifying measurement error is Procrustes ANOVA (Klingenberg and McIntyre 1998; Klingenberg et al. 2002). This method has been implemented in the software MorphoJ and it is popular for analyzing both measurement error (White and Searle 2008; Laffont et al. 2009; Singh et al. 2012; Leamy et al. 2015; Schmieder et al. 2015) and patterns of asymmetry (Breuker et al. 2006; White and Searle 2008; Klingenberg et al. 2010; Leamy et al. 2015). Procrustes ANOVA is based on the fact that the algorithm of generalized Procrustes analysis (Rohlf and Slice 1990)—which is at the core of current geometric morphometric practice—optimally superimposes landmark configurations (i.e. individual observations) by minimizing the sum of squared distances of all objects and the consensus (mean) configuration. For this very reason, then, the sum of squared deviations from the mean configuration of each coordinate can be partitioned in different terms in a two-factor ANOVA. This sum of squares for different effects can then be summed across all the coordinates. Mean squares can then be computed by dividing the total sum of squares for an effect by the relevant degrees of freedom (Klingenberg and McIntyre 1998). The beauty of this approach in the context of the quantification of measurement error is that, by obtaining mean squares for different terms of an ANOVA, one can have an estimate of the relative contribution of each factor to the total variation. That is, one can observe if the mean squares is "large" relative to the biological effect of interest (for instance, a term describing variation in shape among individuals or a term describing levels of asymmetry). This, as mentioned, is particularly relevant in studies of asymmetry where the biologically-relevant effects are often small relative to the variation among individuals and sometimes relative to the variation induced by measurement error. Apart from MorphoJ, this method has been implemented in other software such as the *R* package *geomorph* (Adams and Otárola-Castillo 2013). As for the intraclass correlation coefficient, one potential problem is represented by the difficulty of interpreting the difference in sum of squares when the measurement error mean squares are not much larger than the mean squares for the biological effect of interest. In other words, how much larger should the mean squares for the biological effect of interest should be relative to the measurement error? As in the case of the intraclass correlation coefficient, this is probably a subjective decision. Informally, one could also use the mean squares obtained from the Procrustes ANOVA table to compute repeatability estimates using the formula for the intraclass correlation coefficient (Eqs. 1, 2 and 3 above). Although this would not overcome the problems with the interpretation of the results, it would give a number comprised between zero and one, similar to the intraclass correlation coefficient.

Quantifying the effect of projection in two dimensions

The method recently suggested to measure the effect of using a bi-dimensional projection for three-dimensional structures (Cardini 2014) can be considered in part an extension of the Procrustes ANOVA approach to this specific case. Cardini's approach consists in adding a third zero coordinate to the 2D dataset, performing a common generalized Procrustes analysis using both the 2D and 3D datasets, computing the residuals of shape from the group mean and analyzing the residuals with the same approaches used for other kind of error sources (Procrustes ANOVA, sequential nested agglomerative clustering). In particular, Cardini suggests using Procrustes ANOVA to quantify the amount of shape variation accounted by the

bi-dimensional projection. In terms of the clustering method used, Cardini uses UPGMA—after his procedure that involves subtracting from the 2D and 3D datasets their respective means—as an exploratory tool. The rationale is that if 2D data is an accurate representation of the 3D objects, these would tend to cluster by individual (i.e. 2D and 3D representations of the same individual should be in most or all cases "sister leaves" in the resulting phenogram). One can speculate that other clustering approaches—not creating nested clusters—could also be used. For instance a *k*-means algorithm—which does not construct nested clusters—with *k* (number of clusters) set equal to the number of specimens could be a sensible choice. On the other hand, an algorithm producing nested clusters allows the identification of individuals highly affected by measurement error. For instance, in the case of specimen 14 in Figure 7a of Cardini (2014), not only different representations of the same specimen do not cluster together, but they also occupy very distant portions of the phenogram. One could also compute statistics such as the adjusted Rand index (Hubert and Arabie 1985) to quantify how closely the pattern resulting from the clustering algorithm resembles the expected pattern of all measurements (2D and 3D) of the same individual being in the same cluster. For instance, when computing the adjusted Rand index on the clustering observed in Figure 7a of Cardini (2014), I obtain a value of 0.77, significant ($p < 0.001$; based on 1000 permutations) using a recently proposed procedure (Qannari et al. 2014) to test for the significance of the agreement between expected and observed clustering. These results based on the adjusted Rand index are substantially in agreement with the ones obtained by the Author of the original publication. A more direct approach could also involve simply checking if 2D and 3D versions of the same individual are nearest neighbours in multivariate space. In any case, the part of Cardini's approach using clustering algorithms remains exploratory in nature. As noticed by a reviewer of this paper, Cardini's approach might also produce different results when carried out using partial Procrustes fit (as opposed to the full Procrustes fit implemented in MorphoJ, as used by Cardini). It can be argued that perhaps few researchers have access to devices to acquire 3D data and then decide to use 2D geometric morphometrics. However, I predict that, as methods to acquire 3D data from pictures are developed and perfected (Olsen and Westneat 2015), it will be more and more common for researchers to investigate the costs and benefits of using a 3D or 2D approach on a given biological sample.

Testing for non-random measurement error

Clearly, the procedures presented so far generally quantify the extent of measurement error but do not take into account if measurement error has a specific direction (i.e. bias). For instance, it is entirely possible that different operators impart specific shape features to the data they collect. Similarly, one might also be interested in if and how a certain preservation method affects shape. The typical approach to study this problem is, again, to use a repeated measures design on a set of specimens. That is, in our first example different operators would gather data on the same set of specimens and in our second example the same specimens would be measured fresh and after preservation. The repeated measures design induces non-independence between the two groups of observations and this non-independence should be taken into account. Unfortunately, generally the empirical geometric morphometric studies evaluating measurement error in this kind of design have used hypothesis testing approaches which do not take into account this non-independence. For instance, multiple studies on the effects of preservation techniques (Berbel-Filho et al. 2013; Martinez et al. 2013; Vergara-Solana et al. 2014) have used the permutation procedures based on Hotelling's T^2 or Procrustes distances implemented in MorphoJ (Klingenberg 2011) or in the IMP package (Sheets 2003) to compare the same specimens prior and post-preservation. These procedures are designed to test for difference in mean shape between groups of independent observations. In fact, these approaches use a resampling procedure where the empirical distribution of a test statistic (T^2 or Procrustes distance between groups) is generated under the null hypothesis of no difference between groups by randomly "shuffling" the observations between the two groups. In these procedures, any observation (individual specimen) of one group can, under the null hypothesis, be allocated in the other group and replaced by any other observation of the second group. However, in the case of a repeated measures design both groups contain the same observations/individuals but under two different treatments/conditions. Research in the methods for the statistical analysis of this kind of data—called also "multivariate longitudinal data"—is currently very active (Bandyopadhyay et al. 2011; Verbeke et al. 2014) and outside of the scope of this review. However, the most trivial logical extension of the procedures developed for independent groups of observations is to change the resampling scheme so that treatments are "shuffled" only within subjects. In this case, in fact, the null hypothesis is of no effect due to treatment so, under the null hypothesis, the treatments of the same subjects are exchangeable.

Identifying subjects with high levels of measurement error

When acquiring geometric morphometric data, one might often want to identify the specimens with higher levels of measurement error so that they can be either excluded from the analysis or that better-quality data can be acquired for them prior to the final analysis. Different exploratory procedures have been proposed to identify such specimens. For instance, in an informal online protocol (Adriaens 2007) for testing error in landmark-based geometric morphometrics, Adriaens

suggests inspecting for outliers the plots of Procrustes distances against tangent Procrustes distances available in tpsUtil (Rohlf 2015). Groups of observations lying separately from the others might, in fact, reveal specimens with high measurement error (due, for example, to switching the order of two landmarks during the digitization). Similarly, an inspection of scatterplots of the first two principal components in any of the commonly used morphometric software could also reveal clear outliers. The advantage of this kind of software is that plots of the shape changes accounted for by the first principal component are readily available and immensely helpful in identifying landmarks that are clearly misplaced. MorphoJ implements a more complex exploratory procedure for the identification of outliers. Here, the cumulative curve of pairwise distances between observations expected for a multivariate normal distribution fitted to the data and the cumulative curve observed in the dataset at hand are compared. An inspection of such curve can help identifying the presence of outliers and a list of the specimens with their distance from the average shape is also provided with the possibility of excluding one or more specimens and evaluating how the observed curve changes. Importantly, the deviations for each specimen from the average shape can also be visualized as vectors at each landmark. This is, in turn, very helpful to identify not only landmarks whose labels have been swapped but also more subtle cases of errors in the digitization. These approaches are extremely useful during the preliminary analysis of a new dataset as they allow the identification of many errors. However, for the most part they can be considered procedures for the identification of outliers. Many procedures for the identification of multivariate outliers exist (Rohlf 1975; Penny and Jolliffe 2001; Jobe and Pokojovy 2014) and, although they can certainly help in identifying specimens with high levels of measurement error, they do not analyze measurement error per se. Rather, when repeated measurements of the same individual exist one could simply identify individual observations with high measurement error by looking at the individuals with the highest Procrustes distance among repeated measurements. Plots of the distribution of the distance among repeated measurement of the same individuals might, then, be a useful graphical aid.

Quantifying measurement error at specific landmarks

Sometimes, especially in the first stages of a morphometric study, a researcher might be concerned about the quality of chosen landmarks. In particular, he/she might want to choose to discard specific landmarks with high measurement error to keep a dataset of more reliable landmarks. Identifying landmarks with high measurement error is relatively difficult in common geometric morphometric studies as the generalized Procrustes analysis which is at the core of current geometric morphometric practice induces a non-independence in the error among landmarks. Therefore, a particularly high variation at one landmark is "distributed" to other landmarks—something that has also been dubbed "Pinocchio effect" (Chapman 1990; Zelditch et al. 2004). For this reason, often variation at a particular landmark or coordinate has relatively little meaning. This is also one of the reasons why chiefly multivariate statistical methods are used in the analysis of geometric morphometric data. Sometimes this problem can be overcome through a comparison to a "gold standard". For instance, in a recent study on bird bone shape (Provini et al. 2013) the Authors digitized a first set of landmarks eight times and then—for each landmark—compared the distance between the average landmark location and each of the eight separate measurements, discarding landmarks where the average distance was higher than twice the precision of the acquiring device (as declared by the constructor). Most of the time, however, researchers have been assessing measurement error at each landmark after generalized Procrustes analysis (thus being affected by the non-independence of errors created by this procedure). For instance, Adriaens' protocol (Adriaens 2007) suggests an exploratory approach where one removes from a dataset potentially problematic landmarks one at a time (or two at a time) and each time checks how this change affects the observed patterns of the scores along the first few principal components. Similar exploratory approaches are also common in the literature and include cases in which principal component analyses of repeated measures of a single individual are used to identify the landmarks that are most variable due to measurement error (Bastir et al. 2006). Another approach (Singleton 2002) consists in evaluating the variation across repeated measurements of the distance between each landmark and the centroid of the configuration of points, identifying landmarks with higher level of variation as more problematic. However, it has been shown that this approach can underestimate measurement error at selected landmarks if such error is not collinear with the vector connecting the centroid of the configuration of points and the true position of the landmark (von Cramon-Taubadel et al. 2007). A more recent approach (von Cramon-Taubadel et al. 2007) consists on superimposing the target configuration of points based only on three landmarks so that the measurement error at the other landmarks is retained and not "smeared" across them. Although this method is useful as it sidesteps the "Pinocchio effect", it has a potentially limiting set of assumptions. In fact, the three landmarks used as reference landmarks should have spherical measurement error, all of them should have the same levels of error and, possibly, they should also be far from each other relative to the complete configuration of points (von Cramon-Taubadel et al. 2007). One can imagine that it can be hard in some—or many—empirical datasets to find three reference landmarks to be used for this approach. When this method cannot be used, the exploratory approaches based on principal component analysis

of repeated measurements of a single individual can currently being considered a good working solution. A perhaps more articulated version of these exploratory methods could combine principal component analyses and averaging of repeated measurements (Fig. 2). In particular, one could produce plots of the variation accounted for by the first principal component in both the dataset with repeated measurements and the dataset where repeated measurements have been averaged. The expectation, then, would be that (often) landmarks containing high levels of random measurement error should load heavily along the first principal component in the dataset with repeated measurements but not in the dataset where repeated measurements have been averaged (Fig. 2). It should be noticed, however, that this approach can be considered only an extension of other exploratory approaches and it does not overcome the problems connected to the "Pinocchio effect".

Accounting for and minimizing measurement error

As mentioned, measurement error is ubiquitous in scientific research. One can only hope to account for and minimize the amount of error so that one gains higher biologically-relevant signal to noise ratio and this, in turn, produces higher statistical power. Obviously, many of the techniques used to identify and assess measurement error can also be used to make subjective decisions aimed at decreasing measurement error. For instance, after identifying subjects with high levels of measurement error, one can decide to exclude them from the analysis or acquire again the data. There are, however, two main non-mutually exclusive approaches whose goal is to account for and reduce measurement error: averaging of repeated measures and explicitly modelling error.

Averaging of repeated measurements

Taking multiple measurements and averaging them is perhaps the most commonly used technique to reduce measurement error both in geometric morphometric studies (Arnqvist and Mårtensson 1998; Fruciano et al. 2011b, 2012, 2014) and in other research fields (de Vet et al. 2003; Viswanathan 2005). The general idea behind this approach is that measurement error is random so averaging repeated measurements will approximate the unknown true value better than each separate measurement. One of its advantages is its simplicity as averages of repeated measurements are easily obtained in simple spreadsheets, general statistical software and in the widely used morphometric software MorphoJ. It should, however, be noticed that the rationale of this approach makes it ideal to reduce random measurement error but that in presence of non-random measurement error it can even have detrimental effects (Fig. 3). Another important point is that measurements should be repeated relative to one or more likely sources of measurement error. Perhaps not surprisingly, many empirical geometric morphometric studies (Muñoz-Muñoz et al. 2011; Fruciano et al. 2014) have used averages of repeated digitizations of the same specimens while much fewer studies have computed averages of repeated presentations, which require more effort to be obtained (but see Dvorak et al. 2006; Fruciano et al. 2011b, 2012; Schmieder et al. 2015). However, presentation can be a non-trivial source of error (see above).

Explicitly modelling measurement error

A different strategy—which is clearly more suitable when accounting for non-random measurement error—is to explicitly model the artifactual variation in a sample. This strategy has been used in a number of empirical studies (Fruciano et al. 2011b, 2012, 2014; Franchini et al. 2014; Ingram 2015), although it has been certainly less popular than averaging of repeated measures. It should, however, be stressed that explicitly modelling measurement error and averaging of repeated measures serve two different purposes and they are not mutually-exclusive. This strategy has been often applied to the problem of fish body arching—i.e. the fact that fish, being usually elongated, can be presented in different ways in front of the photo camera, which in turn generates artifactual variation due to different levels of body arching in a sample. An early application to this problem (Glasbey et al. 1995) assumed three landmarks to lie on a straight line in an unbent fish and then derived a geometrical transformation to "straighten" these three and all the other landmarks. These Authors also reported improved correct classification in linear discriminant analysis for the samples subjected to this correction for measurement error. A similar approach to the same

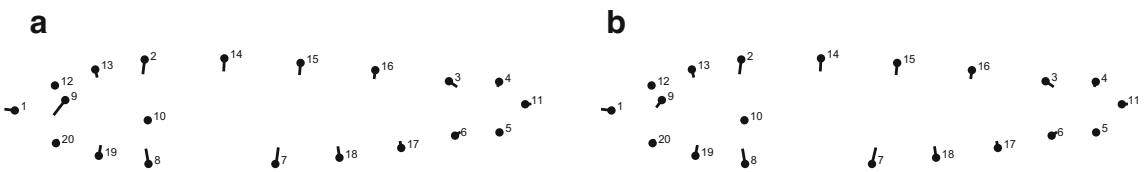

Fig. 2 Schematic representation (based on fictitious data) of an exploratory approach that can be used to identify error at specific landmarks. One can compare plots of the shape variation accounted for by the first principal component when using repeated measurements (**a**) and after averaging repeated measurements (**b**). Notice how the vector at *point 9* changes considerably between the *two plots*, while the rest remains relatively similar. Point 9 could then be a landmark extremely variable for non-biological reasons

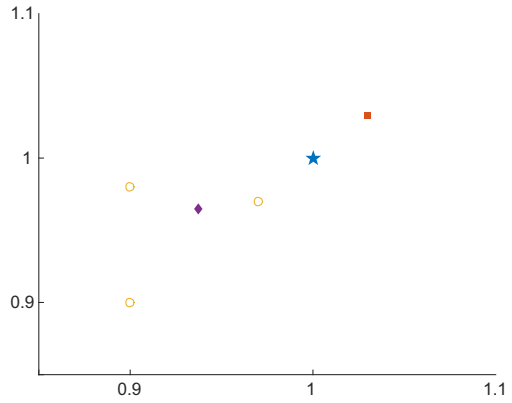

Fig. 3 Hypothetical situation in which averaging across repeated measurements can be detrimental. Each *point* represents a landmark in two dimensions. The *star* represents the true, unknown, position of the landmark, the *square* is the landmark as digitized by a hypothetical expert operator, the *empty circles* are hypothetical digitization from three novice operators. Finally, the *diamond* is the average of both the "novice" and the "expert" digitizations. It can be seen that there is a non-random pattern of measurement error, which differs between the "expert" and "novices". Using only the "expert" digitization would approximate better the true position of the landmark compared to taking an average of all measurements

problem—involving fitting of a quadratic function—has been since long implemented in the software tpsUtil (Rohlf 2015) and allows, by defining three points which should lie on a line in the unbent fish, to correct for a U-shaped arching. More recently, the software has been updated to fit a cubic function —requiring at least 4 points that are assumed to stay on a line in the unbent fish—thus also allowing for the correction of an S-shaped arching. Having been implemented in an easy-to-use software, it is not surprising that this function—aptly named 'Unbend specimens'—has been used in many empirical studies (Genner et al. 2007; Haas et al. 2010, 2015; Arbour et al. 2011; D'Anatro and Lessa 2011; Hirsch et al. 2013; Dennenmoser et al. 2015; Larouche et al. 2015; Santos-Santos et al. 2015). These methods, although powerful, rely on a specific set of assumptions. Namely, that the artifactual variation in presentation can be adequately described by the chosen function and that a certain set of points lies on a straight line in the unbent fish. These assumptions should be carefully considered when using this kind of approach as disregarding them can have negative consequences on downstream analyses. In fact, the position of the chosen reference points can have an effect on the final results and points assumed to lay along a straight line in unbent fish might not be present or easy to identify. Perhaps more importantly, any biologically-relevant variation in the position of the points chosen as reference is transferred to other points in the landmark configuration. A more general and flexible approach to modelling artifactual variation consists in modelling it as a multivariate vector and then projecting the original data to the multivariate subspace orthogonal to said vector. This type of approach has long been suggested as a way to account for variation in a sample due to allometry (Burnaby 1966; Rohlf and Bookstein 1987). More recently, the same idea has been applied to modelling and removing artifactual variation in articulated structures (Adams 1999) and datasets of human heads (Gharaibeh 2005) and fish bodies (Valentin et al. 2008; Fruciano et al. 2011b, 2012, 2014; Franchini et al. 2014; Ingram 2015). In its most recent formulation as a method for removing artifactual variation (Valentin et al. 2008), this approach consists in acquiring data on the same specimen, purposely presenting it at different levels of artifactual variation (for instance, different levels of body arching in fish or different levels of turning in human faces). Then a principal component analysis is performed and if the vast majority of the variance is explained by the first principal component, this is considered a good approximation of the modelled artifactual variation. To avoid that choosing a particular specimen as a model would affect the analysis, it is generally preferred to use multiple specimens for the same procedure (Valentin et al. 2008) and then compare the angles among vectors (i.e. the first principal component obtained for each of the specimens). If the pairwise angles between vectors are reasonably low—meaning that there is limited variation among models—then an average vector is computed. Finally, the Burnaby's procedure (Burnaby 1966) is used to project the original data in the multivariate subspace orthogonal to the average vector. The final result of this procedure is a set of landmark coordinates in which the modelled artifactual variation has been removed and that can be used in downstream analyses. The advantage of this technique is that it is relatively simple yet extremely powerful and flexible. In fact, artifactual variation can be modelled without choosing a priori a specific mathematical function but, rather, producing the expected deformation from the available samples. The set of assumptions of this approach is relatively limited and can be easily tested. In fact, the assumption that most of the artifactual variation can be modelled by a single vector in multivariate space is easily tested by checking the amount of explained variance while pairwise angles between vectors can be easily computed with the formula

$$a_{ij} = \arccos(i \cdot j) \tag{5}$$

where a_{ij} is the angle between each pair of vectors i and j and the operator $\cdot$ is the dot product. The choice of what constitutes an adequate amount of explained variance and how small should pairwise angles between vectors be is partly subjective, although one could test the latter using the formulas (Li 2011) implemented in MorphoJ or simulation procedures. One of the advantages of this method is its flexibility. It can, for instance, be used to remove the potentially confounding effect of sexual dimorphism in a sample comprising observations of unknown sex if sub-samples of individuals of known sex are available (Fruciano et al. 2014). Burnaby's procedure can also be used to project data to the subspace orthogonal not only to a single

direction (axis) but, potentially, even to an entire subspace of the total original multivariate space (comprising more than one direction). Considering the simplicity and flexibility of this approach, it is perhaps unfortunate that it has been used relatively scarcely in empirical studies. Perhaps, the implementation of this approach in easy-to-use software will favour its more widespread adoption. An interesting extension of the approach is to project the original dataset not to the subspace orthogonal to the vector chosen to represent artifactual variation but, rather, to project it on the vector. This would return scores along this vector for each observation of the original dataset which could, in turn, be used to identify specimens particularly affected by a given source of measurement error. These specimens with extreme levels of measurement error might, in fact, have specific artifactual features not present in specimens with milder levels of error.

A practical example

Dataset

Here, I use an existing dataset to show some of the approaches described above. An overview of the methods presented in this review can be found in Table 1, together with examples of software implementing them (including the ones used in the example here). This is not meant to be an exhaustive application of all possible approaches to assess and account for measurement error. The dataset has been previously used in two empirical morphometric studies (Fruciano et al. 2011b, 2012), where further details can also be found—with the addition of two Egyptian individuals used in a genetic study (Fruciano et al. 2011c). It comprises a total of 265 specimens of the labrid fish *Coris julis* collected at 10 Mediterranean sampling sites. Shortly after collection, fish were preserved in 95 % ethanol and individually marked. Pictures of the left side of each specimen were taken using a digital camera mounted on a copy stand with an experimental design in which every specimen had two presentations (two pictures) and two digitizations of landmarks for each presentation, for a total of four sets of coordinates. Twenty points (both landmarks and semilandmarks; Fig. 4) were digitized using the software tpsDig2 (Rohlf 2015). Throughout the data gathering phase, several measures have been taken to avoid, as much as possible, bias and error. In fact, pictures were taken using a relatively long camera-subject distance (495 mm) to minimize the effect of parallax (Mullin and Taylor 2002). Prior to taking pictures, each specimen was kept shortly (approximately 10 min) in water to rehydrate and then kept straight by running a long needle of appropriate length through the right side of the body from the caudal peduncle to the head (Windsor Aguirre, *pers. comm.*), evaluating by eye that the specimen was not dorso-ventrally arched relative to the antero-posterior axis of the body and that it stayed approximately "flat" and orthogonal with respect to the optical axis of the camera. The rehydration step facilitates the insertion of the above-mentioned needle. To better visualize the position of certain landmarks in the pictures, entomological pins were used. Each picture contained a scale bar. The gathering of data for geometric morphometric analyses was set up as a series of "sessions" (which were numbered consecutively for subsequent analyses). Each session comprised a "picture-taking" part and a digitization part. During the "picture-taking" phase each specimen was: put in water to rehydrate (approximately 10 min), straightened with the help of an appropriately-sized needle, marked with entomological pins, put in position and photographed for the first time; then all the pins (both the ones for landmark location and the one to straighten the fish) were removed, the fish put back in water for about 2 min then the whole process was repeated to take the second picture after which the specimen was put back in the storing jar and the process was started again with a new specimen. The aim of such a way of obtaining repeated measures is to take into account both preparation-related error (i.e. slightly different insertion of landmark pins) and presentation-related error (i.e. slightly different ways of presenting the fish relative to the digital camera).

In no case the digitization of landmarks took place on a different day. The same equipment (digital camera, copy stand, laptop computer, mouse, software) was used for all the sessions to avoid introducing further sources of variation; in a similar fashion, all the steps of the analysis were performed by the same operator to avoid inter-observer variation in the data. Individuals from each population were arranged in different sessions (days) so that a manageable number of specimens were photographed and digitized in a single day. Populations were not photographed and digitized as a whole (all the specimens of a certain population on a single day) but, instead, each population was divided in small sub-samples that were photographed and digitized in different sessions and in "rounds" so that each session usually comprised specimens from different populations and sub-samples of a certain population were always photographed in different, non-successive, sessions. The fish were also processed randomly with respect to body size. Such a design was employed to avoid that a potential change in time in the way the operator performed his tasks could produce bias in the analyses as the main biological effect of interest were the variation among populations (Fruciano et al. 2011b) and the morphological variation across growth (Fruciano et al. 2012). After all the specimens had been photographed and landmarks digitized, the position of landmarks/semilandmarks was carefully checked again by eye for each specimen in the dataset looking for sub-optimally placed points. Obtained landmark/semilandmark configurations were then subjected to a generalized Procustes analysis with the sliding of semilandmarks (Bookstein 1997). Sliding of semilandmarks and alignment of configurations were performed with tpsRelw (Rohlf 2007) using ten iterations and setting as sliding criterion the minimization of the squared Procrustes distance; this criterion was chosen because it removes all the tangential variation along

Table 1 Overview of methods for assessing and accounting for measurement error in geometric morphometric studies. The list of software implementation is not meant to be comprehensive and most of the methods, including the ones currently without a proper software implementation, can be implemented either manually or with limited effort in a scripting language

Purpose	Method	Software implementation
Measuring global agreement among measurements	Intraclass correlation coefficient	PAST (Hammer et al. 2001), SPSS R packages *irr* (Gamer et al. 2012), *psy* (Falissard 2012), *psych* (Revelle 2015) Manual (one-way ANOVA/Procrustes ANOVA and formulas based on mean squares)
Measuring global agreement among measurements	Procrustes ANOVA	MorphoJ R package *geomorph*
Quantifying the effect of using a 2D projection for 3D structures	Addition of a zero coordinate, common generalized Procrustes analysis and subtraction of group means followed by typical approaches to quantify/explore measurement error	Partly manual, partly with MorphoJ and general statistical packages
Testing for bias	Permutational procedures based on Procrustes distances with appropriate permutation scheme	–
Identifying subjects with high levels of measurement error	Detection of outliers with exploratory procedures	tpsSmall, MorphoJ, general statistical software (ordinations) R package geomorph
Identifying subjects with high levels of measurement error	Inspection of Procrustes distances among repeated measurements	tpsSmall, NTSYSpc (to compute Procrustes distances)
Identifying landmarks with high levels of measurement error	Inspections of principal component plots of repeated measurements/comparison of the patterns after averaging repeated measurements	Most morphometric and general statistical software
Identifying landmarks with high levels of measurement error	Analysis of the variation across repeated measurements of the distance between each landmark and the centroid of the configuration	–
Identifying landmarks with high levels of measurement error	Superimposition based only on a subset of landmarks	–
Reducing random measurement error	Averaging of repeated measurements	MorphoJ
Removing explicitly modelled measurement error	Fitting a mathematical function and applying the inverse function to the configuration of points	tpsUtil (fish body arching)
Removing explicitly modelled measurement error	Burnaby's method for projecting the data in the subspace orthogonal to a given vector/subspace	NTSYSpc

outlines whereas choosing the criterion of minimizing bending energy some of the tangential variation is retained (Perez et al. 2006). Centroid size was also computed in tpsRelw.

Assessing measurement error

A principal component analysis of this dataset (comprising a total of four configurations of points per specimen) in MorphoJ reveals that variation attributable to body arching is considerable (Fig. 5a), even if extra care had been taken to present the fish as straight as possible. Computing averages of repeated measurements for each specimen slightly reduces the problem but it does not seem to solve it (Fig. 5b). Computing repeatability for each relative warp (principal component) of the raw data by performing a one-way ANOVA in a general statistical package and then using Eqs. 1, 2 and 3 above reveals values that can be

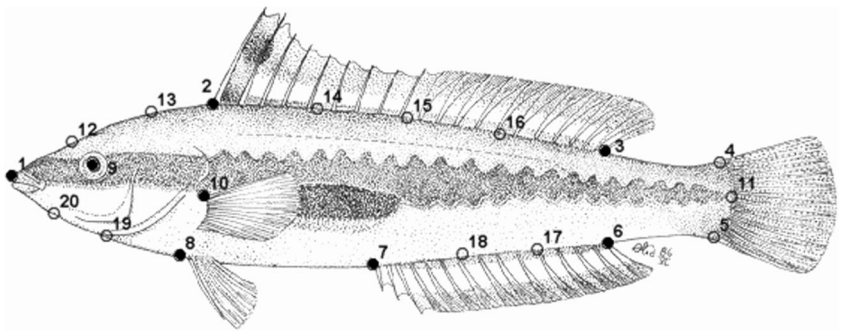

Fig. 4 Configuration of points used in the example. *Filled circles* represent landmarks, *empty circles* semilandmarks

considered medium-high with no pattern of low repeatability being concentrated in the higher-order principal components (Table 2). Performing a Procrustes ANOVA in MorphoJ on the same dataset using as error terms presentation and digitization reveals a non-trivial amount of measurement error (Table 3). In particular, a considerable—and significant—amount of variation among replicates can be attributed to presentation error as the mean squares for this term is more than 10 % than the mean squares for the "individual" term. On the other hand, the residual mean squares are only about 1.5 % than the mean squares for the "individual" term. This suggests that, following a reasonable expectation, presentation error is much larger than digitization error in this dataset. By performing a Procrustes ANOVA using only the "individual" term and the residual term (i.e. pooling the variation attributable to both presentation and digitization) and applying informally the Eqs. 1, 2 and 3 above to obtain a single number comprised between zero and one, we obtain a value of 0.84, similar to the values observed computing repeatability estimates for each relative warp separately. The repeatability for centroid size is 0.996; the estimated repeatability that would be obtained averaging across all four centroid size measurements is 0.999 (i.e. as observed in previous empirical studies, the repeatability for centroid size is higher than the repeatability for shape variables).

Reducing measurement error

The analyses above on the example dataset show that there is a certain amount of measurement error that is related to body arching and that cannot be taken into account by averaging repeated measurements. For this reason, I decided to subject the full dataset comprising four landmark configurations for each specimen to a procedure based on Burnaby's projection to remove the effect of body arching (Valentin et al. 2008), and then average the resulting coordinates of each specimen. The rationale for this choice is that first a certain type of error (which has a precise direction in shape space and it is unlikely to be distributed randomly around the mean in a small number of repeated measures) is removed as much as possible, then the residual error is further reduced by averaging the repeated measures.

To this end, a specimen from each population was photographed in six different positions related to different levels of arching. In particular, while this artifactual variation is normally modelled as a C-shaped dorso-ventral arching of the fish (Valentin et al. 2008), being *C. julis* an elongated fish, I decided to model this variation as both a C-shaped dorso-ventral arching and an S-shaped arching (i.e. head and tail pointing in opposite directions). I then digitized the 20 landmark and semilandmark points on each set of six pictures. I then separately subjected each set of six landmark configurations to a generalized Procustes analysis with the sliding of semilandmarks (Bookstein 1997) followed by a computation of the first eigenvector in NTSYSpc 2.2 (Rohlf 2005). I also computed pairwise angles among the first eigenvectors (scaled to unit size) of each principal component analysis using Eq.. 5 above. Being the angles between vectors relatively small (Table 4), I decided to use the average vector and I subjected the whole dataset comprising the four configurations per individual to a projection into the subspace orthogonal to such vector using the software NTSYSpc. After performing a new generalized Procrustes analysis with sliding of semilandmarks, a principal component analysis on this new dataset—now corrected for body arching—reveals that the main variation corresponds to how deep the body is (Fig. 6), which is a pattern that has clear biological explanations (Fruciano et al. 2011b, 2012) and that is maintained when averages of repeated measurements are computed (Fig. 6b). Computing the repeatability of each relative warp (Table 1) results in most cases in better values. Similarly, when performing a Procrustes ANOVA on this new dataset corrected for body arching (Table 5), we notice that variation among presentations is still the main source of error but now the mean squares for this term is lower (less than 8 % than the mean squares for the "individual" term), while the measurement error associated to digitization is still negligible (around 1.5 %). When pooling the multiple repetitions in a single Procrustes ANOVA term and applying informally the Eqs. 1, 2 and 3 above, we obtain a value of 0.87, slightly higher than the one obtained in the dataset affected by body arching. From these analyses one can conclude that, although the repeatability of each relative warp is good, it is a good idea to take averages of repeated measurements to increase the quality of the data. Perhaps, if this was a pilot study, one could also decide to take only repeated presentations but not repeated digitizations in the full dataset, as the error due to digitization

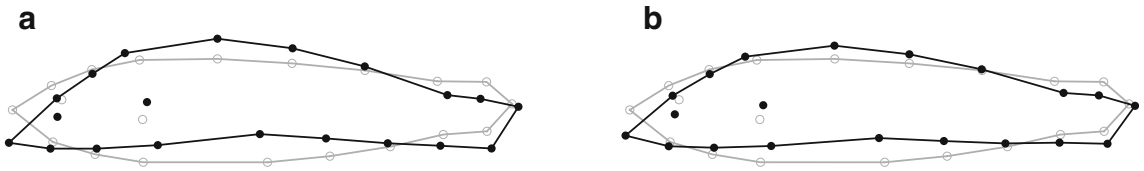

Fig. 5 Shape variation accounted for by the first principal component in the original dataset with repeated measurements (**a**) and after averaging across the repeated measurements (**b**). In both cases, the *points in grey* represent the average shape while the *shape in black* represents the shape predicted by the first principal component at a score about double the observed extreme

Table 2 Repeatability for each of the 36 relative warps both as raw data and after the removal of body arching effect. Estimates of the repeatability after averaging the four repeated measures are also provided

	Raw data		Dataset after removal of body arching	
	Repeatibility	Repeatibility after averaging	Repeatibility	Repeatibility after averaging
RW1	0.789	0.937	0.822	0.949
RW2	0.839	0.954	0.919	0.978
RW3	0.720	0.912	0.721	0.912
RW4	0.825	0.949	0.867	0.963
RW5	0.918	0.978	0.838	0.954
RW6	0.749	0.923	0.762	0.928
RW7	0.859	0.961	0.921	0.979
RW8	0.824	0.949	0.776	0.933
RW9	0.832	0.952	0.773	0.931
RW10	0.667	0.889	0.841	0.955
RW11	0.839	0.954	0.852	0.958
RW12	0.805	0.943	0.860	0.961
RW13	0.801	0.941	0.728	0.915
RW14	0.840	0.955	0.762	0.928
RW15	0.831	0.952	0.837	0.954
RW16	0.698	0.902	0.823	0.949
RW17	0.910	0.976	0.885	0.969
RW18	0.842	0.955	0.811	0.945
RW19	0.890	0.970	0.925	0.980
RW20	0.849	0.958	0.937	0.984
RW21	0.824	0.949	0.840	0.954
RW22	0.916	0.977	0.926	0.980
RW23	0.848	0.957	0.779	0.934
RW24	0.770	0.930	0.760	0.927
RW25	0.897	0.972	0.895	0.972
RW26	0.898	0.972	0.855	0.959
RW27	0.598	0.856	0.748	0.922
RW28	0.795	0.939	0.826	0.950
RW29	0.816	0.947	0.892	0.971
RW30	0.934	0.983	0.874	0.965
RW31	0.773	0.932	0.900	0.973
RW32	0.852	0.958	0.891	0.970
RW33	0.834	0.953	0.907	0.975
RW34	0.798	0.941	0.913	0.977
RW35	0.770	0.931	0.763	0.928
RW36	0.890	0.970	0.875	0.966
Mean	0.821	0.947	0.842	0.954

is extremely low in the first place. Looking at the values of repeatability and to the results of the Procrustes ANOVA before and after the procedure of removal of body arching, one might think that this procedure has a limited effect on the quality of the dataset. Quite to the contrary, the method for the removal of body arching acts on a different type of error, which is not necessarily reflected by measures of repeatability. I also tested for the effect of the removal of body arching by computing in NTSYSpc the tangent Procustes distances among different repetitions of the same individual in each dataset and then computing their averages. I, then, compared tangent Procustes distances among repeated measures before and after the application of the method for removal of body arching effect using in STATISTICA (StatSoft, Inc.) a Wilcoxon signed-rank test (Wilcoxon 1945), which is appropriate in the case paired samples. This test shows a significant

Table 3 Procrustes ANOVA on the raw example dataset

Effect	SS	MS	df	*F*	*p*	Intraclass correlation
Individual	0.662	6.96E-05	9504	9.84	<0.0001	0.84
Photo	0.068	7.08E-06	9540	6.8	<0.0001	
Residual (digitization)	0.020	1.04E-06	19080			

For each effect, the sum of squares (*SS*), mean squares (*MS*), degrees of freedom (*df*), *F* statistics and its significance are provided. Notice that under the "Intraclass correlation" heading, the value based on pooling multiple digitizations is provided, as discussed in the text

reduction ($p < 0.001$) in the average tangent Procustes distance among repeated measurements after performing the adjustment for body arching. This in practice means that, although the values of repeatability are relatively unaffected by the adjustment for body arching, body arching was still an important source of artifactual variation in the dataset.

Other analyses

I decided to use the same dataset for a few more analyses of methodological interest. As mentioned above, each "session" of data gathering was consecutively numbered. I then decided to test if time (session number) or size of the specimen had an effect on the consistency of the operator. To this aim, I used a regression approach by regressing the average tangent Procrustes distance among repetitions (a measure of measurement error for each specimen) on either session number or centroid size. To avoid that a particularly "good" or "bad" day (or a few specimens) would affect the analysis, I have also used an outlier-removal procedure based on studentized residuals. This choice was motivated by the fact that preliminary plots of average tangent Procustes distances among repeated measures on session number and centroid size showed a possible linear relationship but with a few outliers with very high average distance among repeated measures. Given that these outliers did not show any particular association with certain sessions or certain size classes, given that the general pattern of association among variables was of interest, and given that linear regressions are well known to be sensitive to the presence of outliers, an outlier removal procedure was deemed appropriate. Therefore, in STATISTICA, I performed separate regressions for average distance among repeated measures on session number and centroid size and computed studentized residuals. Cases whose studentized residual was higher than 1.9 (value subjectively deemed conservatively appropriate based on a t-distribution table) were then removed. The reduced datasets were then used to perform a new regression. Both the regression of average distances among repeated measurements on session number ($R = 0.4$, $p < 0.001$ for raw data and $R = 0.5$, $p < 0.001$ for body-arching-adjusted data) and on centroid size ($R = 0.2$, $p = 0.001$ for raw data and $R = 0.4$, $p < 0.001$ for body-arching-adjusted data) are significant, indicating that the average distances among repeated measurements tend to drop as the session number or the centroid size grow. This, in practice, confirms that the operator becomes more consistent as he/she accumulates experience and, possibly, that higher consistency is achieved with larger specimens as these are easier to handle.

To test if there was a non-random error affecting shape over time (session) when measures to reduce the measurement error have been taken and the allometric component has been removed, I performed in STATISTICA a regression of shape (relative warp scores) on session number using the dataset of

Table 4 Pairwise angles between eigenvectors of each model used to summarize artifactual dorso-ventral arching. The names correspond to the sampling site where the specimen was collected

	Augusta	Lecce	Mallorca	Mazara	Naples	Oristano	Pantelleria	Riposto	Split
Augusta	–	–	–	–	–	–	–	–	–
Lecce	11.26	–	–	–	–	–	–	–	–
Mallorca	16.85	15.04	–	–	–	–	–	–	–
Mazara	10.75	10.94	18.70	–	–	–	–	–	–
Naples	15.76	16.54	16.28	15.72	–	–	–	–	–
Oristano	16.32	15.95	22.74	13.85	20.25	–	–	–	–
Pantelleria	12.35	12.03	16.11	9.99	16.16	9.72	–	–	–
Riposto	13.76	9.29	13.08	15.09	18.13	19.74	15.08	–	–
Split	15.38	14.81	19.36	10.17	15.31	12.80	11.66	18.23	–

For further details on sampling sites and biologically-relevant results, see Fruciano et al. (2011b); (Fruciano et al. 2011c); Fruciano et al. (2012)

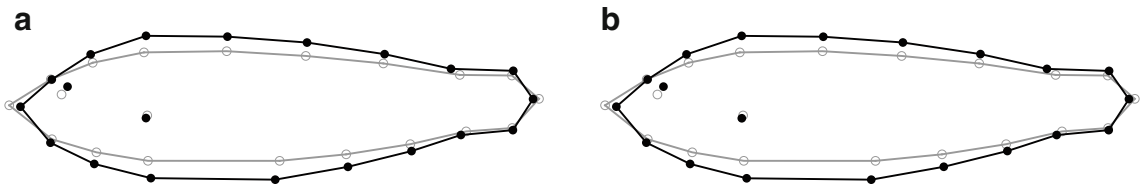

Fig. 6 Shape variation accounted for by the first principal component in the dataset corrected for body arching using repeated measurements (**a**) and after averaging across the repeated measurements (**b**). *Plots* produced as for Fig. 5

shapes after the correction for body arching, averaging of repeated measures and regression-based removal of size-related allometric component [that is, the same dataset used for the biological analysis of shape differences in geographic space; (Fruciano et al. 2011b)]. Regression of shape on session number using shapes with reduced measurement error and size-related allometric component removed is highly significant ($p = 0.00004$) but explains very little variance (1.83 % of total variance). This can be interpreted with a small bias accumulating as the operator gets more experienced. The relatively small amount of explained variance, however, suggests that this aspect might be of very little practical significance.

To study the effect of certain confounding factors on the procedure of removal of body arching, I computed a series of correlations—testing their significance with Mantel tests (Mantel 1967) in NTSYSpc—between on one hand the pairwise angles between the model eigenvectors and, on the other hand, distances among the specimens used to create the models. In particular, a binary matrix encoding if two specimens used as models had the same colour phase [these fish are protogynous hermaphrodites that change colour and body shape as they change sex, see Fruciano et al. (2012) for details] was used to test for the effect of difference in colour phase on the angles, while differences in centroid size were used to test the effect of using specimens of different size to produce the models. Finally, coordinates for the specimens used as models were obtained from the dataset used in biological analyses (that is, coordinates after correction for body arching and averaging of repeated measures) and aligned via a generalized Procustes analysis, computing then pairwise tangent Procustes distances. These pairwise tangent Procustes distances were used to test for the effect of the shape of specimens used as models on the angles. These correlations of pairwise angles between the model eigenvectors and, respectively, difference in centroid size, difference in colour phase and pairwise tangent Procustes distances among model specimens are all low (range 0.19–0.3) and always not significant ($p > 0.05$). These results suggest that none of these factors (size or shape of the specimens used as models) seems to affect the modelling of the artifactual variation, at least in this dataset. It should be, however, noticed that, while the models used here comprise specimens with different shape due to sexual dimorphism, the variation among them is probably still relatively small. It is, then, entirely possible that in analyses of multiple species the choice of species to model artifactual variation could affect the analyses to a larger extent.

Conclusions and final remarks

Here, I have briefly reviewed, in the light of the latest empirical findings, the most common sources of measurement error and commonly used approaches to assess and account for it. It cannot be stressed enough that, being measurement error ubiquitous, sources of error can be found at many different stages of a geometric morphometric study. It is noteworthy that, apart from the obvious sources of measurement error, more subtle sources of measurement error, such as changes in operator consistency over time, do exist. It is, therefore, the responsibility of each researcher to take into consideration possible sources of measurement error and take decisions about which sources of error should be assessed and corrected for. In fact, a 'one fits all' solution does not exist as the feasibility of different procedures depends on the specific experimental design. For instance, accounting for presentation error will be impossible in certain cases, such as when dealing with live specimens to be released as soon as possible to minimize stress. The fact that many of the procedures to assess and account for measurement error are exploratory in nature and provide no

Table 5 Procrustes ANOVA on the dataset corrected for body arching

Effect	SS	MS	df	*F*	*p*	Intraclass correlation
Individual	0.471	4.95E-05	9504	12.65	<0.0001	0.87
Photo	0.037	3.91E-06	9540	5.12	<0.0001	
Residual (digitization)	0.015	7.65E-07	19080			

Abbreviations as in Table 2

exact cut-offs on which to base decisions, makes such decisions also partially subjective. Nonetheless, assessing measurement error allows the researcher—and readers of empirical studies—to have a perception of the robustness of a given dataset and of the inferential results based on it. Although assessing measurement error can be generally considered a cost in terms of time and resources, this practice can also help in saving time and using resources in a more productive way. In this context, the widely held suggestion to assess measurement error in a small pilot study can be certainly recommended as, although estimates of error will be less robust, this approach can help using time and resources in the most productive way. Imagine, for instance, that a pilot study shows that levels of measurement error are negligible relative to the biological effect of interest: in this case, one can perform the same planned study without taking repeated measurements, thus completing the study in less time or using the time and resources to increase sample sizes, which increase statistical power. For these reasons, although measurement error is currently for the most part neglected in empirical geometric morphometric studies, one can only hope that more and more researchers will start measuring and accounting for it in the future.

Acknowledgments My gratitude goes to Venera Ferrito for her continued support. I also thank her because—when she was acting as my supervisor—she has let me pursue my methodological interests even when they were not immediately related to the biological investigation we were carrying out. I am deeply grateful to F. James Rohlf for exposing me to the first few papers I have ever read on the fascinating subject of measurement error, for his comments to a very early version of this manuscript and for his continued support. The insightful comments of two reviewers have greatly contributed to improving this review.

References

Adams DC (1999) Methods for shape analysis of landmark data from articulated structures. Evol Ecol Res 1:959–970

Adams DC, Felice RN (2014) Assessing trait covariation and morphological integration on phylogenies using evolutionary covariance matrices. PLoS One 9: e94335. doi:10.1371/journal.pone.0094335

Adams DC, Otárola-Castillo E (2013) Geomorph: an r package for the collection and analysis of geometric morphometric shape data. Methods Ecol Evol 4:393–399. doi:10.1111/2041-210X.12035

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

Adams DC, Rohlf FJ, Slice DE (2013) A field comes of age: geometric morphometrics in the 21st century. Hystrix Ital J Mammal 24: 7–14. doi:10.4404/hystrix-24.1-6283

Adriaens D (2007) Protocol for error testing in landmark based geometric morphometrics http://www.fun-morph.ugent.be/Miscel/Methodology/Morphometrics.pdf

Ahrens H (1976) Multivariate variance-covariance components (MVCC) and generalized intraclass correlation coefficient (GICC). Biom Z 18:527–533. doi:10.1002/bimj.19760180703

Alibert P, Moureau B, Dommergues JL, David B (2001) Differentiation at a microgeographical scale within two species of ground beetle, *Carabus auronitens* and *C. nemoralis* (Coleoptera, Carabidae): a geometrical morphometric approach. Zool Scr 30:299–311

Anderson TW (1963) Asymptotic theory for principal component analysis Annals of Mathematical Statistics:122-148

Arbour JH, Hardie DC, Hutchings JA (2011) Morphometric and genetic analyses of two sympatric morphs of Arctic char (*Salvelinus alpinus*) in the Canadian High Arctic. Can J Zool 89:19–30. doi:10.1139/Z10-100

Arnqvist G, Mårtensson T (1998) Measurement error in geometric morphometrics: empirical strategies to assess and reduce its impact on measures of shape. Acta Zool Acad Sci Hung 44:73–96

Ayala D, Caro-Riaño H, Dujardin J-P, Rahola N, Simard F, Fontenille D (2011) Chromosomal and environmental determinants of morphometric variation in natural populations of the malaria vector *Anopheles funestus* in Cameroon Infection. Genet Evol 11:940–947

Bandyopadhyay S, Ganguli B, Chatterjee A (2011) A review of multivariate longitudinal data analysis. Stat Methods Med Res 20:299–330

Barrow E, Macleod N (2008) Shape variation in the mole dentary (Talpidae: Mammalia). Zool J Linn Soc 153:187–211

Bastir M, Rosas A, O'Higgins P (2006) Craniofacial levels and the morphological maturation of the human skull. J Anat 209:637–654

Berbel-Filho W, Jacobina U, Martinez P (2013) Preservation effects in geometric morphometric approaches: freezing and alcohol in a freshwater fish. Ichthyol Res 60:268–271. doi:10.1007/s10228-013-0339-x

Bonneau N, Bouhallier J, Simonis C, Baylac M, Gagey O, Tardieu C (2012) Technical note: shape variability induced by reassembly of human pelvic bones. Am J Phys Anthropol 148:139–147

Bookstein FL (1991) Morphometric Tools for Landmark Data vol null. Cambridge University Press, Cambridge/New York/Port Chester/Melbourne/Sydney

Bookstein F (1993) A brief history of the morphometric synthesis. In: Leslie F. Marcus, Elisa Bello, Antonio Garcia-Valdecasas (eds) *Contributions to Morphometrics*. Museo Nacional de Ciencias Naturales, Madrid, p 15–40.

Bookstein FL (1997) Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Med Image Anal 1:225–243. doi:10.1016/S1361-8415(97)85012-8

Bookstein F (2015) Integration, Disintegration, and Self-Similarity: Characterizing the Scales of Shape Variation in Landmark Data Evol Biol 42(4):395–426. doi:10.1007/s11692-015-9317-8

Breuker CJ, Patterson JS, Klingenberg CP (2006) A single basis for developmental buffering of Drosophila wing shape. PLoS One 1: e7

Bromiley PA, Schunke AC, Ragheb H, Thacker NA, Tautz D (2014) Semi-automatic landmark point annotation for geometric morphometrics. Front Zool 11:61

Burnaby T (1966) Growth-invariant discriminant functions and generalized distances Biometrics 22:96–110

Campomanes-Álvarez B, Ibáñez O, Navarro F, Alemán I, Cordón O, Damas S (2015) Dispersion assessment in the location of facial landmarks on photographs. Int J Leg Med 129:227–236

Cardini A (2014) Missing the third dimension in geometric morphometrics: how to assess if 2D images really are a good proxy for 3D structures? Hystrix Ital J Mammal 25:73–81

Cardini A, Elton S (2007) Sample size and sampling error in geometric morphometric studies of size and shape. Zoomorphology 126:121–134. doi:10.1007/s00435-007-0036-2

Cardini A, Elton S (2009) Geographical and taxonomic influences on cranial variation in red colobus monkeys (Primates, Colobinae): introducing a new approach to 'morph'monkeys. Glob Ecol Biogeogr 18:248–263

Cardini A, Jansson AU, Elton S (2007) A geometric morphometric approach to the study of ecogeographical and clinal variation in vervet monkeys. J Biogeogr 34:1663–1678

Chang J, Alfaro ME (2015) Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. Methods Ecol Evol doi:10.1101/023382

Chapman RE (1990) Conventional Procrustes approaches. In: Proceedings of the Michigan Morphometrics Workshop. University of Michigan Museum of Zoology, Ann Arbor, pp 251–267

Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Costa C (2011) The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. J Clin Epidemiol 64:264–269

D'Anatro A, Lessa EP (2011) Phenotypic and genetic variation in the white croaker *Micropogonias furnieri* Desmarest 1823 (Perciformes: Sciaenidae): testing the relative roles of genetic drift and natural selection on population divergence. J Zool 285:139–149. doi:10.1111/j.1469-7998.2011.00823.x

de Vet HC, Terwee CB, Bouter LM (2003) Current challenges in clinimetrics. J Clin Epidemiol 56:1137–1141

Debat V, Béagin M, Legout H, David JR (2003) Allometric and nonallometric components of Drosophila wing shape respond differently to developmental temperature. Evolution 57:2773–2784

Dennenmoser S, Nolte AW, Vamosi SM, Rogers SM (2015) Phylogeography of the prickly sculpin (*Cottus asper*) in northwestern North America reveals parallel phenotypic evolution across multiple coastal–inland colonizations. J Biogeogr 42:1626–1638. doi:10.1111/jbi.12527

Dryden IL, Mardia KV (1998) Statistical shape analysis vol 4. Wiley Chichester

Dujardin J-PA, Kaba D, Henry AB (2010) The exchangeability of shape. BMC Res Notes 3:266

Dvorak V, Aytekin A, Alten B, Skarupova S, Votypka J, Volf P (2006) A comparison of the intraspecific variability of *Phlebotomus sergenti* Parrot, 1917 (Diptera: Psychodidae). J Vector Ecol 31:229–238

El Ghaziri A, Qannari EM (2015) Measures of association between two datasets; Application to sensory data Food Quality and Preference 40, Part A:116-124 doi:http://dx.doi.org/10.1016/j.foodqual.2014.09.010

Escoufier Y (1973) Le traitement des variables vectorielles. Biometrics 29:751–760. doi:10.2307/2529140

Fadda C, Faggiani F, Corti M (1997) A portable device for the three dimensional landmark collection of skeletal elements of small mammals. Mammalia 61:622–627

Fagertun J, Harder S, Rosengren A, Moeller C, Werge T, Paulsen RR, Hansen TF (2014) 3D facial landmarks: inter-operator variability of manual annotation. BMC Med Imaging 14:35

Falissard B (2012) psy: Various procedures used in psychometry

Fisher RA (1958) Statistical methods for research workers. Oliver and Boyd

Fleiss J, Shrout P (1977) The effects of measurement errors on some multivariate procedures. Am J Public Health 67:1188–1191

Franchini P, Fruciano C, Spreitzer ML, Jones JC, Elmer KR, Henning F, Meyer A (2014) Genomic architecture of ecologically divergent body shape in a pair of sympatric crater lake cichlid fishes. Mol Ecol 23:1828–1845. doi:10.1111/mec.12590

Franklin D, Oxnard CE, O'Higgins P, Dadour I (2007) Sexual dimorphism in the subadult mandible: quantification using geometric morphometrics. J Forensic Sci 52:6–10

Fruciano C, Tigano C, Ferrito V (2011a) Traditional and geometric morphometrics detect morphological variation of lower pharyngeal jaw in *Coris julis* (Teleostei, Labridae). Ital J Zool 78:320–327. doi:10.1080/11250003.2010.547876

Fruciano C, Tigano C, Ferrito V (2011b) Geographical and morphological variation within and between colour phases in *Coris julis* (L. 1758), a protogynous marine fish. Biol J Linn Soc 104(148):148–162. doi:10.1111/j.1095-8312.2011.01700.x

Fruciano C, Hanel R, Debes P, Tigano C, Ferrito V (2011c) Atlantic-Mediterranean and within-Mediterranean molecular variation in *Coris julis* (L. 1758) (Teleostei, Labridae). Mar Biol 158:1271–1286. doi:10.1007/s00227-011-1647-1

Fruciano C, Tigano C, Ferrito V (2012) Body shape variation and colour change during growth in a protogynous fish. Environ Biol Fishes 94:615–622. doi:10.1007/s10641-011-9968-y

Fruciano C, Franchini P, Meyer A (2013) Resampling-based approaches to study variation in morphological modularity. PLoS One 8:e69376

Fruciano C, Pappalardo AM, Tigano C, Ferrito V (2014) Phylogeographical relationships of Sicilian brown trout and the effects of genetic introgression on morphospace occupation. Biol J Linn Soc 112:387–398. doi:10.1111/bij.12279

Gamer M, Lemon J, Singh IFP (2012) irr: Various Coefficients of Interrater Reliability and Agreement

Genner MJ, Nichols P, Carvalho GR, Robinson RL, Shaw PW, Turner GF (2007) Reproductive isolation among deep-water cichlid fishes of Lake Malawi differing in monochromatic male breeding dress. Mol Ecol 16:651–662. doi:10.1111/j.1365-294X.2006.03173.x

Gharaibeh W (2005) Correcting for the effect of orientation in geometric morphometric studies of side-view images of human heads. In: Slice D (ed) Modern Morphometrics in Physical Anthropology. Developments in Primatology: Progress and Prospects. Springer US, pp 117-143. doi:10.1007/0-387-27614-9_5

Glasbey CA, Horgan GW, Gibson GJ, Hitchcock D (1995) Fish shape analysis using landmarks. Biom J 37:481–495. doi:10.1002/bimj.4710370408

Gonzalez P, Bernal V, Perez S (2011) Analysis of sexual dimorphism of craniofacial traits using geometric morphometric techniques. Int J Osteoarchaeol 21:82–91

Gunz P, Mitteroecker P (2013) Semilandmarks: a method for quantifying curves and surfaces. Hystrix Ital J Mammal 24:103–109

Haas TC, Blum MJ, Heins DC (2010) Morphological responses of a stream fish to water impoundment. Biol Lett 6:803–806

Haas TC, Heins DC, Blum MJ (2015) Predictors of body shape among populations of a stream fish (*Cyprinella venusta*, Cypriniformes: Cyprinidae). Biol J Linn Soc 115:842–858

Hammer Ø, Harper D, Ryan P (2001) PAST: Paleontological Statistics Software: Package for Education and Data Analysis Palaeontologia Electronica 4

Henry A, Thongsripong P, Fonseca-Gonzalez I, Jaramillo-Ocampo N, Dujardin J-P (2010) Wing shape of dengue vectors from around the world Infection. Genet Evol 10:207–214

Hirsch PE, Eckmann R, Oppelt C, Behrmann-Godel J (2013) Phenotypic and genetic divergence within a single whitefish form—detecting the potential for future divergence. Evol Appl 6:1119–1132

Hood CS, Heins DC, McEachran J (2000) Ontogeny and allometry of body shape in the blacktail shiner, *Cyprinella venusta*. Copeia 2000:270–275

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218. doi:10.1007/BF01908075

Ingram T (2015) Diversification of body shape in Sebastes rockfishes of the north-east Pacific. Biol J Linn Soc 116: 805–818. doi:10.1111/bij.12635

Jobe JM, Pokojovy M (2014) A cluster-based outlier detection scheme for multivariate data. Journal of the American Statistical Association 110:1543–1551 doi:10.1080/01621459.2014.983231

Kitthawee S, Dujardin J-P (2009) The *Diachasmimorpha longicaudata* complex: reproductive isolation and geometric patterns of the wing. Biol Control 51:191–197

Klingenberg CP (2009) Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. Evol Dev 11:405–421. doi:10.1111/j.1525-142X.2009.00347.x

Klingenberg CP (2011) MorphoJ: an integrated software package for geometric morphometrics. Mol Ecol Res 11:353–357

Klingenberg CP, Gidaszewski NA (2010) Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Syst Biol 59: 245–261

Klingenberg CP, McIntyre GS (1998) Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with Procrustes methods. Evolution 52:1363-1375

Klingenberg CP, Leamy LJ, Routman EJ, Cheverud JM (2001) Genetic architecture of mandible shape in mice: effects of quantitative trait loci analyzed by geometric morphometrics. Genetics 157:785–802

Klingenberg CP, Barluenga M, Meyer A (2002) Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. Evolution 56:1909–1920. doi:10.1111/j.0014-3820.2002.tb00117.x

Klingenberg CP, Wetherill L, Rogers J, Moore E, Ward R, Autti-Rämö I, Fagerlund Å, Jacobson SW, Robinson LK, Hoyme HE, Mattson SN, Li TK, Riley EP, Foroud T, CIFASD Consortium (2010) Prenatal alcohol exposure alters the patterns of facial asymmetry. Alcohol 44: 649–657. doi:10.1016/j.alcohol.2009.10.016

Laffont R, Renvoisé E, Navarro N, Alibert P, Montuire S (2009) Morphological modularity and assessment of developmental processes within the vole dental row (*Microtus arvalis*, Arvicolinae, Rodentia). Evol Dev 11:302–311

Langerhans RB, Gifford ME, Joseph EO (2007) Ecological speciation in *Gambusia* fishes. Evolution 61:2056–2074

Larouche O, Cloutier R, Zelditch ML (2015) Head body and fins: patterns of morphological integration and modularity in fishes. Evol Biol 42: 296–311

Leamy LJ, Klingenberg CP (2005) The genetics and evolution of fluctuating asymmetry Annual Review of Ecology. Evol Syst 36:1–21

Leamy LJ, Klingenberg CP, Sherratt E, Wolf JB, Cheverud JM (2015) The genetic architecture of fluctuating asymmetry of mandible size and shape in a population of mice: another look. Symmetry 7:146–163

Li S (2011) Concise formulas for the area and volume of a hyperspherical cap. Asian J Math Stat 4:66–70

Loy A, Ciccotti E, Ferrucci L, Cataudella S (1996) An application of automated feature extraction and geometric morphometrics: temperature-related changes in body form of *Cyprinus carpio* juveniles. Aquac Eng 15:301–311. doi:10.1016/0144-8609(95)00016-X

Maga AM, Navarro N, Cunningham ML, Cox TC (2015) Quantitative trait loci affecting the 3D skull shape and size in mouse and prioritization of candidate genes in-silico. Front Physiology 6:92. doi:10.3389/fphys.2015.00092

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Martinez P, Berbel-Filho W, Jacobina U (2013) Is formalin fixation and ethanol preservation able to influence in geometric morphometric analysis? Fishes as a case study. Zoomorphology 132:87–93. doi:10.1007/s00435-012-0176-x

Mitteroecker P, Bookstein F (2011) Linear discrimination ordination, and the visualization of selection gradients in modern morphometrics. Evol Biol 38:100–114. doi:10.1007/s11692-011-9109-8

Müller R, Büttner P (1994) A critical discussion of intraclass correlation coefficients. Stat Med 13:2465–2476

Mullin SK, Taylor PJ (2002) The effects of parallax on geometric morphometric data. Comput Biol Med 32:455–464

Muñoz-Muñoz F, Sans-Fuentes M, López-Fuster M, Ventura J (2011) Evolutionary modularity of the mouse mandible: dissecting the effect of chromosomal reorganizations and isolation by distance in a Robertsonian system of *Mus musculus domesticus*. J Evol Biol 24: 1763–1776

Nikolakakis S, Bossier P, Kanlis G, Dierckens K, Adriaens D (2014) Protocol for quantitative shape analysis of deformities in early larval European seabass *Dicentrarchus labrax*. J Fish Biol 84:206–224

O'Higgins P, Jones N (1998) Facial growth in *Cercocebus torquatus*: an application of three-dimensional geometric morphometric techniques to the study of morphological variation. J Anat 193:251–272

Olsen AM, Westneat MW (2015) StereoMorph: an R package for the collection of 3D landmarks and curves using a stereo camera setup. Methods Ecol Evol 6:351–356. doi:10.1111/2041-210X.12326

Osis ST, Hettinga BA, Macdonald SL, Ferber R (2015) A novel method to evaluate error in anatomical marker placement using a modified generalized Procrustes analysis. Comput Methods Biomech Biomed Eng 18:1108–1116

Penny KI, Jolliffe IT (2001) A comparison of multivariate outlier detection methods for clinical laboratory safety data. J Royal Stat Soc: Ser D (The Statistician) 50:295–307

Perez SI, Bernal V, Gonzalez PN (2006) Differences between sliding semi-landmark methods in geometric morphometrics, with an application to human craniofacial and dental variation. J Anat 208:769–784

Posnien N, Hopfen C, Hilbrant M, Ramos-Womack M, Murat S, Schönauer A, Herbert SL, Nunes MDS, Arif S, Breuker CJ, Schlötterer C, Mitteroecker P, McGregor AP Aain (2012) Evolution of eye morphology and Rhodopsin expression in the *Drosophila melanogaster* species subgroup. PLoS One 7: e37346. doi:10.1371/journal.pone.0037346

Provini P, Simonis C, Abourachid A (2013) Functional implications of the intertarsal joint shape in a terrestrial (*Coturnix coturnix*) versus a semi-aquatic bird (*Callonetta leucophrys*). J Zool 290:12–18

Qannari EM, Courcoux P, Faye P (2014) Significance test of the adjusted Rand index. Application to the free sorting task. Food Quality and Preference 32, Part A:93-97 doi:http://dx.doi.org/10.1016/j.foodqual.2013.05.005

Revelle W (2015) psych: procedures for psychological, psychometric, and personality Research. Evanston, Illinois

Riaño HC, Jaramillo N, Dujardin J-P (2009) Growth changes in *Rhodnius pallescens* under simulated domestic and sylvatic conditions. Infect Genet Evol 9:162–168. doi:10.1016/j.meegid.2008.10.009

Rohlf FJ (1975) Generalization of the gap test for the detection of multivariate outliers. Biometrics 31:93–101

Rohlf FJ (2002) Geometric morphometrics and phylogeny. In: MacLeod N, Forey P (eds) Morphology, shape and phylogeny:175-193

Rohlf F (2005) NTSYSpc, 2.2 edn. Exeter software, Setauket, New York,

Rohlf FJ (2007) Tpsrelw, relative warps analysis, Version 1.45. Stony Brook, NY: Department of Ecology and Evolution, State University of New York at Stony Brook

Rohlf FJ (2015) The tps series of software. Hystrix, the Italian Journal of Mammalogy 26:9–12

Rohlf FJ, Bookstein FL (1987) A comment on shearing as a method for "size correction". Syst Zool 36:356-367

Rohlf FJ, Corti M (2000) Use of two-block partial least-squares to study covariation in shape. Syst Biol 49:740–753. doi:10.1080/106351500750049806

Rohlf FJ, Marcus LF (1993) A revolution in morphometrics. Trends Ecol Evol 8:129–132

Rohlf FJ, Slice D (1990) Extensions of the Procrustes method for the optimal superimposition of landmarks. Syst Biol 39:40–59

Rüber L, Adams DC (2001) Evolutionary convergence of body shape and trophic morphology in cichlids from Lake Tanganyika. J Evol Biol 14:325–332. doi:10.1046/j.1420-9101.2001.00269.x

Santos-Santos JH, Audenaert L, Verheyen E, Adriaens D (2015) Divergent ontogenies of trophic morphology in two closely related haplochromine cichlids. J Morphol 276:860–871. doi:10.1002/jmor.20385

Schmidt EJ et al (2010) Micro-computed tomography-based phenotypic approaches in embryology: procedural artifacts on assessments of embryonic craniofacial growth and developmen. BMC Dev Biol 10: 18

Schmieder DA, Benítez HA, Borissov IM, Fruciano C (2015) Bat species comparisons based on external morphology: a test of traditional versus geometric morphometric approaches. PLoS One 10: e0127043. doi:10.1371/journal.pone.0127043

Sheets H (2003) IMP-integrated morphometrics package. Department of Physics, Canisius College, Buffalo, NY

Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–428

Sidlauskas B (2008) Continuous and arrested morphological diversification in sister clades of characiform fishes: a phylomorphospace approach. Evolution 62:3135–3156. doi:10.1111/j.1558-5646.2008.00519.x

Simmons LW, Garcia-Gonzalez F (2011) Experimental coevolution of male and female genital morphology. Nat Commun 2:374

Simmons LW, Kotiaho JS (2007) Quantitative genetic correlation between trait and preference supports a sexually selected sperm process. Proc Natl Acad Sci U S A 104:16604–16608

Sinclair C, Hoffmann AA (2003) Monitoring salt stress in grapevines: are measures of plant trait variability useful? J Appl Ecol 40:928–937

Singh N, Harvati K, Hublin J-J, Klingenberg CP (2012) Morphological evolution through integration: a quantitative study of cranial integration in *Homo*, *Pan*, *Gorilla* and *Pongo*. J Hum Evol 62:155–164. doi:10.1016/j.jhevol.2011.11.006

Singleton M (2002) Patterns of cranial shape variation in the Papionini (Primates: Cercopithecinae). J Hum Evol 42:547–578

Smilde AK, Kiers HA, Bijlsma S, Rubingh CM, van Erk MJ (2009) Matrix correlations for high-dimensional data: the modified RV-coefficient. Bioinformatics 25:401–405. doi:10.1093/bioinformatics/btn634

Sokal R, Rohlf F (1995) Biometry: the principles and practice of statistics in biological sciences. WH Free Company, New York

Takahashi KH (2013) Multiple capacitors for natural genetic variation in *Drosophila melanogaster*. Mol Ecol 22:1356–1365

Takahashi KH, Rako L, Takano-Shimizu T, Hoffmann AA, Lee SF (2010) Effects of small Hsp genes on developmental stability and microenvironmental canalization. BMC Evol Biol 10:284

Valentin AE, Penin X, Chanut JP, Sévigny JM, Rohlf FJ (2008) Arching effect on fish body shape in geometric morphometric studies. J Fish Biol 73:623–638. doi:10.1111/j.1095-8649.2008.01961.x

Van Heerwaarden B, Sgrò CM (2011) The effect of developmental temperature on the genetic architecture underlying size and thermal clines in *Drosophila melanogaster* and *D. simulans* from the east coast of Australia. Evolution 65:1048–1067

Verbeke G, Fieuws S, Molenberghs G, Davidian M (2014) The analysis of multivariate longitudinal data: a review. Stat Methods Med Res 23:42–59

Vergara-Solana FJ, García-Rodríguez FJ, De La Cruz-Agüero J (2014) Effect of preservation procedures on the body shape of the golden mojarra, *Diapterus aureolus* (Actinopterygii: Perciformes: Gerreidae), and its repercussions in a taxonomic study. Acta Ichthyologica Piscatoria 44:65

Verhaegen Y, Adriaens D, De Wolf T, Dhert P, Sorgeloos P (2007) Deformities in larval gilthead sea bream (*Sparus aurata*): a qualitative and quantitative analysis using geometric morphometrics. Aquaculture 268:156–168

Viðarsdóttir US, O'Higgins P, Stringer C (2002) A geometric morphometric study of regional differences in the ontogeny of the modern human facial skeleton. J Anat 201:211–229

Viswanathan M (2005) Measurement error and research design. Sage Publishing

von Cramon-Taubadel N, Frazier BC, Lahr MM (2007) The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. Am J Phys Anthropol 134:24–35. doi:10.1002/ajpa.20616

Weisbecker V (2012) Distortion in formalin-fixed brains: using geometric morphometrics to quantify the worst-case scenario in mice. Brain Structure and Function 217:677–685

White TA, Searle JB (2008) Mandible asymmetry and genetic diversity in island populations of the common shrew *Sorex araneus*. J Evol Biol 21:636–641. doi:10.1111/j.1420-9101.2007.01481.x

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bulletin 1:80–83

Wilson LA, Cardoso HF, Humphrey LT (2011) On the reliability of a geometric morphometric approach to sex determination: a blind test of six criteria of the juvenile ilium. Forensic Sci Int 206:35–42

Yezerinac SM, Lougheed SC, Handford P (1992) Measurement error and morphometric studies: statistical power and observer experience. Syst Biol 41:471–482. doi:10.1093/sysbio/41.4.471

Zelditch ML, Swiderski DL, Sheets HD (2004) Geometric morphometrics for biologists: a primer. Academic Press