Biological Journal of the Linnean Society, 2020, **XX**, 1–13. With 3 figures.

Testing repeatability, measurement error and species differentiation when using geometric morphometrics on complex shapes: a case study of Patagonian lizards of the genus *Liolaemus* (Squamata: Liolaemini)

JUAN VRDOLJAK[1,*], KEVIN IMANOL SANCHEZ[1], ROBERTO ARREOLA-RAMOS[1], EMILCE GUADALUPE DIAZ HUESA[2], ALEJANDRO VILLAGRA[3], LUCIANO JAVIER AVILA[1] and MARIANA MORANDO[1]

[1]*Instituto Patagónico para el Estudio de los Ecosistemas Continentales, Consejo Nacional de Investigaciones Científicas y Técnicas (IPEEC-CONICET), Boulevard Almirante Brown 2915, U9120ACD, Puerto Madryn, Chubut, Argentina*
[2]*Instituto de Diversidad y Evolución Austral, Consejo Nacional de Investigaciones Científicas y Técnicas (IDEAUS-CONICET), Boulevard Almirante Brown 2915, U9120ACD, Puerto Madryn, Chubut, Argentina*
[3]*Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Bv. Almte Brown 3051, U9120ACD Puerto Madryn, Chubut, Argentina*

Received 4 February 2020; revised 4 May 2020; accepted for publication 11 May 2020

The repeatability of findings is the key factor behind scientific reliability, and the failure to reproduce scientific findings has been termed the 'replication crisis'. Geometric morphometrics is an established tool in evolutionary biology. However, different operators (and/or different methods) could act as large sources of variation in the data obtained. Here, we investigated inter-operator error in geometric morphometric protocols on complex shapes of *Liolaemus* lizards, as well as measurement error in three taxa varying in their difficulty of digitalization. We also examined the potential for these protocols to discriminate among complex shapes in closely related species. We found a wide range of inter-operator error, contributing between 19.5% and 60% to the total variation. Moreover, measurement error increased with the complexity of the quantified shape. All protocols were able to discriminate between species, but the use of more landmarks did not imply better performance. We present evidence that complex shapes reduce repeatability, highlighting the need to explore different sources of variation that could lead to such low repeatability. Lastly, we suggest some recommendations to improve the repeatability and reliability of geometric morphometrics results.

ADDITIONAL KEYWORDS: complex shapes – geometric morphometrics – inter-operator error – measurement error – replication crisis.

INTRODUCTION

The so-called 'replication crisis' is a hot topic in specialized journals of statistics and psychology (Loken & Gelman, 2017; Fidler & Wilcox, 2018), but is a relatively new field for biologists (Fraser et al., 2018). The 'replication crisis', in broad sense, is associated with the failure to reproduce the results of previous studies. However, most researchers rarely attempt to replicate results, possibly because—motivated by the 'publish or perish' dogma—most scientific journals have explicit policies against publishing replication studies (Schmidt, 2009). Non-repeatability leads to lack of reliability in scientific findings because it compromises our confidence in the generality of scientific theories.

Publication bias, questionable research practices (QRPs) and over-confidence in null hypothesis

*Corresponding author. E-mail: juan.vrdoljak@gmail.com

significance tests (NHSTs) are known to affect repeatability without threatening the generality of scientific facts (Fidler & Wilcox, 2018; Shrout & Rodgers, 2018). NHSTs and *P*-value thresholds are the current paradigms for research, publication and discovery in biological and social sciences (Ioannidis, 2018; Dushoff *et al.*, 2019). This set of ideas can lead to mistakes, and possibly drive publication bias and QRPs. Major mistakes include: the dichotomization of results into 'significant' and 'non-significant'; the focus only on significant results even when they are irrelevant (e.g. descriptive statistics); the omission of other evidence such as magnitude of effect; several misinterpretations of *P*-values; and the implausibility of the null hypothesis when the effects are small, because the possibilities of systematic bias and variation due to highly variable measurements could result in similar small effects (Amrhein *et al.*, 2019; McShane *et al.*, 2019).

Measurement error (ME) is uncontrolled variation that could aggravate the replication crisis (Loken & Gelman, 2017). Given its random nature, ME is frequently associated with statistical 'noise' around true values, and thus if an effect is found in a noisy statistical environment, then one might infer that the actual effect is very strong (Brakenhoff *et al.*, 2018a). However, in these cases, effect size estimates can be exaggerated and the outcomes can be biased by poor measurements (Loken & Gelman, 2017; Brakenhoff *et al.*, 2018b).

Geometric morphometrics is a body of methods based on complex mathematics to identify, quantify and describe shapes independently of size. In brief, three steps are necessary to obtain data from a morphometric protocol: (1) digitize images of the objects by photography or 3D scanning; (2) place landmarks, semilandmarks or outlines in informative positions (Bookstein, 1991); and (iii) superimpose these points (Dryden & Mardia, 1998). Many studies have soiught to help geometric morphometric operators improve their analyses (e.g. Kaliontzopoulou, 2011; Viscosi & Cardini, 2011; Cardini & Loy, 2013; Cardini, 2016), and we have identified 18 400 published papers that made use of these methods (based on a brief search in Google scholar). However, little is known about the sources of variation that could generate spurious results (Arnqvist & Martensson, 1998; Rohlf, 2003; Cardini & Elton, 2007; Fruciano, 2016).

Fruciano (2016) reviewed common sources of error in geometric morphometric studies with an emphasis on ME. He included different methods to assess ME, and concluded that researchers have to consider factors that compromise accurate measurement, such as effort invested in digitization of images (Cardini & Elton, 2007), specimen quality (Fruciano *et al.*, 2020) and maintenance of co-planarity in 3D structure (Cardini, 2014). In complex shapes, highly variable morphological positions could lead to over-inflation of ME due to low-accuracy landmarks (Cummaudo *et al.*, 2013) or high landmarking bias (von Cramon-Taubadel *et al.*, 2007); good treatment of these positions could be the cornerstone to increase repeatability in geometric morphometrics (Fagertun *et al.*, 2014).

The term 'complex shapes' refers to certain configurations for which the placement of landmarks is not trivial, i.e. the informative morphological positions vary drastically among specimens, or the homology among points is difficult to establish (e.g. Toma *et al.* 2009; Fagertun *et al.*, 2014; Campomanes-Álvarez *et al.*, 2015). In this regard, Bookstein (1991) described type I, II and III landmarks that scale from most to least clarity of the anatomical point. Several authors have reported that type III landmarks are clearly associated with high ME (von Cramon-Taubadel *et al.*, 2007; Barbeito-Andrés *et al.*, 2012). However, the analytical procedure in many studies is the same for each of these types of landmarks (see, for example, von Cramon-Taubadel *et al.*, 2007; Ross *et al.*, 2008; Barbeito-Andres *et al.*, 2012), and while this distinction has a strong subjective or arbitrary character (Slice, 2005), many articles fail to make this point. Alternatively, semilandmarks curves and contours are more suitable to describe complex shapes because the superimposition of these types of descriptors is more relaxed than landmark superimposition (Zelditch *et al.*, 2012; Gunz & Mitteroecker, 2013). However, no measurement technique is error-free.

In lizards, dorsal and lateral views of the head are the typical structures analysed using geometric morphometrics, as for example in studies of sexual dimorphism (Kaliontzopoulou *et al.*, 2007), species delimitation (Leaché *et al.*, 2009; Esquerré *et al.*, 2019) and ecological relationships (Kaliontzopoulou *et al.*, 2010). In particular, Patagonian lizards of the genus *Liolaemus*, a highly diverse group with at least 277 species (Reptile Database, May 2020), have different head-scale configurations, including different colour patterns and missing scales between and within species, that make digitization difficult. In this context, *Liolaemus* species are good models to study the relationship among complex shapes, ME and repeatability.

Here, we analyse how digitization of complex shapes influences the repeatability of results, ME and species differentiation using geometric morphometrics. We focus on the following objectives: (1) quantifying spurious (herein termed 'extrinsic') sources of variation, using different geometric morphometric methods to quantify complex shapes such as the dorsal view of lizards' head, and discuss the principal implications for biological inference; (2) evaluating how ME extends to other taxa with well-established

landmark points (a fly and a plant), and compare with lizard head shapes; (3) assessing the potential of each geometric morphometric method to describe and discriminate among closely related species with complex shapes; and (4) integrating the information collected in conjunction with the digitalization effort (measured as processing time) for each of the methods. Given that no technique is error-free, appropriate digitalization effort can be another important factor within the operational criteria workflow. Finally, we advocate the use of a clear and solid statistical framework to address these issues.

MATERIALS AND METHODS

Experimental designs, sample/data collection and methodological approach

The following three approaches were used address the three study objectives. First, we analysed different inter-operator factors that may affect the repeatability of results. To address this, we first made mirror images of 25 photographs of male *Liolaemus elongatus,* and then five operators landmarked/outlined each side of each image twice; our goal here was to assess among-operator design. Mirrored images were added in order to incorporate one extra source of natural variation: the side effect and the interaction with the other sources of variation. Second, we then estimated ME across three very different taxa: the lizard *L. elongatus*, the fly *Drosophila buzzatii* (Vrdoljak *et al*., 2019) and the grape *Vitis riparia* (data available at https://dataverse.harvard.edu/dataverse/VitisLeafVariation; Klein *et al*., 2017). In this case, each of 25 photographs was landmarked/outlined in quadruplicate and analysed for each taxon separately; this second goal was to assess across-taxon design. We do not consider the digitized structures of these last two taxa as complex shapes because they have very simple, high-quality tested and well-established sites to place landmarks (Klingenberg, 2009; Klein *et al*., 2017). Third, we assessed the potential of each morphometric configuration (described below as protocols) to differentiate among similar complex shapes. Here, we sampled 25 males of three closely related *Liolaemus* lizards (*L. elongatus*, *L. shitan* and *L. choique*; Medina *et al*., 2018) with several morphological similarities and then landmarked/outlined; our goal here was to assess related-species design (details of specimen's voucher numbers, collection locality and other data are given in the Supporting Information, Appendix S1).

We took dorsal photographs of the head of each specimen using a Canon 1000D camera mounted on a fixed tripod. For flies, we removed the left wing, mounted them on slides with DPX and photographed them at 40× magnification using a digital camera attached to a microscope (Nikon E200). To characterize the shape, we placed landmarks in four different configurations using TpsDIG 2.31 (Rohlf, 2015). Shape variation was estimated using a generalized Procrustes analysis (Zelditch *et al*., 2012) with sliding of semilandmarks performed by minimizing the Procrustes distance. We then used principal component analyses to summarize the shape information in uncorrelated form. We employed another approach to quantify shape variation based on elliptical Fourier descriptors (Kuhl & Giardina, 1982). Outlines from digital images were used to obtain Fourier coefficients for a polynomial function of the 9th degree for lizards and flies and the 12th degree for plants, normalized for size, rotation and starting point (directly on the matrix of coefficients). We then built a variance–covariance matrix that was used as input in a principal component analysis. Morphometric analyses were performed with the packages Momocs (Bonhomme *et al*., 2014) and geomorph (Adams *et al*., 2018), implemented in the R statistical software (R Core Team, 2019).

For each of the three designs, we developed five morphometric protocols. First, we used two landmark-only protocols with six and ten landmarks for lizards and leaves and ten and 15 landmarks for flies (P-L and F-L protocols for partial-landmark and full-landmark, respectively). We then used two semilandmark protocols (Gunz & Mitteroecker, 2013), both starting from the same P-L configuration, with one and two curves (P-S and F-S protocols for partial-semilandmark and full-semilandmark, respectively). Note that the partial protocols (P-L and P-S) are nested within the full protocols (F-L and F-S). Lastly, we used a contour protocol based on Fourier descriptors, a novelty in herpetological research.

We examined one-side morphologies except in the cases of lizard contours in the across-taxon and related-species designs, where the whole pineal scale was used, while in the among-operator design, half of a pineal scale and whole parietal scale were outlined (see Supporting Information, Appendix S2 for details of contours, and landmark and semilandmark configurations; also see Carreira *et al*., 2008; Klingenberg, 2009; Klein *et al*., 2017).

To implement the among-operators design, the order of the five protocols was randomly selected for each operator. Also, in order to represent the greatest possible variation in ME, operators with different degree of knowledge about morphometric techniques were chosen. For the remaining two designs (across-taxa and related-species), only one of the operators performed all five protocols. Finally, we randomly choose ten specimens from each taxon to compute the data-gathering processing time for each protocol.

Models and statistical analyses

For among-operator design, we used hierarchical models to estimate seven variance effects: (1) specimen variation, (2) operator variation, (3) ME (i.e. intra-operator error), (4) specimen*operator, (5) specimen*side, (6) operator*side and (7) specimen*operator*side variation. Specimen and specimen*side variation (the latter known as fluctuation asymmetry) are two intrinsically natural sources of variation (intrinsic variation), while the remaining five variance effects are influenced by operator error, biased measurement and the consistency of these (extrinsic variation, composed of ME and inter-operator error). As a result of this model, we evaluated different factors affecting the repeatability (operator and operator interaction effects), and thus the reliability of the measurement technique in a relative manner with intrinsic variation. For across-taxon design, to accurately evaluate ME of the five protocols among taxa, we used hierarchical models that included only specimen variance and ME (i.e. each taxon was analysed separately). In this sense, each protocol was applied to grape leaves, fly wings and lizard heads by one operator. For related-species design, we analysed the effects of morphological differences among species (details of the models are given in Supporting Information, Appendix S3).

Because the placement of more points implies more processing time, we used a linear regression to assess the relationship between processing time and centroid size. Centroid size is more suitable to explain processing time than the simple sum of landmark and semilandmark points, because by definition it is a good proxy of the number of points (the square root of the sum of the squared distances of each landmark to the centroid configuration; Dryden & Mardia, 1998), and given that the scale of the images was fixed for each protocol, the operators would probably spend more time in mouse displacement using large than small sizes with an equal number of points.

For all the designs we employed the first principal component (PC) axes that explain at least 60% of the total variation to perform statistical analyses (reason 3 below explains the selected percentage). The single exception was for the related-species design where we explored the minimal number of PC axes needed to differentiate among species. We also investigated the remaining PC axes in search of morphological changes associated with our explanatory effects, but we found no such changes. Some authors criticize the use of single PC axes to perform statistical analyses, rather than the entire dataset, for Procrustes ANOVA (Fruciano *et al.*, 2020). However, in the context of this study, we made this decision for three important interrelated reasons: (1) to incorporate uncorrelated variables in a Bayesian statistical framework (see below); (2) to analyse which morphological changes are more related to measurement or inter-operator error; and (3)to explain more than half of the total variation, reducing dimensionality (from eight to three or from 48 to four axes), thus contributing to the avoidance of overfitting. We know that our methods are not the most common, but they are no less valid and rest on a solid statistical framework for analysis of random variation (Ellison, 2004; Zimova *et al.*, 2016).

The models were fitted within a Bayesian framework that eased implementation of variance components and its uncertainty. Posterior distributions of parameters were estimated using three independent Markov chain Monte Carlo (MCMC) runs for 100 000 iterations and 20% burn-in, each implemented in JAGS 4.3.0 (Plummer, 2003) using the R packages jagsUI (Kellner, 2018) and rjags (Plummer & Stukalov, 2018). The observations were centred and standardized to reduce autocorrelation of chains (Kruschke, 2014). Convergence was assessed using Gelman and Rubin $\hat{R}$ statistics (if converged, $\hat{R} < 1.1$; Gelman & Rubin, 1992), and by visual inspection of trace plots (all fitted data converged successfully). We used weakly informative prior distributions to include small amounts of information on each parameter and hyper-parameter, and to avoid meaningless values (Gelman *et al.*, 2013; McElreath, 2018). Finally, despite using the same analytical procedure, we denoted differences between two samples of the response variable as the standardized difference, whereas we reserved effect size to the posterior distribution of the standardized differences (Maxwell *et al.*, 2015; analysed according to Hedges & Olkin, 1985) and reported the mean of the posterior distribution (hereafter posterior mean) and high posterior density interval (HPD; Hyndman, 1996) using the R package coda (Plummer *et al.*, 2006).

RESULTS

Among-operator design

We found extrinsic variation in all five protocols when considering the first PC that explains the greatest morphological variation (Fig. 1). The contour protocol had the best performance in terms of highest sources of intrinsic variation and smallest sources of extrinsic variation, whereas all other protocols showed a trade-off among different sources of variation. In this sense, both partial protocols showed the highest levels of operator variation, and explained very similar levels of intrinsic variation with a greater uncertainty in the P-S protocol. Full protocols explained more intrinsic variation than partial protocols, but the F-S protocol captured more extrinsic variation than the F-L protocol. Moreover, we found high levels of operator*specimen variation in the F-L protocol

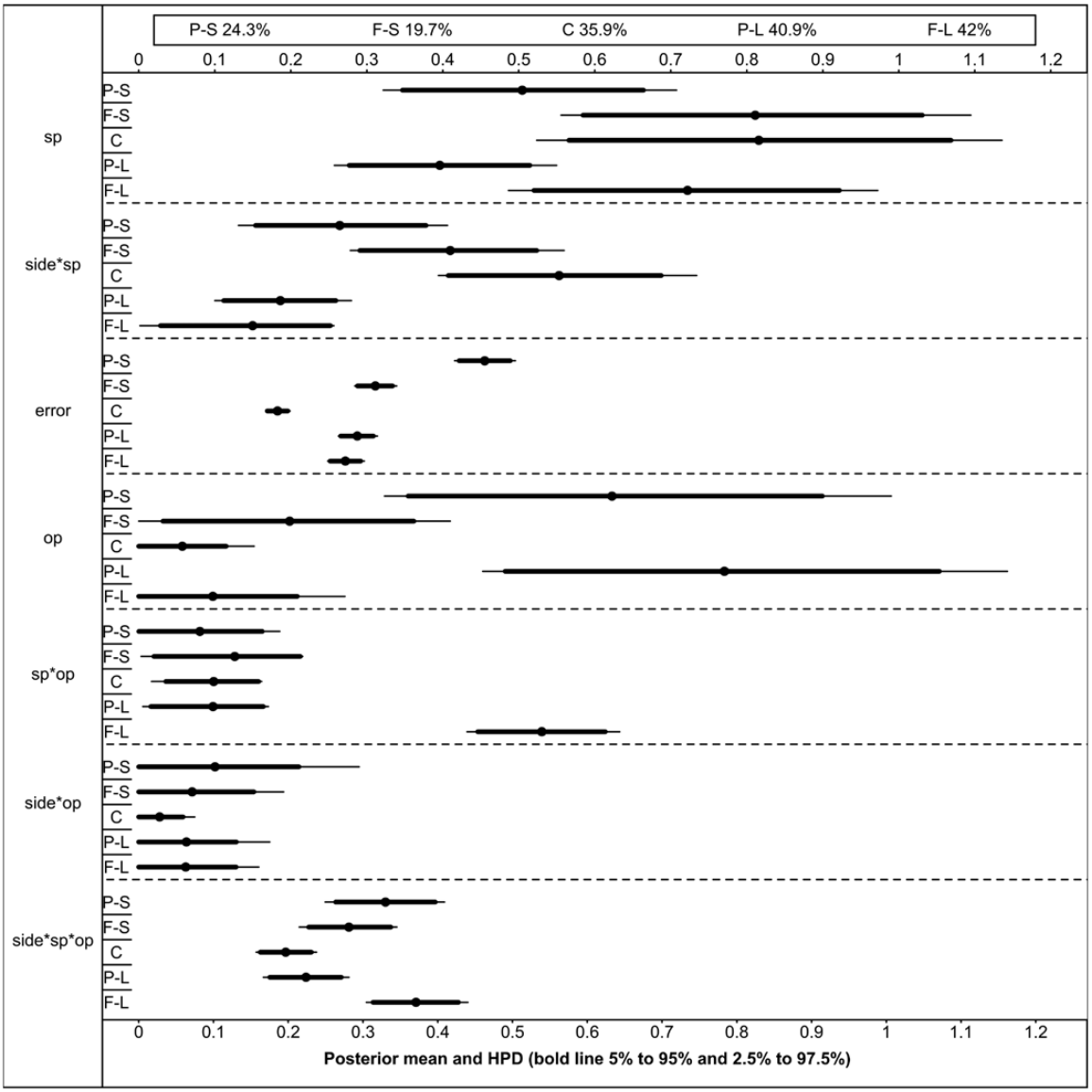

Figure 1. Posterior mean and highest posterior density interval (HPD) of 90% (bold line) and 95% (thin line) for each source of variation of the first principal component: sp, specimen; side, side; op, operator; error, measurement error. *Interaction between sources of variation. Protocols: P-S, partial-semilandmark; F-S, full-semilandmarks; C, contour; P-L, partial-landmark; F-L, full-landmark. Box at the top: percentage of variation explained by the first principal component for each protocol.

due to a systematically erroneous placement of two landmarks by some operators on some specimens (Supporting Information, Fig. S1a), thus highlighting the existence of measurement bias.

Inter-operator error was always greater than ME for all protocols (see PC1 in Fig. 2). More than half of the total variation was explained by inter-operator error in the P-L protocol (56.7%), closely followed by F-L and P-S with almost half of the total variation (48% and 47.7%, respectively). In contrast, inter-operator error contributed to 30.4% of the mean variation to the whole model in F-S, and remarkably less in the contour protocol (19.6%). Nevertheless, ME explained more than negligible variation in all protocols. P-S expressed noticeably greater variation of ME (19.5%) than the contour protocol (9.5%), whereas P-L, F-L, and F-S showed similar variation (14.4, 12.4 and 14.2%, respectively).

Given all PCs analysed, the contour protocol maintained lowest mean values of extrinsic variation (Fig. 2). More than 60% of the total variation was explained by extrinsic factors in the first three PCs

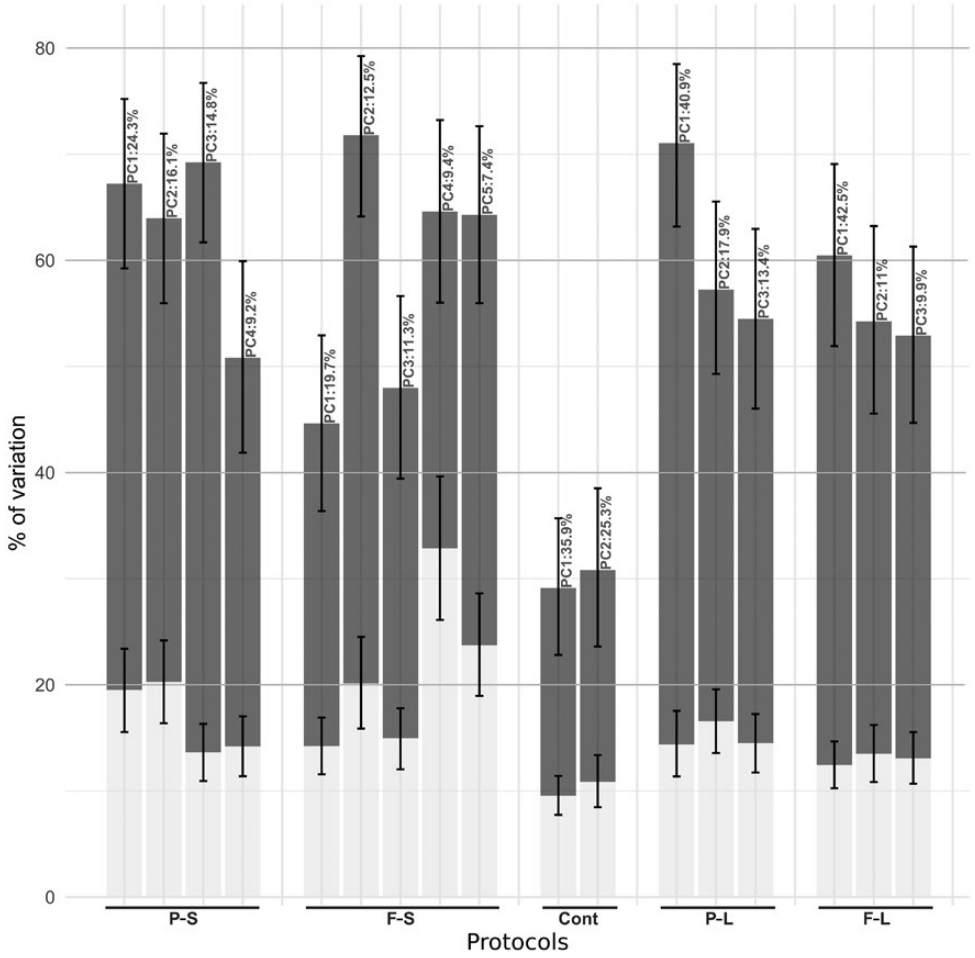

Figure 2. Percentage of variation explained by each factor of extrinsic variation: measurement error (light grey bars) and inter-operator error (dark grey bars). Values represent posterior means and the error bars indicate 95% highest posterior density intervals. Protocols: P-S, partial-semilandmark; F-S, full-semilandmarks; Cont, contour; P-L, partial-landmark; F-L, full-landmark. The percentage of variance explained by each principal component analysed is given on top of the bars. ggplot2 (Wickham, 2016) was used to develop this figure.

of the P-S protocol. In the F-S protocol, the first and third PCs showed smaller extrinsic variation than the others. Both P-L and F-L protocols improved the levels of extrinsic variation through the PCs, but high levels of operator*specimen variation were found in PC2 of the former protocol (Supporting Information, Fig. S2), due to a consistent measurement bias on some specimens (Fig S1b).

Across-taxon design

This design exposed at least three clear patterns (Fig. 3). First and most conspicuously, ME variation was highest in lizards, followed by flies and then leaves (the averages of ME variation weighted by morphological variation explained by each PC were 28.2, 9.8 and 2.6%, respectively). In particular, the protocol with greatest ME contributed 57.3, 19.8 and 7.6% to the total variation, while the protocol with smallest ME contributed 15, 11.3% and 1.5% to the total variation in lizards, flies and leaves, respectively. In other words, a wide range of ME was dependent on both taxon and protocol, indicating that some protocols are more suitable for some taxa than others.

Second, as we expected, processing time was longest in protocols with more points (understanding points to be the number of landmarks plus semilandmarks), i.e. the processing time for all taxa follows from longest to shortest: F-S, P-S, F-L and P-L. Indeed, we found a positive correlation between processing time and specimen size across protocols (excluding the contour protocol for the analysis; Supporting Information, Fig. S3).

Third and more interestingly, the contour protocol showed an independent pattern of processing time with respect to the other protocols. In this sense, contour

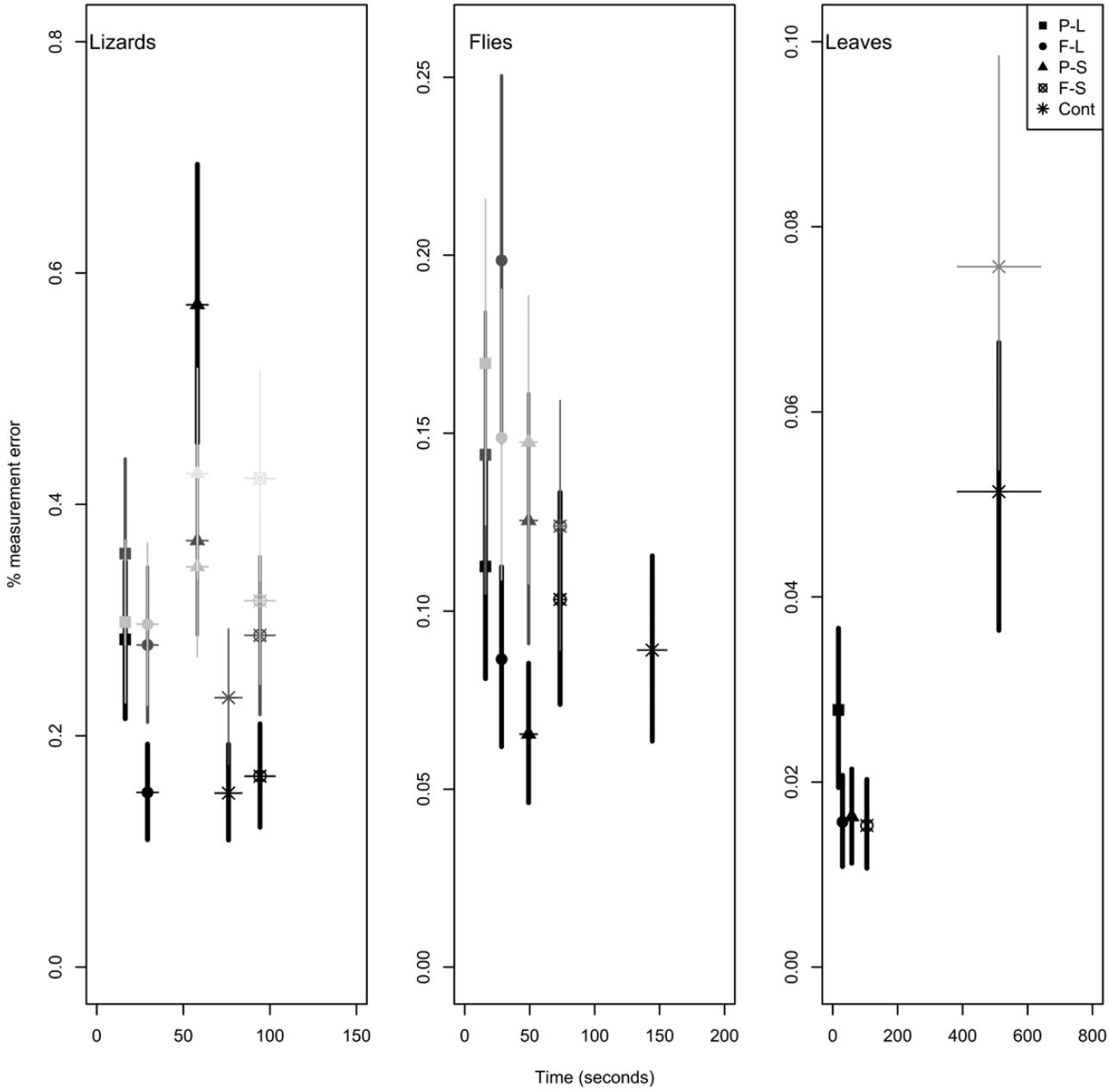

Figure 3. Percentage of measurement error vs. processing time for each protocol and each taxon. Values and error bars of measurement error represent posterior mean and 95% highest posterior density interval, while values and error bars of processing time represent mean and standard deviation. Protocols: P-S, partial-semilandmark; F-S, full-semilandmarks; Cont, contour; P-L, partial-landmark; F-L, full-landmark. The principal components are represented by a grey shade scale, where darker points and lines correspond to the first PC and the subsequent PCs are increasingly clearer.

processing time was shorter than for the F-S protocol in lizards but longer than for the other protocols in flies and leaves. Moreover, the difference between the contour and F-S protocols in mean time elapsed for each digitized image was 0.48, 1.58 and 4.52 min (but 2.12, 7.89 and 4.29 of standardized differences) for lizards, flies and leaves, respectively.

Related-species design

All protocols discriminated among species more or less clearly (Table 1). In this sense, both semilandmark protocols showed very clear differences among species, with a slightly better performance in the partial protocol. However, the morphological information explained by these protocols was redundant

Table 1. Mean values, highest posterior density interval (HPD) of 95 and 90% of the effect size distribution resulting from species comparisons among *Liolaemus*. Note that the distinction between species becomes clearer as the intervals move away from 0. Principal components (PC) shown in the table are those necessary and sufficient for a clear differentiation between species (PC2 of the contour protocol was not sufficient for species differentiation)

		Effect size											
		L. elongatus – *L. shitan*				*L. elongatus* – *L. choique*				*L. shitan* – *L. choique*			
Protocol		HPD 2.5	HPD 5	Mean	HPD 95	HPD 97.5	HPD 2.5	HPD 5	Mean	HPD 95	HPD 97.5	HPD 2.5	HPD 5	Mean	HPD 95	HPD 97.5
PS	PC1	−1.70	−1.61	−1.13	−0.65	−0.57	0.09	0.17	0.63	1.08	1.18	−2.39	−2.29	−1.76	−1.24	−1.14
	PC2	−1.11	−1.02	−0.56	−0.09	−0.01	−2.86	−2.77	−2.20	−1.64	−1.53	1.02	1.13	1.64	2.16	2.25
FS	PC1	−1.54	−1.45	−0.98	−0.51	−0.42	0.07	0.17	0.61	1.08	1.15	−2.19	−2.10	−1.59	−0.97	−1.08
	PC2	−0.92	−0.83	−0.38	0.07	0.16	−2.11	−2.03	−1.51	−1.02	−0.91	0.55	0.65	1.13	1.62	1.71
C	PC1	−0.84	−0.75	−0.30	0.15	0.23	−1.96	−1.86	−1.36	−0.87	−0.78	0.50	0.58	1.06	1.54	1.65
	PC3	−1.65	−1.54	−1.07	−0.59	−0.52	−0.85	−0.77	−0.32	0.13	0.22	−1.31	−1.22	−0.75	−0.29	−0.19
PL	PC1	−1.73	−1.64	−1.16	−0.68	−0.59	−0.34	−0.26	0.18	0.63	0.71	−1.94	−1.83	−1.34	−0.83	−0.76
	PC2	−0.43	−0.33	0.11	0.57	0.65	−2.45	−2.34	−1.81	−1.29	−1.19	1.29	1.39	1.92	2.45	2.56
FL	PC1	−1.05	−0.97	−0.51	−0.06	0.03	0.48	0.58	1.05	1.52	1.61	−2.08	−2.16	−1.56	−1.06	−0.95
	PC2	0.49	0.58	1.05	1.53	1.62	0.42	0.50	0.98	1.45	1.55	−0.47	−0.38	0.08	0.53	0.62

Protocols: PS, partial-semilandmark; FS, full-semilandmark; C, contour; PL, partial-landmark; FL, full-landmark.

(Supporting Information, Figs S4, S5), and more time was necessary for the full protocol. The landmark protocols also showed very clear differences among species, but each protocol explained dissimilar morphological information (Figs S4, S5). It is critical to note that the main differences between species for PC1 of the F-L protocol were due to changes in the same conflictive landmarks that strongly biased the among-operators design (Fig. S1a). In contrast, among-species differences resolved by the contour protocol were slightly less clear than with the other methods, and it was necessary to look beyond the second PC axis.

DISCUSSION

We analysed several inter-operator error and ME factors that could influence the result produced in digitizing complex shapes with geometric morphometric methods, and also the potential of each protocol to discriminate among species. We found disturbing levels of extrinsic variation across all our analyses, highlighting the need for an in-depth inquiry into the repercussions of complex shapes on morphometric studies.

In recent years, different factors that affect the repeatability of geometric morphometric studies have been addressed. Fagertun *et al.* (2014) reported that operator variation (i.e. inter-operator error) was associated with particular landmarks (also reported by Cummaudo *et al.*, 2013; Campomanes-Álvarez *et al.*, 2015; Robinson & Terhune, 2017), and that such variation was similar to variation among individuals. However, what they called the error term (not the ME by model construction) was twice that from operator variation, while Robinson & Terhune (2017), Fruciano *et al.* (2017) and Shearer *et al.* (2017) found that the highest variation was attributable to inter-operator error. In agreement with these last three studies, we show that the inter-operator error was always greater than ME and that, in most cases, the total extrinsic variation was greater than intrinsic variation. Moreover, inter-operator error accounted for at least 19% of the total variation, and this increased to almost 60% in the least-repeatable protocol. A clear operational conclusion should be that digitizing on original images by one operator rather than utilizing datasets collected by multiple operators is preferred (Fruciano *et al.*, 2017; Shearer *et al.*, 2017).

Bias could be defined as systematic error. Unlike any random error, measurement bias could lead to mean differences between groups when there is none. However, Fruciano *et al.* (2017) showed that bias accounts for a small proportion of variation, and becomes significant when highly variable landmarks are removed. In complex shapes, we found biased

measures in two different protocols: F-L and P-L. Variation due to these biased measures were captured by PC1 (42.5%) and PC2 (16.1%), respectively. Curiously, *L. choique* was differentiated from the other species mainly by morphological differences in the two conflictive landmarks involved in the measurement bias of the F-L protocol. If the operator's experience can influence the degree of measurement bias (Shearer *et al*., 2017), then bias in the F-L protocol could become very problematic. More generally, we have shown that configurations with a high risk of measurement bias could exaggerate the true morphological differences, further aggravating the replication crisis.

ME is a widely studied issue in the scientific literature and a concern for a large percentage of publications (Brakenhoff *et al*., 2018b). Some authors predict that with technological advances, ME would probably become a less frequent problem but the large amount of data available that has been obtained by other researchers could lead to additional sources of variation (Ioannidis, 2005; Fruciano *et al*., 2017; Marcy *et al*., 2018). Our findings indicate that there is a relationship between complex shapes and ME. With respect to this, we described the lizard head morphology as a complex shape because informative positions vary among species due to broken or missing scales, or different muscle development and/or colours, which can hinder the digitization process. In contrast, morphological positions are very clear in fly wings and grape leaves, and this is reflected in an ME approximately three and ten times smaller, respectively, compared to lizards.

Another key factor in deciding how to digitize samples is processing time. Despite the fact that this factor was similar for each protocol and taxon, the contour protocol showed a distinct pattern: processing time was positively correlated with size and the 'difficulty' of digitizing. In this sense, the effect of size is expressed in the differentiation between flies and lizards, where the contour of the former occupied almost the entire image while the contour of the latter occupied a part of the image. On the other hand, the fly's wing is a more or less round appendage, and clearly distinguishes itself from the innumerable grape leaf peaks (Supporting Information, Appendix S2). With respect to this, we described the difference in the processing time due to difficulties in digitizing.

Geometric morphometrics are a widely used, well-accepted and practical set of tools to quantify morphological phenotypes (Viscosi, 2015), fluctuating asymmetry (Klingenberg, 2015), acoustic signals (MacLeod *et al*., 2013) and useful forensic patterns (Kimmerle *et al*., 2008), among others. Selecting a configuration that faithfully represents the shape analysed is an obvious but not a trivial notion. Here, we studied the potential of each protocol to discriminate among complex shapes of different species, and found that more landmark points does not necessarily explain more shape information. Indeed, the P-S protocol was better than the F-S protocol in discriminating among species (Table 1). The F-L protocol also differentiates species with high performance, albeit detracted by measurement bias (Supporting Information, Fig. S1a). Although the contour protocol only expressed differences in the pineal scale, species discrimination was successful, indicating that this method deserves to be studied in depth given its high performance here. Nevertheless, it is of note that we only used some of the PC axes for the models, not accounting for all the morphological variation. As a consequence, the components with largest variance accounted for similar shape variation, and this could be the reason why we observed no particular benefit in increasing the number of landmarks. Despite this, we found that some protocols performed better than others, highlighting that each protocol captured morphological changes that are not exactly the same. On the other hand, the analytical procedures typically used to discriminate among species (so-called 'best practices') use all morphological information (e.g. multivariate discriminant or multivariate tests of different means). In this study, the use of only the first two PCs (almost 50% of the total variation in each protocol) was enough to discriminate among species.

Certain recommendations should be noted. First and foremost, although several operators may be involved either to reduce processing time or because of their greater expertise in certain procedures, only one person should perform each stage of the data collection. The greatest variation we found was due to five different operators placing landmarks or outlining images. If five different operators had photographed specimens, for instance, then the extrinsic variation would have been greater (see a similar example in Robinson & Terhune, 2017). Second, bibliographical searches to select homologous positions for landmark placement are good practice to improve replication. However, some landmark configurations from certain publications may not be useful because they were developed to test other hypotheses, or because character homology was not assessed. In this sense, pioneer morphometric studies need to be careful and identify the most stable landmark configurations to test particular hypotheses in pilot tests. Third, geometric morphometric researchers should quantify ME and, if possible, include it in the model. There are many ways to estimate ME in geometric morphometrics (Fruciano, 2016), but most of them entail extra effort such as multiple digitizations, learning about novel methods and good data management. With currently implemented geometric morphometric studies, most researchers focus their efforts on expanding their

dataset, rather than worrying about sources of error. Fourth, researchers should select a method that has a high quality to processing time ratio. Sometimes, long processing times can enhance ME (Engelkes *et al*., 2019). Fifth, complex shapes do not necessarily need more landmarks points. We have shown that there are not many differences between the 'resolution' of partial and full protocols, but the latter require considerably more processing time. Sixth, be careful (or be Bayesian) when the underlying effect is small and sampling error is large, because experiments that achieve statistical significance must have exaggerated effect sizes (Type M error, exaggerated magnitude), and are likely to have the wrong sign (Type S error; Gelman & Carlin, 2014).

Overall, our results call on researchers to reflect on the implications of their conclusions and what these imply, such as in the widespread discourse of scientific truth and scientific unity (Dupré, 1995). Moreover, this problem could become worse if we combine the low reliability of data collection (such as from poorly analysed complex shapes), and some of the current proclamations about the role of subjectivity in the scientist's tasks, as for example the criticism of Garnett & Christidis (2017) on the arbitrariness of taxonomy (but see Raposo *et al*., 2017; Conix, 2019). We invite other researchers to repeat this kind of study in their disciplines, techniques or taxa to understand the depth of the crisis of replication in natural sciences.

ACKNOWLEDGEMENTS

We would like to thank two anonymous reviewers for very constructive comments that helped to improve a previous version of the manuscript. We also thank Jack Sites for English revisions, which helped us to improve the manuscript. We thank the fauna authorities from Río Negro, Neuquén, Mendoza and Chubut Provinces for collection permits. Financial support was provided by the following grants: Agencia Nacional de Promoción Científica y Tecnológica - Fondo Nacional para la investigación Científica y Tecnológica (ANPCYT-FONCYT) 1397/2011 (L.J.A.); ANPCYT-FONCYT 1252/2015, Proyecto de Investigación Plurianual - Consejo Nacional de Investigaciones Científicas y Técnicas (PIP-CONICET) 0336/13 (M.M.), and a doctoral fellowship (J.V.) from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

J.V. designed the experiment, performed the analyses and drafted the manuscript. J.V., K.I.S., R.A.R., E.D.H. and A.V. generated the dataset. J.V., K.I.S. and M.M. corrected and subsequently rewrote the manuscript. L.J.A. and J.V. contributed to field sampling. All authors gave final approval for publication.

REFERENCES

Adams DC, **Collyer ML**, **Kaliontzopoulou A. 2018.** *Geomorph: software for geometric morphometric analyses*. R package version 3.0.6. Available at: https://cran.r-project.org/package=geomorph.

Amrhein V, **Greenland S**, **McShane B. 2019.** Scientists rise up against statistical significance. *Nature* **567:** 305–307.

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

Barbeito-Andrés J, **Anzelmo M**, **Ventrice F**, **Sardi ML. 2012.** Measurement error of 3D cranial landmarks of an ontogenetic sample using computed tomography. *Journal of Oral Biology and Craniofacial Research* **2:** 77–82.

Bonhomme V, **Picq S**, **Gaucherel C**, **Claude J. 2014.** Momocs: outline analysis using R. *Journal of Statistical Software* **56:** 1–24.

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

Brakenhoff TB, **Mitroiu M**, **Keogh RH**, **Moons KG**, **Groenwold RH**, **van Smeden M. 2018a.** Measurement error is often neglected in medical literature: a systematic review. *Journal of Clinical Epidemiology* **98:** 89–97.

Brakenhoff TB, **Van Smeden M**, **Visseren FL**, **Groenwold RH. 2018b.** Random measurement error: why worry? An example of cardiovascular risk factors. *PLoS One* **13:** e0192298.

Campomanes-Álvarez BR, **Ibáñez O**, **Navarro F**, **Alemán I**, **Cordón O**, **Damas S. 2015.** Dispersion assessment in the location of facial landmarks on photographs. *The International Journal of Legal Medicine* **129:** 227–236.

Cardini AL. 2014. Missing the third dimension in geometric morphometrics: how to assess if 2D images really are a good proxy for 3D structures? *Hystrix* **25:** 73–81.

Cardini A. 2016. Lost in the other half: improving accuracy in geometric morphometric analyses of one side of bilaterally symmetric structures. *Systematic Biology* **65:** 1096–1106.

Cardini A, **Elton S. 2007.** Sample size and sampling error in geometric morphometric studies of size and shape. *Zoomorphology* **126:** 121–134.

Cardini A, **Loy A. 2013.** Virtual morphology and evolutionary morphometrics in the new millenium. *Hystrix* **24:** 1–140.

Carreira VP, **Soto IM**, **Fanara JJ**, **Hasson E. 2008.** A study of wing morphology and fluctuating asymmetry in interspecific hybrids between *Drosophila buzzatii* and *D. koepferae*. *Genetica* **133:** 1–11.

Conix S. 2019. In defence of taxonomic governance. *Organisms Diversity & Evolution* **19:** 87–97.

von Cramon-Taubadel N, **Frazier BC**, **Lahr MM. 2007.** The problem of assessing landmark error in geometric

morphometrics: theory, methods, and modifications. *American Journal of Physical Anthropology* **134:** 24–35.

Cummaudo M, **Guerzoni M**, **Marasciuolo L**, **Gibelli D**, **Cigada A**, **Obertovà Z**, **Ratnayake M**, **Poppa P**, **Gabriel P**, **Ritz-Timme S**, **Cattaneo C. 2013.** Pitfalls at the root of facial assessment on photographs: a quantitative study of accuracy in positioning facial landmarks. *The International Journal of Legal Medicine* **127:** 699–706.

Dryden IL, **Mardia KV. 1998.** *Statistical shape analysis*. Chichester: Wiley.

Dupré J. 1995. *The disorder of things: metaphysical foundations of the disunity of science*. Boston: Harvard University Press.

Dushoff J, **Kain MP**, **Bolker BM. 2019.** I can see clearly now: reinterpreting statistical significance. *Methods in Ecology and Evolution* **10:** 756–759.

Engelkes K, **Helfsgott J**, **Hammel JU**, **Büsse S**, **Kleinteich T**, **Beerlink A**, **Gorb SN**, **Haas A. 2019.** Measurement error in μ CT-based three-dimensional geometric morphometrics introduced by surface generation and landmark data acquisition. *Journal of Anatomy* **235:** 357–378.

Ellison AM. 2004. Bayesian inference in ecology. *Ecology Letters* **7:** 509–520.

Esquerré D, **Ramírez-Álvarez D**, **Pavón-Vázquez CJ**, **Troncoso-Palacios J**, **Garín CF**, **Keogh JS**, **Leaché AD. 2019.** Speciation across mountains: phylogenomics, species delimitation and taxonomy of the *Liolaemus leopardinus* clade (Squamata, Liolaemidae). *Molecular Phylogenetics and Evolution* **139:** 106524.

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 Medical Imaging* **14:** 35.

Fidler F, **Wilcox J. 2018.** *Reproducibility of scientific results in the Stanford encyclopedia of philosophy*. California: Stanford University Press.

Fraser H, **Parker T**, **Nakagawa S**, **Barnett A**, **Fidler F. 2018.** Questionable research practices in ecology and evolution. *PLoS One* **13:** e0200303.

Fruciano C. 2016. Measurement error in geometric morphometrics. *Development Genes and Evolution* **226:** 139–158.

Fruciano C, **Celik MA**, **Butler K**, **Dooley T**, **Weisbecker V**, **Phillips MJ. 2017.** Sharing is caring? Measurement error and the issues arising from combining 3D morphometric datasets. *Ecology and Evolution* **7:** 7034–7046.

Fruciano C, **Schmid D**, **Ramírez Sanchez MM**, **Morek W**, **Avila Valle Z**, **Talijančić I**, **Pecoraro C**, **Schermann Legionnet A. 2020.** Tissue preservation can affect geometric morphometric analyses: a case study using fish body shape. *Zoological Journal of the Linnean Society* **188:** 148–162.

Garnett ST, **Christidis L. 2017.** Taxonomy anarchy hampers conservation. *Nature News* **546:** 25.

Gelman A, **Carlin J. 2014.** Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* **9:** 641–651.

Gelman A, **Carlin JB**, **Stern HS**, **Dunson DB**, **Vehtari A**, **Rubin DB. 2013.** *Bayesian data analysis*. New York: Chapman and Hall/CRC.

Gelman A, **Rubin DB. 1992.** Inference from iterative simulation using multiple sequences. *Statistical Science* **7:** 457–472.

Gunz P, **Mitteroecker P. 2013.** Semilandmarks: a method for quantifying curves and surfaces. *Hystrix* **24:** 103–109.

Hedges LV, **Olkin I. 1985.** *Statistical methods for meta-analysis*. Orlando: Academic Press.

Hyndman RJ. 1996. Computing and graphing highest density regions. *The American Statistician* **50:** 120–126.

Ioannidis JP. 2005. Why most published research findings are false. *PLoS Medicine* **2:** e124.

Ioannidis JP. 2018. The proposal to lower P value thresholds to .005. *JAMA* **319:** 1429–1430.

Kaliontzopoulou A. 2011. Geometric morphometrics in herpetology: modern tools for enhancing the study of morphological variation in amphibians and reptiles. *Basic and Applied Herpetology* **25:** 5–32.

Kaliontzopoulou A, **Carretero MA**, **Llorente GA. 2007.** Multivariate and geometric morphometrics in the analysis of sexual dimorphism variation in *Podarcis* lizards. *Journal of Morphology* **268:** 152–165.

Kaliontzopoulou A, **Carretero MA**, **Llorente GA. 2010.** Intraspecific ecomorphological variation: linear and geometric morphometrics reveal habitat-related patterns within *Podarcis bocagei* wall lizards. *Journal of Evolutionary Biology* **23:** 1234–1244.

Kellner K. 2018. *jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses*. R package version 1.5.0. Available at: https://CRAN.R-project.org/package=jagsUI.

Kimmerle EH, **Ross A**, **Slice D. 2008.** Sexual dimorphism in America: geometric morphometric analysis of the craniofacial region. *Journal of Forensic Sciences* **53:** 54–57.

Klein LL, **Caito M**, **Chapnick C**, **Kitchen C**, **O'Hanlon R**, **Chitwood DH**, **Miller AJ. 2017.** Digital morphometrics of two North American Grapevines (*Vitis*: Vitaceae) quantifies leaf variation between species, within species, and among individuals. *Frontiers in Plant Science* **8:** 373.

Klingenberg CP. 2009. Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. *Evolution & Development* **11:** 405–421.

Klingenberg C. 2015. Analyzing fluctuating asymmetry with geometric morphometrics: concepts, methods, and applications. *Symmetry* **7:** 843–934.

Kruschke J. 2014. *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan*. Boston: Academic Press.

Kuhl FP, **Giardina CR. 1982.** Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing* **18:** 236–258.

Leaché AD, **Koo MS**, **Spencer CL**, **Papenfuss TJ**, **Fisher RN**, **McGuire JA. 2009.** Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (*Phrynosoma*). *Proceedings of the National Academy of Sciences of the United States of America* **106:** 12418–12423.

Loken E, **Gelman A. 2017.** Measurement error and the replication crisis. *Science* **355:** 584–585.

MacLeod N, **Krieger J**, **Jones KE. 2013.** Geometric morphometric approaches to acoustic signal analysis in mammalian biology. *Hystrix* **24:** 110–125.

Marcy AE, **Fruciano C**, **Phillips MJ**, **Mardon K**, **Weisbecker V. 2018.** Low resolution scans can provide a sufficiently accurate, cost-and time-effective alternative to high resolution scans for 3D shape analyses. *PeerJ* **6:** e5032.

Maxwell SE, **Lau MY**, **Howard GS. 2015.** Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist* **70:** 487.

McElreath R. 2018. *Statistical rethinking: a Bayesian course with examples in R and Stan*. New York: Chapman and Hall/CRC.

McShane BB, **Gal D**, **Gelman A**, **Robert C**, **Tackett JL. 2019.** Abandon statistical significance. *Journal of the American Statistical Association* **73:** 235–245.

Medina CD, **Avila LJ**, **Sites JW Jr**, **Santos J**, **Morando M. 2018.** Alternative methods of phylogenetic inference for the Patagonian lizard group *Liolaemus elongatus-kriegi* (Iguania: Liolaemini) based on mitochondrial and nuclear markers. *Molecular Phylogenetics and Evolution* **120:** 158–169.

Plummer M. 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vol. 124*, 1–10. March. doi:10.1.1.13.3406.Available at: https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Plummer M, **Best N**, **Cowles K**, **Vines K. 2006.** CODA: convergence diagnosis and output analysis for MCMC. *R News* **6:** 7–11.

Plummer M, **Stukalov A. 2018.** *rjags: Bayesian graphical models using MCMC*. Available at: https://cran.r-project.org/web/packages/rjags/index.html.

Raposo MA, **Stopiglia R**, **Brito GRR**, **Bockmann FA**, **Kirwan GM**, **Gayon J**, **Dubois A. 2017.** What really hampers taxonomy and conservation? A riposte to Garnett and Christidis (2017). *Zootaxa* **4317:** 179–184.

R Core Team. **2019.** *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Robinson C, **Terhune CE. 2017.** Error in geometric morphometric data collection: combining data from multiple sources. *American Journal of Physical Anthropology* **164:** 62–75.

Rohlf FJ. 2003. Bias and error in estimates of mean shape in geometric morphometrics. *Journal of Human Evolution* **44:** 665–683.

Rohlf FJ. 2015. *tpsDig, version 2.16. Ecology and Evolution*. Suny at Stony Brook Available at: http://life.bio.sunysb.edu/morph.

Ross AH, **Williams S. 2008.** Testing repeatability and error of coordinate landmark data acquired from crania. *Journal of Forensic Sciences* **53:** 782–785.

Schmidt S. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology* **13:** 90–100.

Shearer BM, **Cooke SB**, **Halenar LB**, **Reber SL**, **Plummer JE**, **Delson E**, **Tallman M. 2017.** Evaluating causes of error in landmark-based data collection using scanners. *PLoS One* **12:** e0187452.

Shrout PE, **Rodgers JL. 2018.** Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. *Annual Review of Psychology* **69:** 487–510.

Slice DE. 2005. Modern morphometrics. In: Slice DE eds. *Modern morphometrics in physical anthropology. Developments in primatology: Progress and prospects*. Boston, MA: Springer, 1–45.

Toma AM, **Zhurov A**, **Playle R**, **Ong E**, **Richmond S. 2009.** Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images. *Orthodontics & Craniofacial Research* **12:** 33–42.

Viscosi V. 2015. Geometric morphometrics and leaf phenotypic plasticity: assessing fluctuating asymmetry and allometry in European white oaks (*Quercus*). *Botanical Journal of the Linnean Society* **179:** 335–348.

Viscosi V, **Cardini A. 2011.** Leaf morphology, taxonomy and geometric morphometrics: a simplified protocol for beginners. *PLoS One* **6:** e25630.

Vrdoljak J, **Padró J**, **De Panis D**, **Soto IM**, **Carreira VP. 2019.** Protein–alkaloid interaction in larval diet affects fitness in cactophilic *Drosophila* (Diptera: Drosophilidae). *Biological Journal of the Linnean Society* **127:** 44–55.

Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. New York: Springer.

Zelditch ML, **Swiderski DL**, **Sheets HD. 2012.** *Geometric morphometrics for biologists: a primer*. London: Elsevier Academic Press.

Zimova M, **Mills LS**, **Nowak JJ. 2016.** High fitness costs of climate change-induced camouflage mismatch. *Ecology Letters* **19:** 299–307.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. Detailed information on each specimen used.
Appendix S2. Landmark and semilandmark placement, and contours outlined on *Liolaemus elongatus* (Figure app 2.1), *Drosophila buzzatii* (Figure app 2.2), and *Vitis riparia* (Figure app 2.3).
Appendix S3. Statistical models in JAGS programming language: (1) among-operator design, (2) across-taxon design, (3) related-species design, and (4) regression model for processing time and centroid size.

Figure S1. Different shape changes. (a) Changes between the same specimen digitized by different operators with the full-landmark protocol. (b) Changes between the same specimen digitized by different operators with the partial-landmark protocol. Arrows indicate conflictive landmarks (i.e. those that were erroneously placed by some operators).

Figure S2. Posterior mean and high posterior density interval (HPD) of 90% (bold line) and 95% (thin line) for each source of variation in the second principal component: sp, specimen; side, side; op, operator; error, measurement error. *Interaction between sources of variation. Protocols: P-S, partial-semilandmark; F-S, full-semilandmarks; C, contour; P-L, partial-landmark; F-L, full-landmark. Box at the top: percentage of variation explained by the second principal component for each protocol.

Figure S3. Correlation between time (in seconds) and size for landmark and semilandmark protocols in each taxon. Green: lizards (*Liolaemus elongatus*); blue: flies (*Drosophila buzzatii*); red: leaves (*Vitis riparia*). Continuous lines represent the lineal regression whereas dashed lines represent simulated normal envelopes from posterior means and standard deviations.

Figure S4. Shape changes across the first principal component of related-species designs for: (a) the partial-semilandmark protocol, (b) full-semilandmark protocol, (c) full-landmark protocol and (d) partial-landmark protocol. The scatter plots of the first two principal components of each design are shown.

Figure S5. Shape changes across the second principal component of related-species design for: (a) the partial-semilandmark protocol, (b) full-semilandmark protocol, (c) full-landmark protocol and (d) partial-landmark protocol.

SHARED DATA

The data underlying this work are available at doi.org/10.6084/m9.figshare.10022657.