Evolutionary Biology (2024) 51:179–207
https://doi.org/10.1007/s11692-024-09627-6

RESEARCH ARTICLE

# Interrogating Random and Systematic Measurement Error in Morphometric Data

Michael L. Collyer[1] · Dean C. Adams[2]

Received: 18 July 2023 / Accepted: 21 January 2024 / Published online: 16 February 2024

## Abstract

Measurement error is present in all quantitative studies, and ensuring proper biological inference requires that the effects of measurement error are fully scrutinized, understood, and to the extent possible, minimized. For morphometric data, measurement error is often evaluated from descriptive statistics that find ratios of subject or within-subject variance to total variance for a set of data comprising repeated measurements on the same research subjects. These descriptive statistics do not typically distinguish between random and systematic components of measurement error, even though the presence of the latter (even in small proportions) can have consequences for downstream biological inferences. Furthermore, merely sampling from subjects that are quite morphologically dissimilar can give the incorrect impression that measurement error (and its negative effects) are unimportant. We argue that a formal hypothesis-testing framework for measurement error in morphometric data is lacking. We propose a suite of new analytical methods and graphical tools that more fully interrogate measurement error, by disentangling its random and systematic components, and evaluating any group-specific systematic effects. Through the analysis of simulated and empirical data sets we demonstrate that our procedures properly parse components of measurement error, and characterize the extent to which they permeate variation in a sample of observations. We further confirm that traditional approaches with repeatability statistics are unable to discern these patterns, improperly assuaging potential concerns. We recommend that the approaches developed here become part of the current analytical paradigm in geometric morphometric studies. The new methods are made available in the `RRPP` and `geomorph` R-packages.

**Keywords** Morphometrics · Landmarks · Measurement error

## Introduction

Quantitative inferences in evolutionary biology are made by estimating biological signal from empirical observations, and evaluating that signal relative to expectation under a particular hypothesis (Houle et al., 2011). However, this seemingly straightforward endeavor is compromised by the fact that our observations are impacted by measurement error (Fleiss & Shrout, 1977; Kreutz et al., 2013). Measurement error affects one's ability to distinguish signal from noise, and is a pervasive problem in all quantitative disciplines. The field of morphometrics is no exception. Here, the biometer quantifies anatomical shapes from sets of linear measurements, or increasingly, from landmark points representing discrete anatomical locations, curves and surfaces of structures, as commonly found in geometric morphometric data (Adams et al., 2013; Bookstein, 1991; Mitterœcker & Schæfer, 2022). From these measurements, one may characterize the shape of anatomical objects, summarize patterns of shape variation for a sample of observations, and describe the covariation of shape with other explanatory variables. Yet our morphometric data contain uncertainty associated with the values assigned to each landmark, which can inflate the inter-specimen variation in a sample (Arnqvist & Mårtensson, 1998; Bailey & Byrnes, 1990; Yezerinac et al., 1992). This can have potentially serious consequences for making downstream statistical and biological inferences, and thus it is incumbent upon the biometer to ensure that the effects of measurement error are minimized, as much as possible.

✉ Michael L. Collyer
m.collyer@chatham.edu

[1] Department of Science, Chatham University, Pittsburgh, PA, USA

[2] Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA

To do so first requires an understanding of the major components of measurement error, and how they manifest in a sample of observations. In the field of measurement theory, measurement error is defined as the deviation between a measured quantity and its true value (sensu Rabinovich 2005). This deviation exists in part because the actual value of any physical attribute is unknown, and thus quantitative values assigned to it are inexact estimates (Hand, 1996; Krantz et al., 1971; Kyburg, 1984; Luce et al., 1990; Rabinovich, 2005; Suppes et al., 1989). Additionally, imprecision in these estimates—due to instrumentation inaccuracies, how observers take readings, or inconsistencies in experimental procedure—further contribute to these deviations. Collectively, these deviations result in measurement error (ME). Importantly, measurement error may occur randomly across observations, or it may deviate systematically in some manner (Hand, 1996; Rabinovich, 2005). Random ME corresponds to stochastic variation in the magnitude or direction of deviations from observation to observation. Statistically, random ME has a well-known and obvious effect; it increases the variance in a sample, and thus increases the potential for type II errors in hypothesis tests (Yezerinac et al., 1992). In other words, random ME impinges on the biometer's ability to detect a signal when it is present in a sample. By contrast, systematic ME corresponds to differences that vary in regular fashion in repeated measurements of the same observations. Because these deviations are nonrandomly distributed, systematic ME can result in estimation bias of model coefficients and can manifest as a measurable signal, thereby altering the actual biological signal present in the dataset. Thus, from a statistical standpoint, systematic ME is a far more insidious problem, as it has the potential to lead biological inferences astray.

That measurement error exists in morphometric data is not in dispute. Rather, for the biometer, the concerns are: (1) How to detect it? and (2) How to minimize it? With respect to the former, deviations from the true value cannot typically be used to estimate ME, because the true value cannot be known precisely (Rabinovich, 2005). Instead, ME is most commonly characterized by taking repeated measurements of the same observations, and summarizing the within-subject (i.e., among-replicate) variation. Here, smaller within-subject variation implies less ME, and thus greater repeatability of the estimated measurements (Bailey & Byrnes, 1990). To assess this, a repeated measures analysis of variance (ANOVA) model may be used to attribute variance to model effects, and to isolate the within-subject variance component (Arnqvist & Mårtensson, 1998; see Fleiss & Shrout, 1977). The latter may be conveyed as the intra-class correlation, or *ICC* (Bartko, 1966; Fisher, 1950; Haggard, 1958; Liljequist et al., 2019), which describes the among-subject variance relative to the total variation in the sample. The *ICC* expresses the degree to which repeated measurements are similar, and thus, higher values imply lower ME. Multivariate analogs have been proposed for *ICC* using canonical correlation analyses between covariance matrices (e.g., Konishi et al., 1991), but these approaches compare the covariance matrices of inherently related subjects (like parents and offspring) rather than repeated measurements of the same subjects. Similarly, the within-subject variance component itself, or its associated coefficient of determination ($R^2$), may be used as a heuristic to describe the percentage of variation attributable to ME in a dataset (Galimberti et al., 2019; Klingenberg et al., 2002). Taken together, these summary measures ($ICC$, $R^2$) are relatively straightforward to calculate, and not surprisingly, are used in a wide variety of disciplines. However, it should be recognized that they are agnostic to the type of ME present in a sample. As typically implemented, they characterize the overall magnitude of ME, but are generally incapable of disentangling any random and systematic components that may be present.

In 1998, Arnqvist and Mårtensson brought the topic of measurement error to the attention of practitioners of geometric morphometrics (GM), and highlighted the importance of investigating measurement error in landmark data. Their seminal review described in detail how ME permeates the various steps of our digitization and analytical pipelines, proposed strategies for minimizing ME, and advocated that summary measures such as the *ICC* be regularly used to gauge the extent of ME in a morphometric sample. Since then, an increasing number of GM studies have incorporated an evaluation of ME as part of their data analytic procedures. Typically, these studies leverage repeated measurements of observations, and utilize one or more of the summary measures mentioned above. In fact, a survey of the recent literature reveals a rather diverse set of publications, which includes studies that assess the overall level of ME in a sample (e.g., Fox et al., 2020; Vrdoljak et al., 2020), studies that evaluate the precision of particular landmarks (Barbeito-Andrés et al., 2012; Cramon-Taubadel et al., 2007), and studies that evaluate inter-observer error and device-specific differences (e.g., Fruciano et al., 2017; Giacomini et al., 2019; Marcy et al., 2018; Menéndez, 2016; Robinson & Terhune, 2017; Shearer et al., 2017). Thus, it appears that Arnqvist and Mårtensson's (1998) call to arms has been heeded by the morphometric community, and evaluations of measurement error are now much more routine. We view this to be a positive development.

Since the publication of Arnqvist and Mårtensson's treatise 25 years ago, the field of geometric morphometrics has witnessed a veritable explosion of analytical advances in many topical areas, developed to address a wide array of biological hypotheses (Adams, 2014; Adams & Collyer, 2019; Bookstein, 2015; Bookstein et al., 2003; Collyer & Adams, 2013; Conaway & Adams, 2022; Gunz et al., 2005;

Klingenberg & Gidaszewski, 2010; Mitterœcker & Bookstein, 2009; Mitterœcker et al., 2004; Rohlf & Corti, 2000, to name a few). Yet curiously, little has changed in terms of the recommendations regarding how one should evaluate measurement error in GM studies. For instance, a current review of the topic (Fruciano, 2016) offers: (1) a careful scrutiny of one's digitizing procedures, (2) visual inspection of one's data to identify problematic landmarks and dispersion among within-subject replicates, (3) the use of summary measures as heuristics to evaluate the extent to which ME may be present, and (4) evaluation of differences between observers or devices when such data are available. Yet this is essentially the same advice as advocated by Arnqvist and Mårtensson in 1998, with a modern focus on available software. Other reviews of the subject (Daboul et al., 2018; Fruciano et al., 2017) proffer similar suggestions without alteration. In fact, apart from an alternative permutation scheme for testing inter-observer or inter-device differences (Fruciano et al., 2017), no new analytical procedures have been forwarded that explore aspects of ME from a new perspective. In short, the analytical machinery for investigating ME in geometric morphometric data has remained rather static for two and a half decades, and has not kept pace with analytical advances achieved in other areas of the field. We feel it is imperative to reacquaint the field of GM with analysis of ME, utilizing some of the statistical tools that have been developed in the last decade.

We contend that interrogating measurement error in GM studies should have the same degree of quantitative rigor as is currently attained in other areas of the field. To do so requires a more synthetic view of ME that is capable of decomposing it into its constituent components, and simultaneously evaluating the attributes of ME in terms of their magnitude, and their direction. By relating trends in ME to patterns present in one's data, the biometer can properly discern how ME influences their statistical, and thus biological conclusions.

In this article, we develop a novel set of analytical procedures and graphical tools that establish a new paradigm for how empiricists should investigate patterns of measurement error in multivariate data. Our approach dissects the random and systematic components of ME from one another, and extracts any group-specific systematic ME that may be present. Multivariate test measures are proposed to characterize these patterns, which are evaluated with appropriate permutation procedures. A set of graphical tools accompanies these procedures to provide additional insights. First we formalize the algebra of our approach. Then, through a series of motivating examples, we illustrate how different aspects of ME manifest in GM data, and demonstrate how our new analytical paradigm detects these patterns. Computer simulations are then used to verify that associated permutation tests display appropriate statistical properties. An important outcome of these simulations is the observation that Procrustes superimposition buffers against the negative impacts of systematic ME, rather than enhancing them. Next, a reanalysis of an empirical dataset illustrates the dissection of ME into its random and systematic components, and reveals that the main direction of systematic ME in this example coincides with the direction of biological signal; obfuscating interpretation of the latter. This highlights the importance of performing a more comprehensive interrogation of ME in morphometric datasets, which our analytical and visual tools provide. Finally, all methods developed in this article are available in the R-packages `geomorph` (Adams et al., 2023; Baken et al., 2021) and `RRPP` (Collyer & Adams, 2018, 2023) libraries.

## Methods and Results

We present updated and novel methods for the analysis of ME by first introducing the conceptual basis for the methods, explaining what systematic and random components of ME mean, and how they manifest in GM data. We introduce examples for simulation experiments, which create plausible contexts for varied amounts of systematic and random ME, based on repeated digitizations of the same landmark configuration. The examples covered in the simulation experiments help ground the conceptual basis for the methods we propose in a realistic way by syncing graphical patterns to statistical results. Statistical methods include a novel resampling procedure used to create empirical sampling distributions of test statistics for Procrustes ANOVA and multivariate ANOVA (MANOVA), plus a graphical tool to assist in assessing and interpreting the amounts and patterns of systematic and random ME in a GM-ME experiment. In the work below, a GM-ME experiment is any study that selects specimens for digitizing and uses a systematic method of repeated digitizations of the same landmark configuration on each specimen, resulting in GM data.

### Conceptual Basis for the Analysis of ME

In the purest sense, ME is a quantifiable divergence from a true value or suite of values made by a process intended to replicate the true value. An example that might be easy to appreciate for researchers who use landmark-based GM data involves several machines in a factory that are used to drill holes in wood planks for assembling furniture. Machines are programmed to drill a specific configuration of holes. There is, therefore, a known "true" configuration from which departures can be measured for each of the machines. ME is the measured result of any tendency for machines to misplace holes in the locations they were programmed to be placed. The amount of ME is directly related to the imprecision of hole-placement in the drilling process. However, the imprecision can be defined in different ways. One could measure the displacement of a particular hole from its target, both in the distance from the true location and the direction in which it was displaced. Alternatively, and more relevant for GM data, one could attempt to measure the mismatch of the entire configuration to the true configuration. Even if the reason for any ME is localized to one hole (landmark), the difference between true and replicated configurations can be observed at every hole, after the configurations have been aligned to best match all corresponding holes to each other.

If the drilling of configurations was replicated several times, per machine, ME might be consistent, for example, as a displacement of a specific hole to the left of its true location. This would be indicative of a systematic bias or prejudice of the machine. Because there is some repeatability of this type of error, the resulting displacement of the hole is referred to as systematic ME. This is an obvious trend, unlike random ME, a tendency for misplacement of one or more holes, but not in a predictable way. Both systematic and random ME could be measured, provided replication in measurements is made on sample planks, for multiple individual machines.

The practicality of the machine example breaks down perhaps with the realization that in just about any GM study, the true configuration is not known. However, as presented, this example is not the only way to assess ME. It can be implied from the example that machines are research subjects and replication of the wood-drilling process occurs for multiple configuration-drillings by each research subject. This might seem practical if there is only one configuration of points to consider. If, however, various different configurations could be programmed into each machine, a GM-ME experimental design like the one above, repeated for every configuration, would require many observations (which might be costly), and would allow inference only to be made, configuration by configuration, and machine by machine. Rather, if the configurations were considered research subjects and the machines replications of the process applied to each subject, the tendency of any one machine to misplace holes could be assessed, irrespective of configurations. Furthermore, knowing the true configuration that is programmed into each machine would not be as necessary as understanding the tendency for machines to drill the same configurations, especially if evaluating the consistency of machines to perform the same process—regardless of configuration—was the purpose of the study[1]

This alternative design draws more parallels to GM studies. Research subjects are specimens on which landmark configurations are placed, and replications are repeated digitizations, that are distinct in some way. For example, two or more researchers digitize the same photos; a researcher digitizes the same configuration on separate photos of the same specimen; a researcher and automated digitizer digitize the same configuration on research specimens; two different scanners are used to collect 3D surface points on the same object; and other scenarios are certainly possible. Assessment of ME is consistent with an assessment of the repeatability of digitizing a landmark configuration on the same specimen and getting the same results. There is no need to have a "true" configuration. Rather, an assessment of the tendency for repeated digitizations on the same specimens to produce shapes in a shape space that are in close proximity, compared to the shapes of disparately shaped specimens, is the goal. ME is the measurable disparity among replicated measures of the same research subjects. Quantifying ME is challenging, because there is no appreciable range of expectation without relativizing the variation among replicated measurements to some other source of shape variation. Regardless, a design that has the same configuration digitized multiple times on the same specimen—the measurements nested within a research subject—also repeated for multiple specimens, allows assessment of ME in GM studies.

Unfortunately, the data of landmark-based GM—the Procrustes coordinates[2] from generalized Procrustes analysis (GPA) (Adams et al., 2013; Rohlf & Slice, 1990)—involve transformation that can obfuscate specific digitizing phenomena. ME most typically will be measured on Procrustes coordinates, as the elements of configuration size, orientation, and position would make an analysis on the raw coordinates of digitized landmarks impractical. However, it is the impact that a digitizing prejudice—the tendency of a digitizing process to impose a consistent change in the location of one or more landmarks in a configuration compared to another—has on the estimation of the shape of specific research subjects or the groups that contain them that is probably of most interest. For example, if a researcher digitizes a landmark configuration on 2D photos of research specimens (first replicate) and an automated digitizer places the same landmarks on the same photos, and it is revealed that the landmarks of the automated digitizer are misplaced in the same direction by the same amount (accounting for

[1] If only one machine was the cause of inconsistency, it would be clear which machine it was, regardless of the exactness of any machine to produce the true configuration.

[2] Often the terms, "Procrustes residuals" and "Procrustes coordinates" are used almost interchangeably. Procrustes coordinates are the mean configuration after GPA, plus the Procrustes residuals, which are the deviations of configuration-specific coordinates from the mean. Either can be used in most analyses, producing the same results, as the mean shape would be constant for every research observation.

specimen orientation), then there might be little concern. If every landmark was perfectly displaced, the resulting configurations would have the same size and there would be no difference between the coordinates after GPA.[3] However, if the displacement occurs for one or few landmarks, only, the configurations would have different size and mismatched coordinates after GPA, but not only for the landmarks where the mistake occurred. Even though the digitizing prejudice is an attribute of the process that places raw landmarks, it is in most cases the change in Procrustes coordinates that result from that process that is a concern. Procrustes coordinates are the data from which ME is measured.

Digitizing prejudice should translate to systematic ME that can be quantified in an analysis of ME performed on Procrustes coordinates. If the effect of systematic ME can be measured, the shape change associated with this effect can be envisioned by mapping the mean configuration of Procrustes coordinates onto a configuration changed by the effect, which might reveal which landmarks are most likely changed as a result of a digitizing prejudice. Alternatively, random ME has no specific directional shape change but signifies that different shapes are observed among digitizing replicates of the same subject. For example, if the same research specimens are digitized by two researchers, one who is meticulous and one who is sloppy, pairs of shapes for research subjects might appear displaced in a principal component (PC) plot, but in no consistent way. This is in contrast to systematic ME, which would be revealed more so as a tendency for consistent displacement. Greater ME, whether systematic or random, will be revealed by greater disparity between corresponding points of subject replicates in a PC plot. Random ME might not be of much concern, if small, as it might not have much impact on the estimation of subject shapes. Systematic ME can be of great concern, however, even if small, as it could lead to biased shape estimates for some but not all research subjects, which would have implications for analyses that target estimation of shape change among groups. An analysis of ME ideally evaluates the impact of systematic ME, in addition to measuring the amount of ME, whether random or systematic. As we show below, systematic and random ME can be partitioned, and systematic ME tested, with an appropriate analytical paradigm. First we outline a few hypothetical examples for the types of systematic ME one might wish to detect.

## Motivating Examples (and Simulation Experiment Set-Up)

In this paper, we use simulation experiments to assess type I error rates and statistical power for testing for systematic ME, based on six examples of varied but realistic systematic and random ME. In each case, random landmark configurations were simulated (more detail below) that were practically invariant to positional and rotational differences (except if simulated by chance, in which case they would be slight). As is typical with most GM-ME experiments, we eventually perform statistical analysis on Procrustes coordinates, following generalized Procrustes analysis (GPA) (Rohlf & Slice, 1990). However, because our simulation experiments did not vary position and rotation of configurations, it was also possible to perform statistical tests on raw landmarks for comparison.

The six experiments (Table 1) sought to evaluate the efficacy of ME tests for scenarios that varied the amounts of systematic and random ME, whether research subjects were sampled from different groups with specific shape differences (like sampling individuals from different species), whether a digitizing prejudice was applied to all specimens or specific groups of specimens, and varied how the digitizing prejudice might be applied to different groups.

Random subjects were simulated via the distortion of a landmark configuration template,

$$\mathbf{Y}_i = \mathbf{Y}_0\mathbf{H}_i, \tag{1}$$

where $\mathbf{Y}_0$ was the $p \times 2$ template (resembling a fish) and $\mathbf{H}_i$ was a $2 \times 2$ symmetric transformation matrix for the $p$ points in $k = 2$ dimensions ($x$ and $y$ Cartesian coordinates) found in $\mathbf{Y}_0$. $\mathbf{H}_i$ was randomly sampled for subject $i$, by modifying a $2 \times 2$ identity matrix by adding values sampled from a normal distribution ($\delta$) with a mean of 0 to elements of the identity matrix; i.e.,

$$\mathbf{H}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \delta_x & \delta_{xy} \\ \delta_{xy} & \delta_y \end{bmatrix}, \tag{2}$$

where $\delta_x \sim \mathcal{N}(\mu = 0, \sigma_x)$, $\delta_y \sim \mathcal{N}(\mu = 0, \sigma_y = 0.5\sigma_x)$, and $\delta_{xy} \sim \mathcal{N}(\mu = 0, \sigma_{xy} = 0.25\sigma_x)$. This approach allowed more shape change in the $x$-direction (lengthening) than in the $y$ -direction (deepening), and allowed the covariance between $x$ and $y$ coordinates to remain consistent and comparatively muted to the lengthening or deepening of the configuration. By randomly sampling $\mathbf{H}_i$ in Eq. (2) for each simulated research subject, initial (first replicate) inter-subject variation in shape among subjects was simulated. We varied the amount of inter-subject variation by simply changing the value of $\sigma_x$. Figure 1 demonstrates how variation in fish shapes could be generated.

[3] Despite the imprecision of the automated digitizer compared to the researcher, the configurations it produces are accurate with respect to the researcher's.

**Table 1** Explanation of simulation experiments, indicating purpose, how systematic and random ME were varied, and whether group differences in shape were included in analysis

| Experiment | Systematic ME | Random ME | Group differences in shape | Purpose |
|---|---|---|---|---|
| 1 | None | Progressive, from small to large | None | To determine if the amount of random ME (digitizing noise) influences tests for systematic measurement error, before or after GPA |
| 2 | None | Constant and relatively small | Progressively larger group differences | To determine if sampling research subjects from distinctly different shaped groups could influence tests of systematic measurement error, before and after GPA |
| 3 | Progressive, from small to large, applied to each research subject | Constant and relatively small | Three levels: no group differences, small group differences, and large group differences | To determine the responsiveness of tests for systematic ME based on the amount of digitizing prejudice applied, before and after GPA. Additionally, to determine whether group differences affect tests, both for a global systematic ME and a systematic ME by group interaction |
| 4 | Progressive, from small to large, applied to each research subject in only one group (enhancing group difference) | Constant and relatively small | Three levels: no group differences, small group differences, and large group differences | To determine the responsiveness of tests for systematic ME based on the amount of digitizing prejudice applied, only to a particular group, in a direction of group differences (increased group difference), before and after GPA. Additionally, to determine whether group differences affect tests, both for a global systematic ME and a systematic ME by group interaction |
| 5 | Progressive, from small to large, applied to each research subject in only one group (retarding group difference) | Constant and relatively small | Three levels: no group differences, small group differences, and large group differences | To determine the responsiveness of tests for systematic ME based on the amount of digitizing prejudice applied, only to a particular group, in a direction opposite of group differences (decreased group difference), before and after GPA. Additionally, to determine whether group differences affect tests, both for a global systematic ME and a systematic ME by group interaction |
| 6 | Progressive, from small to large, applied to each research subject in only one group (not in a direction of group difference) | Constant and relatively small | Three levels: no group differences, small group differences, and large group differences | To determine the responsiveness of tests for systematic ME based on the amount of digitizing prejudice applied, only to a particular group, in a direction orthogonal to group differences (changed group but not in a direction that defines group differences), before and after GPA. Additionally, to determine whether group differences affect tests, both for a global systematic ME and a systematic ME by group interaction |

To simulate inherent group differences (for example, by sampling research subjects from different species), an update to Eq. (1) was performed for the **first replicate** as,

$$\mathbf{Y}_i = \mathbf{Y}_0\mathbf{H}_i, +\mathbf{G}_j, \tag{3}$$

where $\mathbf{G}_j$ was a $p \times 2$ matrix comprising mostly 0s (no displacement) except for the elements found at $(p-1, 1)$ and $(p, 1)$ to allow shifting of two tail landmarks, consistently, only along $x$ Cartesian directions, pertaining to expected group shape difference for groups. For group $a$ ($j = 1$), these values were 0. A predefined group difference ($d$) was assigned for $(p-1, 1)$ and $(p, 1)$ for group $b$ ($j = 2$), and $2d$ was assigned for group $c$ ($j = 3$). In other words, if group differences were included ($d > 0$), differences in shape were attained by shifting $x$ Cartesian coordinates for two landmarks by an amount, $d$, for group $b$ and $2d$ for group $c$. If no group differences were assigned, $\mathbf{G}_j$ was a matrix of 0s, meaning the simulated $\mathbf{Y}_i$ was unchanged. An example of the outcome of this simulation protocol, based on $\sigma_x = 0.20$ is shown in Fig. 2.

To simulate random ME, an update to Eq. (3) was performed for the **second replicate** as,

$$\mathbf{Y}_i = \left(\mathbf{Y}_0\mathbf{H}_i + \mathbf{G}_j\right) + \mathbf{R}_i, \tag{4}$$

where $\mathbf{R}_j$ was a $p \times 2$ matrix comprising $2p$ random values sampled from a normal distribution, $\mathcal{N}\left(\mu = 0, \sigma_r\right)$, where $\sigma_r$ defined how variable random digitizing error could be. These values were simulated independently (isotropic scatter). The parentheses around $\mathbf{Y}_0\mathbf{H}_i + \mathbf{G}_j$ indicate the fixed value for the first replicate, changed for the second replicate by the addition of $\mathbf{R}_j$. Figure 3 shows how random ME as digitizing error can be simulated.

A digitizing prejudice (systematic ME) could also be added to Eq. (4) with an additional update,

$$\mathbf{Y}_i = \left(\mathbf{Y}_0\mathbf{H}_i + \mathbf{G}_j\right) + \mathbf{R}_i + \mathbf{S}_j, \tag{5}$$

where $\mathbf{S}_j$ resembles $\mathbf{G}_j$ but with different displacement of the $x$ or $y$ Cartesian coordinates for the same landmarks that are shifted for group differences. In our simulations, either all of $\mathbf{S}_j$ were 0, if not simulating systematic ME, contained consistent displacements for the $p-1$ and $p$ landmarks (in either $x$ or $y$ directions) to simulate the same digitizing prejudice applied to all research subjects, or contained displacements only for group $a$ (0 values for groups $b$ and $c$) to simulate a digitizing prejudice applied only to one group (e.g., species). Figure 4 illustrates how digitizing prejudice in the second replicate can manifest as shape changes (without group differences).

By simulating configurations with Eq. (5) it was possible to obtain landmarks and Procrustes coordinates for the consideration of every scenario in Table 1. Tests of systematic ME for these scenarios involve both univariate-like (Procrustes) ANOVA, based on the dispersion of shapes, or multivariate-ANOVA (MANOVA) statistics, based on linear model covariance matrices (using principal component scores). We describe these in more detail in the next four sections.

## A Resampling Procedure to Test Systematic Measurement Error

An analysis of ME foremost is a test of systematic ME. A null hypothesis of no systematic ME is not exactly the same as a null hypothesis of no difference in shape between replicated measurements of shape from the same research subject; it is a null hypothesis of no consistent shape change between replicates, among research subjects. This distinction is important as it distinguishes systematic ME from total ME. For a test of systematic ME, it is imperative that an evaluation of within-subject variation in shape can be assessed, despite variation among subjects. This might seem counter-intuitive, as the variation in shape among subjects is often used a basis for measuring ME in a relative way (as a percentage of subject or total variation). Although understanding subject variation might be important, the point made here is that a test that generates a sampling distribution of a statistic should not introduce changes in subject variation.

Randomization of residuals in a permutation procedure (RRPP) has become a common method for ANOVA in research using GM data (Collyer et al., 2015; Collyer & Adams, 2018), especially because of its ability to handle high-dimensional data (number of shape variables exceed the number of observations). RRPP generates empirical distributions of various ANOVA or pairwise test statistics, and its statistical properties (parameter estimates, empirical sampling distributions, type I error rates and statistical power) have been extensively vetted (Adams & Collyer, 2018, 2022; Collyer et al., 2022). The assertion that subject variation should remain constant in the analysis means that a sampling distribution of a statistic for systematic ME is developed for a process that produces the same subject variance in every random RRPP permutation. This is possible by restricting the randomization of residuals within subjects.

For example, for an $n \times (pk)$ matrix, $\mathbf{Z}$, containing $n$ vectors, $z_i^T$ for the $i = 1, 2, ..., n$ observations of Procrustes coordinates containing $p$ points in $k$ dimensions ($k = 2$ or 3), a linear model to estimate the overall mean takes the form, $\hat{\beta}_{null} = \bar{\mathbf{z}}^T = \left(\mathbf{X}_{null}^T\mathbf{X}_{null}\right)^{-1}\mathbf{X}_{null}^T\mathbf{Z}$, where $^T$ means vector or matrix transposition, and $^{-1}$ means matrix inversion. The linear model design matrix, $\mathbf{X}_{null}$, is a vector of 1s. The mean is a vector of coefficients ($\hat{\beta}_{null}$) that if multiplied times the linear model design matrix produces an $n \times p$ matrix of mean

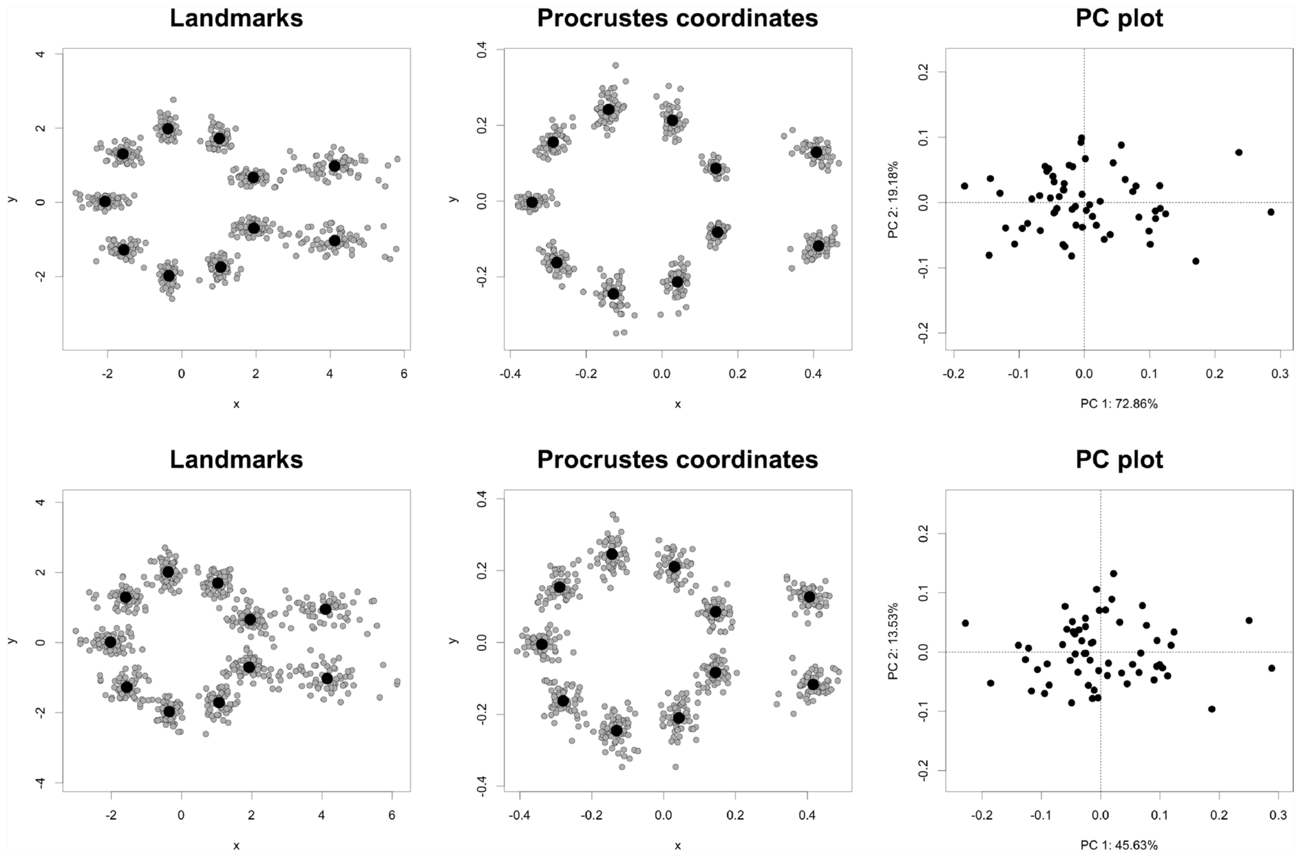

**Fig. 1** Example of simulated research subjects, with different inter-subject variation, based on $\sigma_x$. Top row: small variation, $\sigma_x = 0.02$. Bottom row: large variation, $\sigma_x = 0.16$. Left column: raw landmarks. Middle column: Procrustes coordinates, following GPA. Right column: plot of principal component scores. There are 60 subjects in each case

values; i.e., $\bar{\mathbf{Z}} = \mathbf{X}_{null}\hat{\beta}_{null}$. To estimate subject means, $\mathbf{X}_{null}$ can be updated by concatenating $s - 1$ columns of dummy variables for the $s$ subjects represented in the data. (Dummy variables comprise 0s and 1s, with 1s indicating subject match.) We assume that this resulting matrix, $\mathbf{X}_{subject}$ is balanced,[4] meaning there are an equal number of replicated observations within subjects; i.e., $n = sr$, where $r$ is the number of replicates. In this way, the column sums except the first of $\mathbf{X}_{subject}$ equal $r$ (the first equals $n$). We can estimate coefficients for subject means as $\hat{\beta}_{subject} = \left(\mathbf{X}^T_{subject}\mathbf{X}_{subject}\right)^{-1}\mathbf{X}^T_{subject}\mathbf{Z}$, and subject means as $\hat{\mathbf{Z}}_{subject} = \mathbf{X}_{subject}\hat{\beta}_{subject}$. The difference between subject means and the overall mean, $\hat{\mathbf{Z}} - \bar{\mathbf{Z}}$ is the basis for the subject variance. The covariance matrix is found as, $\hat{\boldsymbol{\Sigma}}_{subject} = (s - 1)^{-1}\left(\hat{\mathbf{Z}} - \bar{\mathbf{Z}}\right)^T\left(\hat{\mathbf{Z}} - \bar{\mathbf{Z}}\right)$, and its trace (sum of diagonal elements equal to the sum of variable variances) is the variance based on dispersion, the summed squared differences between the points of subject means and the overall mean. The $(s - 1)$ degrees of freedom represent the additional parameters in $\mathbf{X}_{subject}$ required to estimate subject means compared to the overall mean.

RRPP applied to the null model has first- and second moment exchangeability (Adams & Collyer, 2018; Commenges, 2003), meaning if residuals of the null model, $\mathbf{Z} - \bar{\mathbf{Z}}$, are randomly shuffled to produce random pseudodata, $\mathcal{Z} = \bar{\mathbf{Z}} + \left(\mathbf{Z} - \bar{\mathbf{Z}}\right)^*$, where $^*$ represents a randomized form of the residuals, the mean (first moment) and variance (second moment) of the pseudodata, $\mathcal{Z}$, will be the same as for the real data, $\mathbf{Z}$, in any random permutation. The same is not true with respect to the subjects model, if it is applied to $\mathcal{Z}$. Indeed, this is the basis for ANOVA, and how one might test for subject effects, if this would be of interest. The many permutations of $\mathcal{Z}$ makes it possible to generate sampling distributions of ANOVA statistics, so it is possible to evlauate a null hypothesis for subject variance. Rather, an analysis of ME seeks to preserve subject effects, not explicitly test for them. It might seem intuitive to randomize the residuals of the subjects model in a similar way; i.e.,

[4] There is not a strict need for replicate balance in the research design (see "Discussion"). However, issues like heterogeneity of variance among subjects might be more difficult to interpret with replicate imbalance.

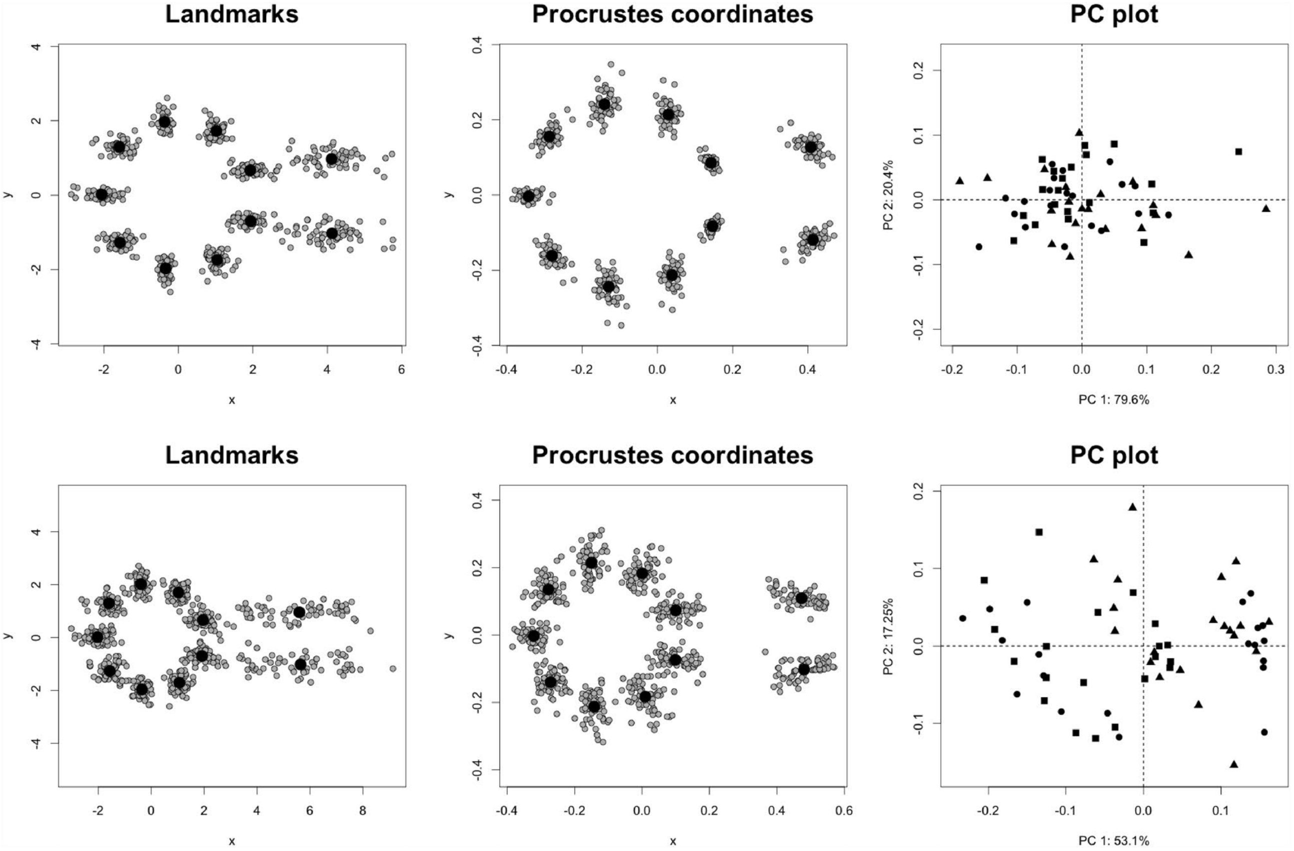

**Fig. 2** Example of simulated research subjects, with group differences. Top row: no group differences for 60 research subjects. Bottom row: group differences simulated for three groups of 20 subjects, via tail-lengthing. Left column: raw landmarks. Middle column: Procrustes coordinates, following GPA. Right column: plot of principal component scores, with different symbols corresponding to different groups

$\mathcal{Z} = \hat{\mathbf{Z}}_{subjects} + \left(\mathbf{Z} - \hat{\mathbf{Z}}_{subjects}\right)^{*}$, but RRPP this way would not have exact first- and second-moment exchangeability, even if approximately the same means and variance are found across permutations. However, a slight alteration makes it possible to achieve first- and second-moment exchangeability. If RRPP is restricted within subjects, subject means and subject variance will remain constant across permutations, for either model. This should be obvious. Changing the order of replicates within one subject will not change the subject mean or variance among observations for that subject. However, RRPP that randomizes the order of replicates many times for every subject makes it possible to evaluate the consistency of replicate changes in shape among all subjects. Thus, restricted (within-subject) RRPP makes it possible to test for systematic ME.

A test of systematic ME involves comparison of sums of squares and cross-products between two models: one that includes coefficients for subject means, and one that includes coefficients to estimate replicate means in addition to subject means. The latter model involves adding $r - 1$ parameters (dummy variables) to $\mathbf{X}_{subject}$ to form $\mathbf{X}_{subject+replicate}$. (We henceforth use $\mathbf{X}_s$ to mean $\mathbf{X}_{subject}$ and $\mathbf{X}_{sr}$ to mean $\mathbf{X}_{subject+replicate}$, for simplicity.) Coefficients can be estimated with a least-squares criterion, as before, and the fitted values compared between the two models, i.e.,

$$\begin{aligned}\mathbf{S}_r &= \left(\hat{\mathbf{Z}}_{sr} - \hat{\mathbf{Z}}_s\right)^T \left(\hat{\mathbf{Z}}_{sr} - \hat{\mathbf{Z}}_s\right) \\ &= \left(\mathbf{X}_{sr}\hat{\beta}_{sr} - \mathbf{X}_s\hat{\beta}_s\right)^T \left(\mathbf{X}_{sr}\hat{\beta}_{sr} - \mathbf{X}_s\hat{\beta}_s\right)\end{aligned} \tag{6}$$

where $\mathbf{S}_r$ is a $pk \times pk$ symmetric sums of squares and cross-products (*SSCP*) matrix, with variable (coordinate) sums of squares along the diagonal and cross-products between variables in the off-diagonal elements. In every subject-restricted RRPP permutation, $\hat{\mathbf{Z}}_s^T\hat{\mathbf{Z}}_s$ will be constant. If test statistics require inverting $\mathbf{S}_r$ (more on this below), a problem arises because $\mathbf{S}_r$ will be singular if using Procrustes coordinates, due to invariance in size, orientation, and position of configurations imposed by GPA (and potential redundancies due to use of sliding semi-landmarks). In such cases, finding vectors of principal component scores ($\mathbf{P}$) of $\mathbf{Z}$ (explaining

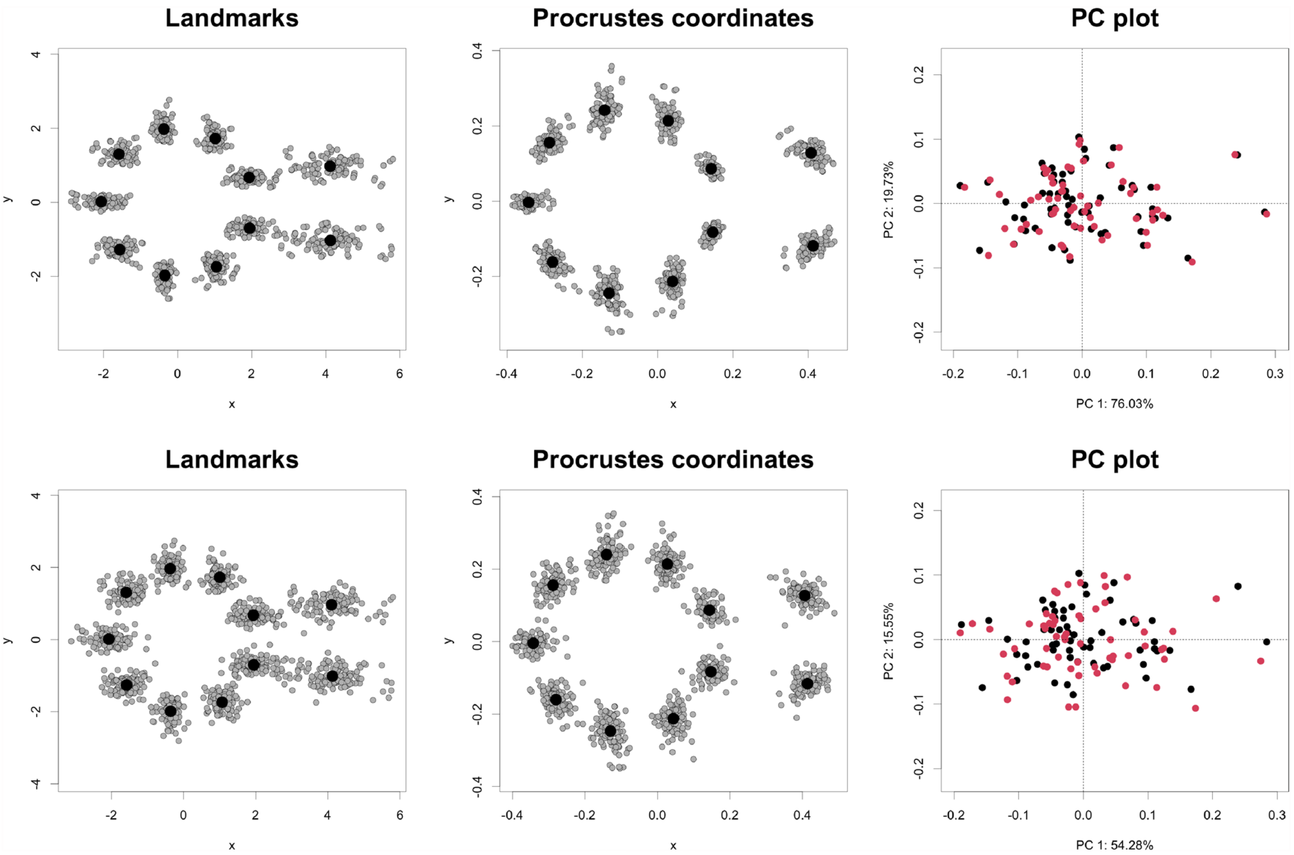

**Fig. 3** Example of simulated research subjects with second replicates (60 research subjects), with different levels of random ME. Top row: small random ME ($\sigma_r = 0.06$). Bottom row: large random ME ($\sigma_r = 0.18$). Left column: raw landmarks. Middle column: Procrustes coordinates, following GPA. Right column: plot of principal component scores, with black dots representing first replicates and red dots representing second replicates

either 100% of the shape variation, or as close to 100% as is reasonable) and using these in place of **Z** in all equations above, would be required. The calculation of $\mathbf{S}_r$ in Eq. (6) with subject-restricted RRPP makes it possible to test for systematic ME via univariate-like (Procrustes) ANOVA or multivariate-ANOVA (MANOVA). These are discussed in more detail below.

## Procrustes ANOVA

Procrustes ANOVA (Goodall, 1991; Klingenberg & McIntyre, 1998) is a term used for analysis that resembles univariate ANOVA, based on the dispersion of linear model estimates in either the shape space, or as we will assume for our discussion here, an orthogonal projection of values into a space tangent to shape space, where Euclidean interpretations of dispersion are appropriate. Four sums of squares (*SS*) calculations are required from four *SSCP* matrices for Procrustes ANOVA; *SS* is the trace (sum of diagonal elements) of these matrices, each calculated as in Eq. (6). Thus, the four *SS* calculations are as follows:

$$\begin{aligned} SS_{total} &= trace(\mathbf{S}_{total}) \\ &= trace\left( \left(\mathbf{Z} - \bar{\mathbf{Z}}\right)^T \left(\mathbf{Z} - \bar{\mathbf{Z}}\right) \right), \end{aligned} \tag{7}$$

$$\begin{aligned} SS_{subject} &= trace(\mathbf{S}_{subject}) \\ &= trace\left( \left(\hat{\mathbf{Z}}^T_{sr|r} - \hat{\mathbf{Z}}_r\right)^T \left(\hat{\mathbf{Z}}^T_{sr|r} - \hat{\mathbf{Z}}_r\right) \right), \end{aligned} \tag{8}$$

$$\begin{aligned} SS_{replicate} &= trace(\mathbf{S}_{replicate}) \\ &= trace\left( \left(\hat{\mathbf{Z}}^T_{sr|s} - \hat{\mathbf{Z}}_s\right)^T \left(\hat{\mathbf{Z}}^T_{sr|s} - \hat{\mathbf{Z}}_s\right) \right), \end{aligned} \tag{9}$$

$$\begin{aligned} SS_{residuals} &= trace(\mathbf{S}_{residuals}) \\ &= trace\left( \left(\mathbf{Z} - \hat{\mathbf{Z}}_{sr}\right)^T \left(\mathbf{Z} - \hat{\mathbf{Z}}_{sr}\right) \right). \end{aligned} \tag{10}$$

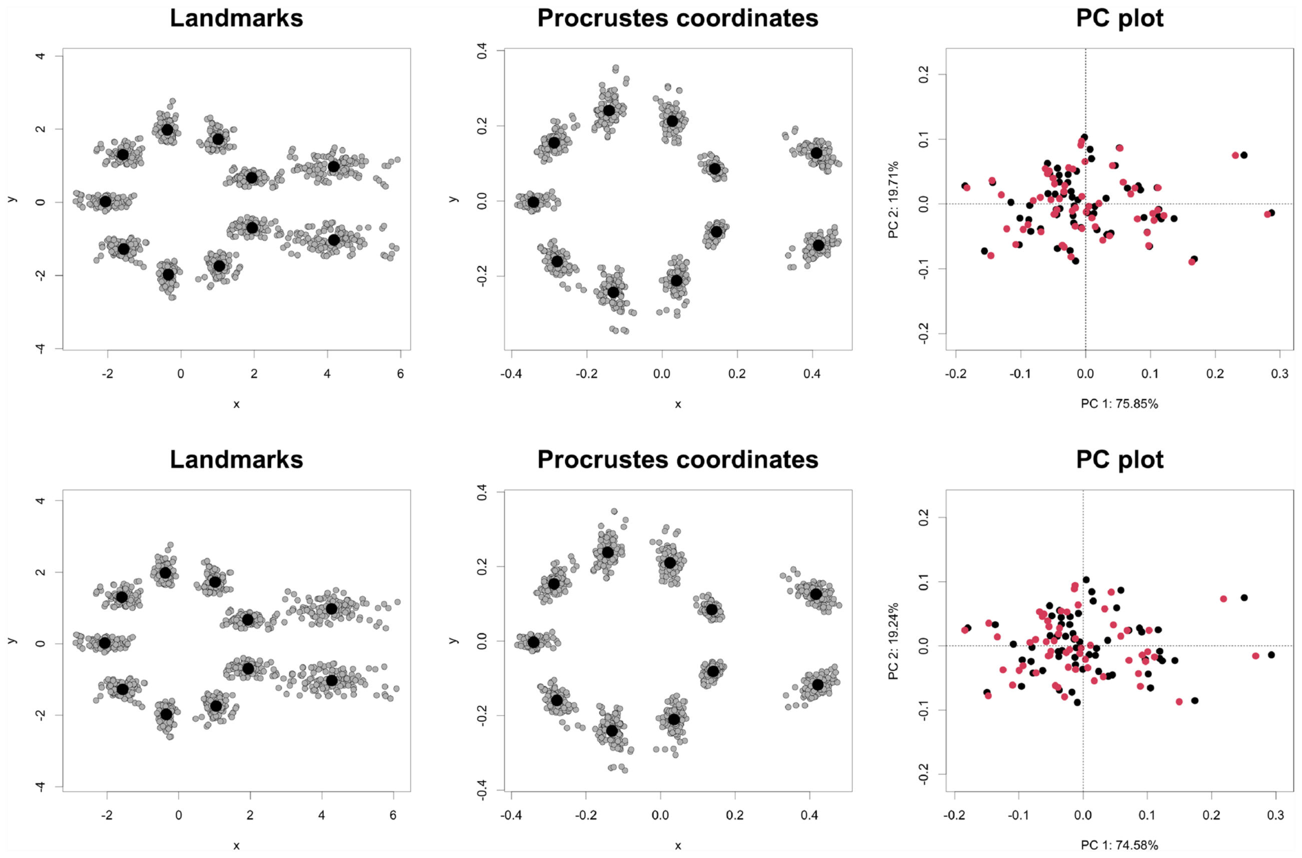

**Fig. 4** Example of simulated research subjects with second replicates (60 research subjects), with digitizing prejudice (systematic ME) and a small amount of random ME. Digitizing prejudice shifted tail landmarks in the second replicate. Top row: small systematic ME. Bottom row: large systematic ME. Left column: raw landmarks. Middle column: Procrustes coordinates, following GPA. Right column: plot of principal component scores, with black dots representing first replicates and red dots representing second replicates

The notation is important to define, precisely. The subscripts, $_{sr|r}$ and $_{sr|s}$ in Eqs. (8) and (9), respectively, indicate that fitted values are obtained for combinations of subjects and replicates, but in different ways. The $|r$ or $|s$ indicates both the restriction for RRPP and estimates of the appropriate null model, for replicates or subjects, respectively. In the formulae above for $SS_{subject}$ and $SS_{replicate}$, $\hat{\mathbf{Z}}_r$ and $\hat{\mathbf{Z}}_s$ are constant across RRPP permutations, respectively, because of the RRPP restriction. There is no specific need to restrict RRPP permutations within replicate to test for subjects, but this provides some consistency for tests. Additionally, it is worth noting that these $SS$ estimates are obtained from $SSCP$ matrices, estimated with a type II $SSCP$ method of estimation. This is important, as it ensures that assessment of systematic ME is conditioned on the subjects chosen for investigation. As such, the mode of restriction and method of estimation are commensurate, even if explicit subject tests are not the principal goal. The final formula, for the calculation of $SS_{residuals}$ does not produce unique values within any RRPP permutation. Because the estimates of $\hat{\mathbf{Z}}_{sr}$ will differ with the different null models used for different terms (whether $sr|r$ or $sr|s$), so too will the residual $SS$. With respect to random ME, it is the version of $SS_{residuals}$ that holds constant subject means that is used in any calculation requiring $SS_{residuals}$.

As with typical ANOVA statistics, the $SS$ values could also be converted to mean-square ($MS$) values by dividing $SS$ by the degrees of freedom, $s-1$ or $r-1$ for subjects and replicates, respectively. $SS_{subjects}$ and $SS_{replicates}$, could also be converted to coefficients of determination as,

$$R^2_{effect} = \frac{SS_{effect}}{SS_{total}}, \tag{11}$$

where *effect* refers to the effect of adding either $s-1$ subject or $r-1$ replicate parameters to their corresponding null models. Henceforth, we replace *replicates* with *SystematicME* and *residuals* with *RandomME* to directly associate $SS$ with these types of ME. The $R^2$ statistics are helpful for understanding the partitioning of the total $SS$ by effects. It is important to realize that with type II $SSCP$s, $SS_{subjects} + SS_{SystematicME} + SS_{RandomME} \neq SS_{total}$, because of the non-sequential addition of model terms. Therefore, the sum of the $R^2$ values is not expected to equal 1.

Generally for ANOVA, an $F$-statistic would also be calculated, and most likely used as a test statistic, for which an empirical sampling distribution could be generated across all RRPP permutations. Although an $F$-statistic would be appropriate as a test statistic in this procedure, we recommend against it for two reasons. First, an $F$-statistic should not convey any interpretation that one might have with a parametric $F$-distribution, both because the data are not univariate despite the calculation of statistics based on distances (Anderson, 2001; Anderson & Walsh, 2013) and the non-independence of observations would call for adjustment of a typical $F$-statistic, if a parametric probability distribution could be invoked (which is unnecessary). Rather, the non-independence of observations is handled by the restricted RRPP strategy, so at best, the distribution of random $F$-statistics could be used to calculate a $P$-value, even though the value of $F$ would not make much sense. Second, a better statistic that would be perfectly rank-correlated with random $F$-statistics across RRPP permutations could be used. We recommend inclusion of this alternative statistic that has appeal as both a descriptive measure and as a test statistic: a systematic ME (signal)-to-random ME (noise) ratio (SNR), which is calculated for the effect of systematic ME as,

$$SNR = \frac{SS_{SystematicME}}{SS_{RandomME}} = F\frac{r-1}{n-s-r}. \tag{12}$$

$SNR$ could be calculated likewise for subject $SS$ and in either case, is a statistic that describes variation in shape associated with a signal relative to variation in random ME (noise). As Eq. (12) illustrates, $SNR$ is also no different as a test statistic than $F$ in a permutation procedure (because $\frac{r-1}{n-s-r}$ would be constant in every random permutation). However, an $F$-statistic would have a varied expectation based on the number of research subjects and replicates, but $SNR$ is a statistic that could more logically be compared across studies. For example, one ME experiment that finds an $SNR$ of 0.5 for systematic ME would elicit more concern than one that finds $SNR = 0.1$.

It might be of interest to also calculate partial coefficients of determination ($\eta^2$) just for $ME$, however, we must realize that $\eta^2_{SystematicME} \neq \frac{SS_{SystematicME}}{SS_{SystematicME}+SS_{RandomME}}$ and $\eta^2_{RandomME} \neq \frac{SS_{RandomME}}{SS_{SystematicME}+SS_{RandomME}}$, because of the type II $SS$ estimation. However, this limitation is easily overcome. By holding constant the effect of research subjects, the residuals from a null model with subjects as the only factor can be subjected to analysis with a single-factor linear model that contains replicate parameters. By doing this, $SS_{\epsilon|subjects_{SystematicME}} + SS_{\epsilon|subjects_{RandomME}} = SS_{\epsilon|subjects_{total}}$, where $\epsilon|subjects$ corresponds to residuals from the single-factor subjects model. Thus,

$$\eta^2_{\epsilon|subjects_{SystematicME}} = \frac{SS_{\epsilon|subjects_{SystematicME}}}{SS_{\epsilon|subjects_{total}}}, \tag{13}$$

and

$$\eta^2_{\epsilon|subjects_{RandomME}} = \frac{SS_{\epsilon|subjects_{RandomME}}}{SS_{\epsilon|subjects_{total}}}, \tag{14}$$

where $SS_{\epsilon|Subjects_{total}} = trace(\epsilon^T_{subjects}\epsilon_{subjects})$, for the matrix of residuals obtained from the single-factor subjects model, $\epsilon_{subjects}$. These descriptive statistics simply convey the portion of systematic and random components of ME in the absence of subject variation. This might be practical if, for example, $R^2_{SystematicME}$ is small but highly significant, because $R^2_{subjects}$ is large, due to sampling disparately shaped subjects.

The $SNR$ and partial $\eta^2$ statistics might seem unnecessarily redundant. Indeed, we would expect that $SNR \approx \frac{\eta^2_{SystematicME}}{\eta^2_{RandomME}}$. Although partial $\eta^2$ statistics are more commonly associated with ANOVA and MANOVA, and $SNR$ statistic might seem like a complicated introduction here, a multivariate generalization of the $SNR$ statistic is more consistent with the basis for MANOVA statistics, which we discuss in more detail, below. Therefore, despite the redundancy, calculating both statistics is helpful.

A $P$-value for the $SNR$ statistic for systematic ME is the probability of finding as large or larger $SNR$, by chance, based on the frequency of outcomes that larger $SNR$ is generated, randomly by RRPP, divided by the number of RRPP permutations. It is worth re-iterating that $R^2_{SystematicME}$ can be misleading as a descriptive statistic. If very great disparity in shape is sampled inherently by the subjects chosen for an evaluation of ME—something a researcher could augment to feel better about the impact of ME in their study—the observed $R^2_{SystematicME}$ might be deceptively small, but the $SNR$ statistic could be large, as it is measured independent of subject variation. Nonetheless, as a test statistic, it remains difficult to adjudicate an $SNR$ statistic without understanding the probability of observing as large of a $SNR$ statistic by chance (the $P$-value). As an effect size, this is a bit problematic, since the same $SNR$ could be either signficant or not significant in two different studies. However, by normalizing the distribution of random $SNR$ statistics, so that $\theta = f(SNR)$, a standardized effect size can be calculated as,

$$Z = \frac{\theta_{observed} - \mu_\theta}{\sigma_\theta}, \tag{15}$$

where, $\mu$ and $\sigma$ are the mean and standard deviation of the normalized distribution, respectively. *Z*-statistics are more reliable for comparison of the effect of systematic ME, both to other sources of variation (more on this below) and systematic ME from other ME experiments. For example, a test of systematic ME might be performed for different configurations associated with different anatomical structures, digitized on the same research specimens, and *Z*-scores compared to ascertain if a digitizing prejudice is found more so for one configuration compared to another.

The statistics calculated for ANOVA can also lend themselves well to calculations of intraclass correlations (Arnqvist & Mårtensson, 1998; Fruciano, 2016), which rather than measuring the amount of ME, provide an effect size for the reliability of research subjects to represent themselves in repeated digitizations, in spite of ME. As will be apparent in the subsequent section, however, reliability can be artificially augmented by simply choosing subjects with quite different shapes. However, compared to previous descriptions of the intraclass correlation for shape data, we provide methods for the calculation of alternative coefficients, which can help reveal systematic ME.

## Intraclass Correlation

The intraclass correlation coefficient ($ICC$) has been proposed previously for use with GM data in studies with repeated digitizations, as a measure of "repeatability" or "reliability", the consistency of research subjects to resemble themselves in repeated digitizations (Arnqvist & Mårtensson, 1998; Fruciano, 2016). $ICC$ has been defined for GM data as,

$$ICC = \frac{E(MS)_A}{E(MS)_W + E(MS)_A}, \tag{16}$$

where $E(MS)$ is the expected mean squares (variance components), and the subscripts $_A$ and $_W$ refer to among-subject and within-subject variance, respectively. Previous descriptions of $ICC$ have asserted that $E(MS)_A = (MS_s - MS_W)/r$ and $E(MS)_W = MS_W$. The within-subject variance, $MS_W$, is calculated as $(SS_r + SS_{residuals})/(n - s)$, for the $n$ total observations, which is the cumulative shape variation within subjects, disregarding the effect of replicates; i.e., it only measures variance among repeated digitizations but is neither concerned with the order of the digitizations nor the classification of digitizations (e.g., unit 1 vs. unit 2). It should be clear that a balanced design is required for $ICC$, as $r$ is part of the calculation. Equation (16) can be thus updated to define $ICC$ based on $MS$ values rather than $E(MS)$ values as,

$$ICC = \frac{MS_s - MS_W}{MS_s + (r-1)MS_W}, \tag{17}$$

as detailed by Liljequist et al. (2019).

By calculating $ICC$ this way, it is clear that if subject variation is large (shapes vary greatly among subjects) and the variation among digitizations within subjects is small, $ICC$ will tend toward a maximum value of 1, indicating good repeatability. It should also be clear that if the expected within-subject shape variation is somewhat constant, despite additional subjects added to the study (adding new subjects does not change the expected variation between digitizations, as a practice), then $ICC$ can be inflated by merely sampling a more disparate representation of subject shapes. Because the within-subject variance does not focus on replicate assignment, there is no accounting for systematic ME, rather, ME, whether systematic or random, is only a measurement of imprecision, $MS_W$, with respect to subject variation. However, $ICC$ can be updated to better evaluate the tendency for systematic ME due to digitizing prejudice.

Liljequist et al. (2019) presented two alternative $ICC$ calculations that would not change from the former $ICC$ if ME was 100% random ME. The first calculation is,

$$ICC_A = \frac{MS_s - MS_{residuals}}{MS_s + (r-1)MS_{residuals} + r/s(MS_r - MS_{residuals})}, \tag{18}$$

which updates $ICC$ if absolute agreement between different digitizations is desired. $MS_r$ is the estimated variance due to replicates (systematic ME) and $MS_{residuals}$ is the estimated residual variance (random ME). The second calculation is,

$$ICC_C = \frac{MS_s - MS_{residuals}}{MS_s + (r-1)MS_{residuals}}, \tag{19}$$

which updates $ICC$ to focus on the consistency of repeated digitizations. Careful examination of the three formulae in Eqs. (17), (18), and (19), illustrates that $MS_W$ can be partitioned into $MS_r$ and $MS_{residuals}$ but if there is no systematic ME, then $MS_r = 0$, $MS_W = MS_{residuals}$, and the three $ICC$ values converge. $ICC_A$ calculates a weighted average of $MS_W$ in the denominator and $ICC_C$ excludes variation due to systematic ME. If these $ICC$ values diverge, systematic ME can be implicated. It would be challenging to associate a particular level of disparity between $ICC$ values with an interpretable amount of systematic ME. Any $ICC$ value measured this way is based on dispersion in perhaps many dimensions, and the number of subjects or number of variables might affect the $ICC$ values. However, if a test of systematic ME finds significant systematic ME, disagreement among the $ICC$ values should be apparent.

Both ANOVA and $ICC$ calculations performed this way focus on the dispersion of shapes among and within subjects, and because distances of vectors are univariate despite the number of dimensions in which they are measured, these analyses are univariate solutions for multivariate problems.

Statistical tests are not a concern if based on RRPP, since a parametric probability density function is not required to obtain *P*-values. However, there may be cases where a fully multivariate analysis that focuses on the covariances among landmarks is desired. The analyses above can be generalized with eigenanalysis for such cases.

## Multivariate Generalizations and Visualizations

The *SNR* statistic introduced above is a useful statistic because it has a multivariate generalization that is commonly used in MANOVA:

$$\mathbf{\Phi}_{SNR} = \mathbf{S}^{-1}_{RandomME}\mathbf{S}_{SystematicME}, \tag{20}$$

where $\mathbf{S}$ is an *SSCP* matrix and $\mathbf{\Phi}$ is the multivariate generalization of the ratio, *SNR*. Various MANOVA statistics can be calculated from eigenanalysis of $\mathbf{S}^{-1}_{RandomME}\mathbf{S}_{SystematicME}$, the simplest being Roy's maximum root, the largest eigenvalue obtained from eigenanalysis. With respect to MANOVA, a null hypothesis for the signal evaluated relative to noise is typically tested with an *F*-distribution proxy, which is not appropriate here for the same reasons *F*-statistics are discouraged with Procrustes ANOVA. Rather, a sampling distribution of Roy's maximum root can be generated with the same RRPP strategy used for Procrustes ANOVA[5] *P*-values for Roy's maximum root are calculated as the percentile of observed statistics in their corresponding sampling distributions and effect sizes are calculated as in Eq. (15).

With respect to an ordination plot of *SNR*, mean-centered Procrustes coordinates can be projected onto the eigenvectors of $\mathbf{S}^{-1}_{RandomME}\mathbf{S}_{SystematicME}$, which have a maximum number of $\min(s-1, r-1)$, to visualize systematic ME patterns. For example, if two replicates are used in the ME experiment, points will fall on one line. The paired points for subjects will indicate if there is a consistent left-right pairing, which would be indicative of systematic ME. More than two replicates increases the dimensions in which systematic ME can manifest, but the principle is the same; systematic ME is a consistent divergence of points in such a plot. Multivariate SNR plots will reveal, perhaps better than PC plots, the pattern of systematic ME, as the orientation of the vectors is specific to systematic ME, relative to random ME. This could be helpful compared to a PC plot, where other factors can influence the rotation of eigenvectors and thus, the dispersion of points is a space reduced to the first 2–3 vectors. It might be of interest to standardize the systematic error-to-noise ratio as, $\mathbf{S}^{-1/2}_{RandomME}\mathbf{S}_{SystematicME}\mathbf{S}^{-1/2}_{RandomME}$, which yields a symmetric matrix that produces orthogonal eigenvectors. Although eigenanalysis will produce the same eigenvalues, their distribution will be different (see Bookstein and Mitteroecker (2014) for details), so caution would be needed to assure the order of eigenvectors is appropriate. The concern for orthogonal vectors is not strongly needed, however, as the points in these projections should not be interpreted as shape variation in the space tangent to shape space. The plots simply reveal the consistency of systematic ME relative to noise (random ME).

An example of how *SNR* plots can be used is shown in Fig. 5. In these example plots, the same digitizing prejudice is applied to two sets of data, the second also applying group difference shifts (tail lengthening) to the first set. An interesting attribute to this example is that systematic ME seems to differ between the two data sets, even though the same digitizing prejudice was simulated. It is difficult to fully appreciate the utility of the *SNR* plots in this example, but this is because the group differences in shape that were also simulated obscure interpretation. We will return to this issue after considering how *ICC* statistics can also be generalized.

The equations for *ICC* can also be generalized and eigenanalysis performed in a similar manner. The *ICC* generalizations are as follows:

$$\mathbf{\Phi}_{ICC} = (\mathbf{MS}_s - \mathbf{MS}_W)^{-1}(\mathbf{MS}_s + (r-1)\mathbf{MS}_W); \tag{21}$$

$$\begin{aligned}\mathbf{\Phi}_{ICC_A} =& (\mathbf{MS}_s - \mathbf{MS}_{Residuals})^{-1}(\mathbf{MS}_s + (r-1)\mathbf{MS}_{residuals} \\ &+ r/s(\mathbf{MS}_r - \mathbf{MS}_{residuals}));\end{aligned} \tag{22}$$

and

$$\begin{aligned}\mathbf{\Phi}_{ICC_C} =& (\mathbf{MS}_s - \mathbf{MS}_{Residuals})^{-1}(\mathbf{MS}_s \\ &+ (r-1)\mathbf{MS}_{residuals});\end{aligned} \tag{23}$$

where, $\mathbf{MS}$ is the covariance matrix form of *MS* and $\mathbf{\Phi}$ is the multivariate generalization of a ratio, for *ICC*. However, as a generalization, it is not clear how useful $\mathbf{\Phi}$ matrices are, since the same covariance matrices ($\mathbf{MS}$) are used multiple times in the calculation of these matrix generalizations, meaning they are singular (not positive-definite). Eigenanalyses of these matrices might be helpful, producing a distribution of eigenvalues that are *ICC* scores for corresponding eigenvectors, with *ICC* maximized in the first vector, but this value will be most likely inflated compared to an *ICC* statistic based on dispersion, making it challenging to use as descriptive statistic. A generalized *ICC* value can be found as $\prod |\lambda_i|$ for the distribution of eigenvalues$(\lambda_i)$ (sensu Bookstein & Mitterœcker 2014), but because the matrices are

[5] It is important to realize that the same strategy (within-subject RRPP) is used to obtain sampling distributions, whether Roy's maximum root or *SNR* are used as test statistics. Alternative statistics could also be used. Generally, *P*-values and *Z*-scores will be similar in terms of interpretation but not perfectly rank-correlated unless they are linear transformations of each other, like *SNR* and *F*. However, alternative sampling distribution strategies are not needed if different statistics are used.

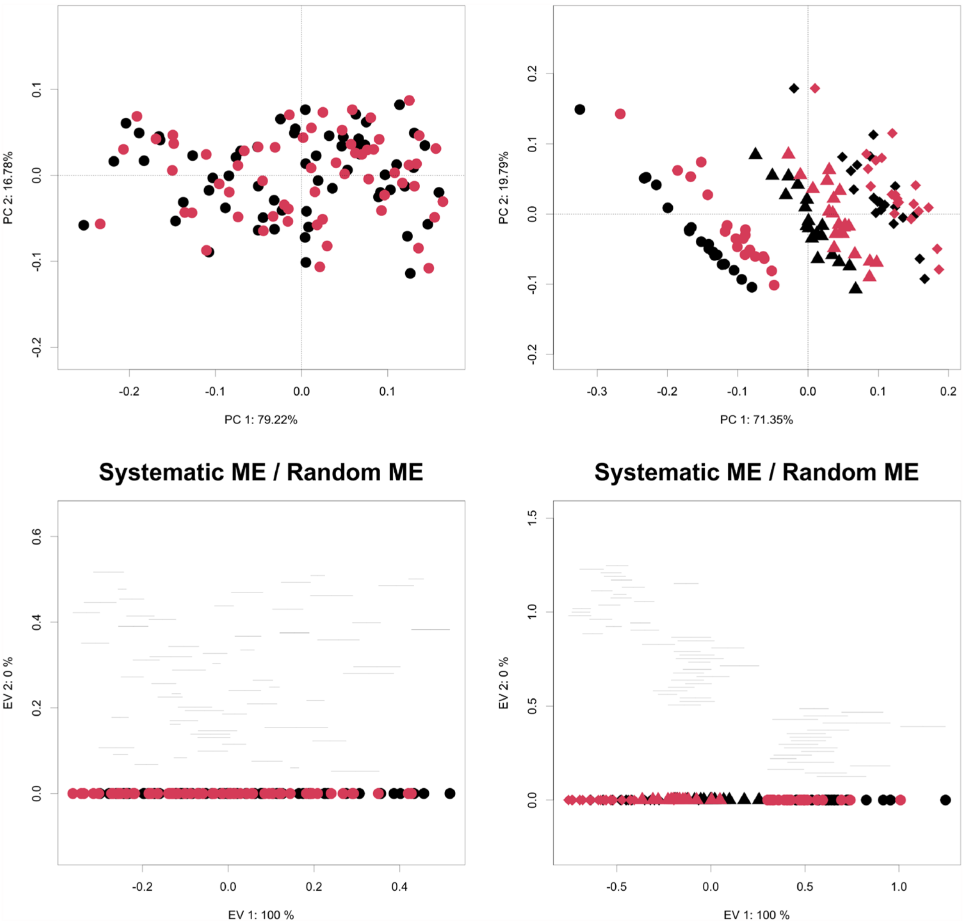

**Fig. 5** Principal component plots (top row) and *SNR* eigenvector plots (bottom row) for two examples of systematic ME: no group differences in shape (left) and obvious group differences in shape (right). The same, per-subject digitizing prejudice was simulated for both data sets. Points are colored by replicates in each plot and different symbols correspond to different groups. The *SNR* eigenvector plots contain vectors above points, showing the connection of subject points in the plot. The scale of the *SNR* axes are different, with group differences appearing to make the amount of systmatic ME look smaller than it actually is

singular, the generalized statistic is certain to be 0. However, we recommend examining the cumulative product by eigenvector, i.e., $\prod_{i=1}^{i=j} |\lambda_i|$ for eigenvalues, $\lambda_1, \lambda_2, ..., \lambda_j$, allowing the generalized *ICC* statistic to be examined before it attenuates. It will be challenging to garner an appreciation for the values, themselves, but it should still be possible to evaluate the divergence between agreement and consistency of *ICC* values, at least in the first few vectors. For the concerns we addressed with these matrices, we do not recommend projection of mean-centered Procrustes coordinates on these vectors for graphical results. The *SNR* eigenvectors explicitly maximize systematic ME relative to random ME in the first vector, so a graphical representation cannot be improved with *ICC* ordination plots.

We have indicated multiple times that *ICC* values could be improved by sampling disparately-shaped subjects, and therefore, caution against reliance on these statistics is warranted. However, it is worth considering how sampling research subjects from groups with known shape differences can obfuscate interpretations of ME. Just as disparately shaped groups of subjects might be separated in a PC plot, so too might they be separated in a *SNR* eigenvector plot, effectively reducing the length of vectors between subject replicates compared to the spread of subject shapes in the plot (see Fig. 5, for example). Additionally, sampling individuals of both sex from sexually dimorphic species, or sampling several individuals from vastly differently shaped species can result in rather clustered sets of points in an *SNR* eigenvector plot, making interpretation of systematic ME challenging. Although this might seem like a sampling problem, it is perhaps one to embrace, because it is possible that systematic ME as a result of digitizing prejudice is not homogeneous across all specimens; digitizing prejudice might differ among groups of specimens. Although previous ME analytical strategies have focused on evaluating the amount of ME relative to subject variation, variation in ME associated with different groups or strata sampled along with subjects have not been explored, to the best of our knowledge. We argue, however, including potential group differences should be a welcomed analytical consideration, and can be accomplished with simply adding a grouping factor to analyses and accounting for the grouping factor in calculation of *ICC* or generation of *SNR* eigenvector plots.

## Accounting for Group Differences in the Analysis of ME

It is not unreasonable that the subjects used in a GM-ME experiment come from groups with different shapes (like species). It is also not unreasonable—rather, recommended—that a GM-ME experiment includes disparately-shaped research subjects, so that any pattern of systematic ME that might pertain to research subjects of a particular group can be recognized. All the statistics and analysis presented thus far would not be easily capable of revealing varied systematic ME by groups, unless data are subsetted to different groups for analysis, a practice that is neither needed nor recommended.

If it is known before analysis that subjects are sampled from different groups (as in Fig. 5), a grouping factor can be included in all analyses. The subject factor subsumes the effect of group, when holding subject variation constant, as research subjects are unique to groups. However, it is possible to test a systematic ME × group interaction as part of the analysis. By using type II *SSCP*, a test of this interaction would hold constant the effects of both subjects and replicates, meaning variation that would be normally considered random ME could be parsed into a systematic ME × group component and smaller random ME component. For calculating *ICC*, the group effect can be removed from the subject variation by using the residual shapes from groups to estimate the subject variation (tantamount to centering all group means at the origin). This step can also assist *SNR* eigenvector plots by removing the scatter due to group differences from interpretation of paired differences in shape among subjects.

For the example in Fig. 5, Table 2 provides most of the statistics discussed (excluding multivariate *ICC* generalized values, which would have to be considered by component) and Fig. 6 provides an updated *SNR* plot for the set of data that have inherent group differences. These data had a consistent digitizing prejudice (tail lengthening in one replicate) applied to all research subjects, so no group-specific digitizing prejudice was made.

We start by summarizing results for the data set without group structure, in which a consistent digitizing prejudice was simulated. The systematic ME $R^2$ was the same, regardless of whether a group factor was included in the linear model, and it was small ($R^2 = 0.0076$). Random *ME* was also small and together, it might not be alarming that only $R^2 \approx 0.028$ of the shape variation was due to ME. However, systematic ME was highly significant and had a fairly large effect, whether using *SNR* or Roy's maximum root ($Z_{SNR} = 5.9144$; $Z_{Roy} = 7.9485$, $P = 0.001$ in both cases). Approximately 27% of the ME was systematic, resulting in a *SNR* of 0.3788, which changed little by adding groups (0.3926). Although all *ICC* values were $\approx 0.94$ or higher, there was a little disparity between $ICC_A$ and $ICC_C$, perhaps indicative of a systematic ME signal, but not as obvious as the ANOVA results. These values were little changed by adjusting for groups, meaning the *ICC* values were not excessively augmented by sampling subjects from different groups.

By contrast, the same digitizing prejudice simulated for subjects that differed much more in shape because they were sampled from differently shaped groups resulted in greater systematic ME, overall. Without considering group differences in the analysis, the *SNR* rose to 1.8600; random ME was similar as in the previous data so this value indicates an increase in systematic ME. Effect sizes (*Z*-scores) decreased despite the increase in *SNR*, but *ICC* values changed little. However, including a group factor in the analysis added a highly significant and large Systematic ME × groups effect ($Z_{SNR} = 5.7285$; $Z_{Roy} = 4.1256$, $P = 0.001$ in both cases), increasing *SNR* (to 2.1894). Interestingly, adding group effects substantially increased the systematic ME effect size, just for MANOVA (from 10.4763 to 13.5606) and the effect was more pronounced for the systematic ME effect for MANOVA, although the systematic ME × groups effect was larger for ANOVA.

*ICC* values were slightly reduced for *ICC* and $ICC_A$ when including group effects, reflecting the tendency for disparately shaped groups to inflate subject variation. The disparity between $ICC_A$ and $ICC_C$ was also larger than for the data set without group structure, suggesting the systematic ME from the same digitizing prejudice was larger, which was confirmed with ANOVA and MANOVA.

At first blush, it might be disheartening that an analysis would find both strong systematic ME and strong systematic ME × group effects for a consistent digitizing prejudice, irrespective of group. However, this result is not surprising. The digitizing prejudice was made by a shift in tail landmarks, regardless of whether subjects were sampled from short-tailed or long-tailed groups. The same shift in an individual from a short-tailed species will more profoundly increase the relative tail size than the same shift in an individual from a long-tailed species. This example elucidates what should be a standard principle: digitizing prejudice does not translate to equitable systematic ME; the choice of subjects matters. This example also revealed that a digitizing prejudice in the direction of group differences can augment or retard estimated group shape differences. Not accounting for group in the ME analysis might mean overlooking this phenomenon. A comprehensive evaluation of the methods in the ME analysis in this example is explored with simulation experiments for the six scenarios in Table 1.

**Table 2** Example of results obtained from four different analyses of ME, for two data sets (Fig. 6)

| Statistic | ME1 | $\text{ME1}_g$ | ME2 | $\text{ME2}_g$ |
|---|---|---|---|---|
| $R^2$, systematic ME | 0.0076 | 0.0076 | 0.0207 | 0.0207 |
| $R^2$, systematic ME × groups | – | 7e-04 | – | 0.0017 |
| $R^2$, random ME | 0.02 | 0.0193 | 0.0112 | 0.0095 |
| $\eta^2$, systematic ME | 0.2747 | 0.2747 | 0.6503 | 0.6503 |
| $\eta^2$, systematic ME × groups | – | 0.0254 | – | 0.0526 |
| $\eta^2$, random ME | 0.7253 | 0.6998 | 0.3497 | 0.297 |
| *SNR*, systematic ME | 0.3788 | 0.3926 | 1.86 | 2.1894 |
| *SNR*, systematic ME × groups | – | 0.0363 | – | 0.1771 |
| $Z_{SNR}$, systematic ME | **5.9144** | **5.9** | **4.0048** | **4.0274** |
| $Z_{SNR}$, systematic ME × groups | – | 0.1779 | – | **5.7285** |
| Roy's $\lambda_{max}$, systematic ME | 8.9486 | 8.9531 | 11.7138 | 35.8667 |
| Roy's $\lambda_{max}$, systematic ME *imes* groups | – | 0.5107 | – | 2.7862 |
| $Z_{Roy}$, systematic ME | **7.9485** | **7.6324** | **10.4763** | **13.5606** |
| $Z_{Roy}$, systematic ME × groups | – | −0.5596 | – | **4.2156** |
| *ICC* | 0.9457 | 0.9457 | 0.9372 | 0.9372 |
| $ICC_A$ | 0.9461 | 0.9461 | 0.9385 | 0.9385 |
| $ICC_C$ | 0.9597 | 0.9597 | 0.9772 | 0.9772 |
| *ICC*, group-adjusted | 0.9457 | 0.9444 | 0.9372 | 0.8354 |
| $ICC_A$, group-adjusted | 0.9461 | 0.9448 | 0.9385 | 0.8437 |
| $ICC_C$, group-adjusted | 0.9597 | 0.9587 | 0.9772 | 0.9382 |

One data set has no inherent group differences in shape (even if there is a presumptive group factor); the other data set has inherent difference in shape (like species differences). Table columns correspond to: measurement error analysis for data set 1 (no group structure), not including groups as a factor in the analysis (ME1); measurement error analysis for data set 1, including groups as a factor in the analysis (ME1g); measurement error analysis for data set 2 (group structure simulated), not including groups as a factor in the analysis (ME2); and measurement error analysis for data set 2, including groups as a factor in the analysis (ME2g). Values in bold correspond to significant test results ($\alpha = 0.05$), based on RRPP with 1000 random permutations

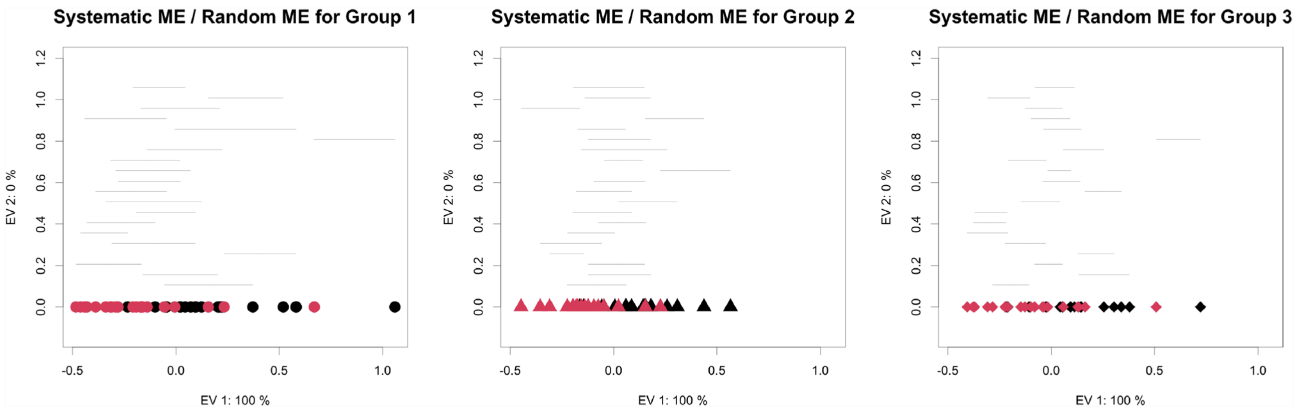

**Fig. 6** For the same data with group difference in Fig. 5, plots of subject scores on the *SNR* eigenvectors for data that removes group shape differences. Three plots are shown for subjects, by groups, to facilitate an understanding that systmatic ME tends to be greater for one group

## Statistical Properties Assessed from Simulation Experiments

Simulation experiments were performed for every example in the *Motivating examples* section, above. In every experiment, $\mathbf{H}_i$ and $\mathbf{R}_i$ were randomly simulated for every research subject in every run, varied by the amount of inter-subject shape variation and random ME, respectively. The experiments varied the composition of elements in $\mathbf{G}_j$ and $\mathbf{S}_j$ in a non-random, specific way, based on experiment objectives. This model allowed us to collectively consider the six experiments for the six examples, described above. We used 20 research subjects within three groups for all experiments (60 research subjects, total). Landmark configurations contained 11 landmarks, but only two of which were changed in $\mathbf{G}_j$ or $\mathbf{S}_j$. A total of 500 simulation runs were performed in all cases, and 1000 RRPP permutations were performed for each ME analysis, for both raw landmarks and Procrustes coordinates, following GPA, within every run. The $P$-value was recorded for the effects, systematic ME and systematic ME:groups (if appropriate), and the portion of cases a null hypothesis of no systematic ME was rejected at a significance level of $\alpha = 0.05$ (if $P < \alpha$) was recorded. For evaluation of type I error rates, 95% confidence intervals for a true rejection rate of $\alpha = 0.05$ were calculated from a binomial probability distribution, *sensu* Anderson and Walsh (2013; Supplementary Material), using the `prop.test` function of `R` (R Core Team, 2023).

The results from simulation experiments are too numerous to present comprehensively, but are available in the Supplementary Material, in their entirety. The Table 3 below summarizes the results in practical terms. There are also `R` scripts in the Supplementary Material that can be used to replicate simulation experiments.

Summarizing across experiments, it is clear that *ICC* statistics are not that valuable for detecting the relative portions of systematic and random components of ME, and whether systematic ME varies among groups; that *SNR* statistics and plots are valuable tools for understanding how ME manifests in shape data; that GPA can actually buffer the effects of a digitizing prejudice; that ANOVA and MANOVA tend to offer consistent interpretation, although the effect sizes can vary a little; and finally, one should not assume a consistent digitizing prejudice results in consistent systematic ME, especially if there are subjects sampled from disparately shaped groups. Type I error rates were universally appropriate, regardless of the amount of random ME or whether there were group shape differences, whether GPA was performed, and whether using ANOVA or MANOVA. The analytical paradigm had good statistical power, regardless of data type, for detecting effects that were simulated.

As a more comprehensive demonstration of the methods presented in this paper, an empirical example is more practical. We next re-evaluate a previously published example below with the techniques we have outlined, discussing the strengths and weaknesses of each approach.

## Empirical Example: Reanalysis of Fruciano et al. (2017)

To illustrate the utility of the procedures developed here, we performed a reanalysis of the empirical dataset found in Fruciano et al. (2017). The original study was conducted to examine the effects of combining landmark data from multiple observers and scanning devices. The dataset consisted of three-dimensional landmark data obtained from the crania of 23 marsupial species. Surface scans were obtained from each cranium using three different scanning technologies (devices), and each scan was digitized by two different observers, who recorded the locations of 31 three-dimensional landmarks on each (seven landmarks were subsequently removed following initial inspection). Thus, the final dataset contained 138 landmark configurations, comprising six replicates (2 observers × 3 devices) for each of 23 species, with 24 landmarks digitized on each. Fruciano et al. (2017) correctly noted that this experimental design had the potential for ME to be introduced at several levels, and conducted a series of analyses to inspect this possibility. Two of their analyses are most relevant here. First, they used an analysis of variance on the Procrustes-aligned coordinates to extract variance components (species, side, species × side, device, observer), and to calculate $R^2$ values for each model effect. The $R^2$ values for `device` and `observer` were then treated as estimates of ME for comparison with other model effects. Second, they conducted tests of 'bias' on subsets of the data using a series of pairwise comparisons (e.g., among devices for the same observer, and between observers for the same device). Here they performed separate Procrustes alignments for each subset of data, and used a permutation test to evaluate pairwise group differences (Fruciano et al., 2017). Significant differences between groups were treated as evidence of systematic digitizing bias between observers or devices.

The analytical approach employed by Fruciano et al. (2017) was not fully capable of interrogating the effects of ME in this dataset. One reason is that they utilized a standard symmetry-based ANOVA design (as found in Klingenberg, 2010), which only described overall ME for each specified error term. That is, the procedure implemented by Fruciano et al. (2017) identified variation among devices and among observers, but did not parse ME into its random and systematic components, nor did it consider any group-specific systematic ME. In addition, the pairwise comparisons among groups that they calculated were obtained from separate Procrustes alignments on different subsets of the data. As such, the resulting summary values were incomparable across tests, rendering any synthetic generalizations based on them inconclusive. Our reanalysis below provides additional insights regarding the nature of ME in this dataset that were not easy to consider prior to the methodological development in this paper.

For our reanalysis, we first performed a Procrustes alignment of all specimens, and following Fruciano et al. (2017) extracted the symmetric component of shape variation (Fig. 7A). We then conducted a principal component analysis

**Table 3** Conclusions from simulation experiments

| Experiment | Purpose | Conclusions |
|---|---|---|
| 1 | Effect of digitizing noise on systematic ME | 1. Increasing random ME had no observable effect on ANOVA or MANOVA effect sizes or *SNR* statistics<br>2. Increasing random ME reduced dispersion-based *ICC* scores, more so for Procrustes coordinates than raw landmarks. *ICC*, $ICC_A$, and $ICC_C$ were all consistent, irrespective of the amount of random ME or whether GPA was performed<br>3. Dispersion-based *ICC* scores could be reassuringly large despite a large amount of random ME, provided subjects were different in shape<br>4. Multivariate *ICC* eigenvector scores were difficult to interpret, especially because $ICC_C$ could become negative (with large ME or GPA performed), owing to singularities imposed by matrix products. *ICC* were nearly all equal to 1 in the first few components, regardless of the amount of random ME or whether GPA was performed<br>5. *SNR* plots did not reveal any patterns<br>6. Type I error rates were appropriate, regardless of the amount of random ME, or whether GPA was performed |
| 2 | Effect of sampling from differently shaped groups on systematic ME | 1. Increasing group differences tended not to induce meaningful changes in *SNR*, or *Z*-scores for either systematic ME or the systematic ME by group interaction of ANOVA, or the *Z*-scores of MANOVA, regardless of the amount of group difference or whether GPA was performed<br>2. Type I error rates were appropriate regardless of the amount of group shape difference, whether GPA was performed, or whether ANOVA or MANOVA was used<br>3. Dispersion-based *ICC* statistics were consistent among the three types and increased as group differences increased. These stats were mitigated by adjusting for group differences, but were still reassuringly (and perhaps, unreasonably) large<br>4. Multivariate *ICC* stats were again difficult to interpret. The scores were nearly 1 in all cases in the first component. In lower components, the same trends as the dispersion stats seemed to take place, unless *ICC* scores were negative<br>5. $ICC_A$ and $ICC_C$ stats tended to be consistent, when adjusting for groups |

**Table 3** (continued)

| Experiment | Purpose | Conclusions |
|---|---|---|
| 3 | Effect of the same digitizing prejudice applied to different groups of subjects | 1. When there were no group shape differences, small systematic ME did not tend to produce a significant systematic ME effect, but large systematic ME did. No amount of systematic ME tended to induce a significant systematic ME:group effect. This was true for both ANOVA and MANOVA<br>2. When there were group shape differences, the same tendencies were observed for systematic ME effects as with no group shape differences, but larger systematic ME also induced significant systematic ME:group effects, for Procrustes coordinate data (not for raw landmarks)<br>3. The statistical power associated with detecting systematic ME increased fast with increased distizing prejudice, regardless of method or data type<br>4. The statistical power associated with detecting systematic ME:group increased more moderately, but only for Procrustes coordinate data, and more so for ANOVA than MANOVA<br>5. *ICC* stats followed the same trends as before with these exceptions: disparity between $ICC_A$ and $ICC_C$ scores increased with the amount of systematic ME (although all scores were large, regardless); and, larger group shape differences exacerbated the disparity<br>6. *SNR* plots revealed that a larger difference between shapes in digitizations could be found for one group versus another, for Procrustes coordinates, for the same digitizing prejudice.—A consistent digitizing prejudice should not be expected to produce consistent measurement error if speciemns are sampled from disparately shaped groups |
| 4 | Effect of a digitizing prejudice applied to one group, in the direction of group shape differences | 1. ANOVA and MANOVA results were consistent with Experiment 3 with one exception: sytematic ME:group effects were larger than sytematic ME effects. Nevertheless, a digitizing prejudice applied to only one group of subjects induced both systematic ME and systematic ME:group effects, both increasing with the size of the digitizing prejudice<br>2. Increasing group shape difference did not have any appreciable change in the statistical power curves, even though applying the digitizing prejudice to only one group would impact the shape differences among groups, if averaged over replicates<br>3. The statistical power increased at a slightly faster rate for the systematic ME:group effect than the systematic ME effect, also more so for raw landmarks than Procrustes coordinates, and more so for ANOVA than MANOVA<br>4. There were no appreciable differences between *ICC* scores from Experiments 3 and 4, despite large differences between ANOVA and MANOVA effect sizes. However, the disparity between $ICC_A$ and $ICC_C$ scores was reduced, suggesting systematic ME was of little concern<br>5. *SNR* plots demonstrated a good ability to detect the digitizing prejudice localized to one group<br>6. The ANOVA $\eta^2$ and *SNR* statistics remained rather consistent, despite changing group shape differences, and highlighted well the tendency for digitizing prejudice to be localized to one group.—Collectively, the results in this experiment demonstrate that GPA can buffer systematic error from a digitizing prejudice, and ANOVA or MANOVA can reveal the extent to which a digitizing prejudice is varied among different groups of organisms |

**Table 3** (continued)

| Experiment | Purpose | Conclusions |
|---|---|---|
| 5 | Effect of a digitizing prejudice applied to one group, in the direction opposite of group shape differences | 1. All conclusions from Experiment 4 are exactly the same for Experiment 5.—Collectively, the results in this experiment demonstrate that digitizing prejudices in a direction of group shape differences—whether increasing or decreasing shape differences – have similar analytical results, and can confirm the group to which the digitizing prejudice was applied |
| 6 | Effect of a digitizing prejudice applied to one group, in a direction orthogonal to group shape differences | 1. Most conclusions in Experiments 4 and 5 were retained in Experiment 6 except for three alternative conclusions: the systematic ME:group effects were large but only slightly larger than systematic ME effects, regardless of data type or method; the *SNR* plot continued to reveal the greater systematic ME in one group, despite less ability for digitizing prejudice to change shape differences among groups; and, the *ICC* stats became more consistent (between $ICC_A$ and $ICC_C$), suggesting digitizing prejudice was not a problem.—Collectively, these results elucidate that a digitizing prejudice that does not augment or retard group shape differences is still detectable, and the amount of systematic ME applied to one group was still obvious in *SNR* plots. These results are not available with *ICC* statistics |

to inspect patterns of shape variation among species in morphospace, and to visually discern whether device differences or observer differences were evident. Next we performed a series of measurement error analyses, using the analytical procedures developed in this paper. Our first analysis extracted the overall components of systematic and random ME by treating the six repeated observations for each species (2 observers × 3 devices) as within-subject replicates. Next we performed analyses that included *clade* as a grouping factor, in which different subjects could be assigned to subclade A, subclade B, or a one-species outgroup. (This factor was not included in measurement error analyses by Fruciano et al. (2017) but was important for evaluating the effect of measurement error on estimates of phylogenetic signal.) The goal in the second analysis was to consider whether random ME as estimated in the first analysis could be cloaked as group-specific systematic ME. 10,000 within-subject RRPP permutations were used for these analyses. The test of the among-subject effect restricted RRPP permutations within replicates for consistency.

Finally, we examined the extent to which the direction of systematic ME aligned with other aspects of biological signal in this dataset, by examining the correlation of principal vectors for different effects. The biological signals that could be considered were the effects of species or clade, which are inherently correlated as clades comprise species within them (a species or subject effect inherently includes a clade effect). Either a species effect or clade effect is constant across RRPP permutations that sample within subjects, as subjects are species, in this case. Therefore, the principal eigenvector of the sums of squares and cross-products (SSCP) matrix for either species or clade is unchanged across permutations. Adding parameters for observers, devices, or observer × device interactions will result in different principal eigenvectors for each SSCP across RRPP permutations, as replicates are randomized within species. The same RRPP procedure used to evaluate components of systematic and random measurement error allows a permutation test of vector correlations between biological signal and sources of systematic ME. For these tests, a null hypothesis of vector independence would be rejected if the correlation between vectors—the cross-product between unitized eigenvectors—is larger than expected by chance (i.e., the angle between vectors, which is the arccosine of the vector correlation, is smaller than expected by chance). We performed permutation tests based on the 10,000 RRPP fits used in the previous analysis (not including clade as a factor that interacts with replicates), parsing the parameters for replicates into operator, device, and interaction parameters, in order to calculate separate SSCP matrices, and thus, eigenvectors.

For all tests, a level of significance of $\alpha = 0.05$ was used. The functions, `measurement.error` and `plot.measurement.error` from the `RRPP R` package (Collyer & Adams, 2023) were employed, along with `gpagen` in the `geomorph R` package to perform GPA (Baken et al., 2021). We also used the functions, `focusMEonSubjects`, `interSubVar` and `plot.interSubVar` from the `RRPP R` package to evaluate how ME for specific subjects might cause concerns for estimates of species shapes.

*Empirical Results* The principal component plot (Fig. 7B) was identical to that presented by Fruciano et al. (2017) (Fig. S4), and revealed that replicate observations within species were generally tightly clustered compared to interspecies variation. The visual evidence was also supported by

**Table 4** Analysis of variance tables evaluating random and systematic components of measurement error, for the empirical example

| | *Df* | $R^2$ | $\eta^2$ | *SNR* | *Z* | *P* |
|---|---|---|---|---|---|---|
| A: Analysis without clade effect | | | | | | |
| Subjects | 22 | 0.9658 | | 33.3843 | 20.0540 | 0.0001 |
| Systematic ME | 5 | 0.0053 | 0.1551 | 0.1835 | 7.5934 | 0.0001 |
| Random ME | 110 | 0.0289 | 0.8449 | | | |
| Total | 137 | | | | | |
| B: Analysis with clade effect | | | | | | |
| Subjects | 22 | 0.7082 | | 31.5462 | 23.4077 | 0.0001 |
| Systematic ME | 5 | 0.0053 | 0.1551 | 0.2365 | 7.7152 | 0.0001 |
| Systematic ME:groups | 10 | 0.0065 | 0.1892 | 0.2885 | 2.0120 | 0.0216 |
| Random ME | 100 | 0.0225 | 0.6557 | | | |
| Total | 137 | | | | | |

traditional Procrustes ANOVA statistics. For instance, 96.6% of the total variation was described by among-species differences, but only 3.4% of the variation was attributable to ME (Table 4). Additionally, there was high repeatability across replicate observations (*ICC* > 0.960). (The three *ICC* statistics were also consistent, and the multivariate generalized *ICC* statistic was 0.9996 for each of the three statistic types in the first component of each generalized matrix.) Nevertheless, using the novel statistics and their evaluation, as presented in this paper, revealed some reason for concern. First, 15.5% of the ME was systematic ME, which was significant and displayed a large effect, whether for the univariate analysis of dispersion ($Z = 7.4545; P = 0.0001$; Table 4) or the multivariate analysis ($Z = 7.8823; P = 0.0001$; Table 5). Additionally, the systematic error-to-noise ratio (*SNR*) was 18.4%, which was only small if compared to the *SNR* of subjects (3338.4%), illustrating how sampling from disparately shaped groups can obfuscate interpretation.

Moreover, adding clade as a grouping factor to the measurement error analysis had an interesting effect. First, the subject variation reduced from $R^2 = 0.9658$ to $R^2 = 0.7082$. (This is the shape variation among subjects, accounting for clade differences.) The amount of variation explained as systematic ME remained the same, $R^2 = 0.0053$, however, the former $R^2 = 0.0289$ for random ME was now partitioned into $R^2 = 0.0065$ and $R^2 = 0.0224$, for the systematic ME:clade interaction and random ME, respectively. Thus, 18.9% of the total ME could be explained by the systematic ME:clade interaction, meaning the SNR statistics for systematic ME and systematic ME:clade were 23.7% and 28.9%, respectively. The effect sizes for systematic ME were slightly changed by adding clades (increased for ANOVA but decreased for MANOVA). However, the effect of adding clades meant that a significant systematic ME:clade effect was observed for both dispersion ($Z = 2.8630; P = 0.0014$) and multivariate analysis ($Z = 3.3087; P = 0.0001$). That the systematic ME effect size increased for ANOVA but decreased for MANOVA, but the effect size for the systematic ME:clade effect was greater in MANOVA, suggests that the group effect was more associated with the changes in covariances among Procrustes coordinates; i.e., differences between replicates could be more associated with the direction of replicate vectors rather than the length of the vectors in a PC plot.

*ICC* statistics were again misleading. Accounting for clades reduced *ICC* dispersion statistics, but only slightly (*ICC* = 0.943–0.949, for all three types.) *ICC* statistics were all 0.999 in the first component for the multivariate analysis. The *ICC* statistics merely confirmed that subjects were so different in shape that even obvious differences from digitizing could be dismissed. This was not a consistent interpretation when viewing SNR plots (Fig. 7C:F).

The *SNR* plots revealed that in three cases, ME was a concern for the subjects sampled compared to other subjects: *Dendrolagus goodfellowi*, *Setonix brachyurus*, and *Aepyprymnus rufescens*. The concerns were not as apparent in the PC plot, or were not strongly apparent compared to other clusters of points for subjects. For example, in the PC plot, point-scatter for *Onychogalea fraenata* and *O. unguifera* compared to most other subjects might elicit some concern, but it was apparent in the PC plot that the scanning devices clustered as pairs, meaning the spread of points was comparatively reduced for these two species in the *SNR* plots. The three species that stood out tended to have inconsistent patterns compared to other species, which might explain why a significant systematic ME:clade interaction was observed. For both *Dendrolagus goodfellowi* and *Aepyprymnus rufescens*, there was a strong operator difference associated with the first *SNR* eigenvector, but additionally, the most divergent (*A. rufescens*) or nearly most divergent (*D. goodfellowi*) estimates of shape came between the two operators while using photogrammetry as the method of data acquisition (even more so than between operators with different devices). By contrast, only one operator had a divergent estimate of shape with photogrammetry for *Setonix brachyurus*, otherwise the estimates of shape were rather clustered

**Table 5** Multivariate analysis of variance tables evaluating random and systematic components of measurement error, for the empirical example

| | $\lambda_{max}$ | $Z$ | $P$ |
|---|---|---|---|
| A: Analysis without clade effect | | | |
| Subjects/random ME | 3015.3400 | 9.0960 | 0.0001 |
| Systematic ME/random ME | 5.5374 | 7.8823 | 0.0001 |
| B: Analysis with clade effect | | | |
| Subjects/random ME | 1939.6671 | 2.6229 | 0.0001 |
| Systematic ME/random ME | 7.7918 | 6.3449 | 0.0001 |
| Systematic ME:groups/random ME | 20.6687 | 3.3087 | 0.0001 |

(Fig. 7F). Interestingly, these three species were all found in a similar portion of the shape space, divergent in shape from most other species. These results suggest that systematic ME can be localized (appear only for certain subjects) because of divergent digitizing prejudices only for certain subjects, and as resoundingly suggested already, sampling from a broader set of subjects can hide such concerns, if conclusions are based on statistics that relativize ME by subject variation.

These results allude to shape estimation concern because the choice of operator-scanner combination that can affect the estimates of shape differences among subjects. Although neither vectors for operator digitizing prejudices nor device digitizing prejudices were significantly correlated with either species or clade vectors, the interaction between operator and device was significantly correlated with both species ($Z = 3.1220; P = 0.0001$) and clade ($Z = 2.3068; P = 0.0011$) (Fig. 8A). Furthermore, a heat map of variances (Fig. 8B) among inter-species (Euclidean) shape distances revealed concern about the estimates of *Dendrolagus goodfellowi* and *Aepyprymnus rufescens* shapes, as there was greater variability in shape distances between these and other species, meaning choice of an operator-device combination could affect estimates of shape, and thus, shape variation. The concern for *Setonix brachyurus* was not as evident in this plot, suggesting that outside of the one aberrant estimate, shape estimates were consistent.

## Discussion

This article provides a conceptual and mathematical investigation of the subject of measurement error as it pertains to geometric morphometric data. We argued that the current state of the field does not arm empiricists with the tools required for determining whether ME should be of concern in their datasets, largely because of their inability to distinguish between systematic and random ME. Through several motivating examples we developed a set of analytical procedures and graphical tools that dissect the random and systematic components of ME from one another, and extract any group-specific systematic ME that may be present. Through simulation and empirical examples we demonstrated that relying on simple summary measures such as the *ICC* or $R^2$ is insufficient for determining whether ME is a problem, and that inter-subject variation can obfuscate the effects of systematic ME in a sample. By contrast, we illustrated that our new procedures are capable of detecting how and where ME affects patterns of shape variation, and thus downstream biological inferences made from such data. Overall our procedures provide a deeper interrogation of ME than is currently accessible, thereby formalizing a new paradigm for how empiricists should investigate the effects of measurement error in multivariate data.

From the extensive simulations performed here, we can conclude that the analytical paradigm we have proposed does not produce spurious results and has appropriate statistical properties. We were able to determine from the simulation experiments that (1) random ME does not produce significant patterns of systematic ME, irrespective of the amount of ME, but (2) the same digitizing prejudice applied to subjects sampled from groups with disparate shapes might not only produce significant systematic ME in an hypothesis test, but also a significant systematic ME by group interaction. This possibility is important. It means that as a practice, a consistent digitizing prejudice might not be negligible for GM data, if applied to all research specimens. It made sense that with the simulation experiments the digitizing prejudice could have varied results, as the groups differed in tail shape and the prejudice of lengthening or shortening a tail by an absolute amount with respect to landmark placement would impact short-tailed and long-tailed species differently. It is perhaps no surprise that a consistent digitizing prejudice could spur varied types of systematic ME. Researchers familiar with generalized Procrustes analysis (GPA) are probably universally aware of the "Pinocchio effect", whereby a displacement of a single landmark (e.g., tip of Pinocchio's nose) in one landmark configuration, in which all alternative landmarks are in the same location in a replicate configuration, will result in different locations of every Procrustes coordinate in the configuration, following GPA (Klingenberg, 2021). If a nose tip was shifted exactly $x$ units in the same direction for two landmark configurations—but the configurations already differed in terms of nose length—the changes in relative nose length would differ between the configurations and distribution of change across all landmarks should not be expected to be the same.

However, for GM studies, measurement error should be focused on the precise estimate of shapes, and thus, shape differences, so a direct link between process and pattern is

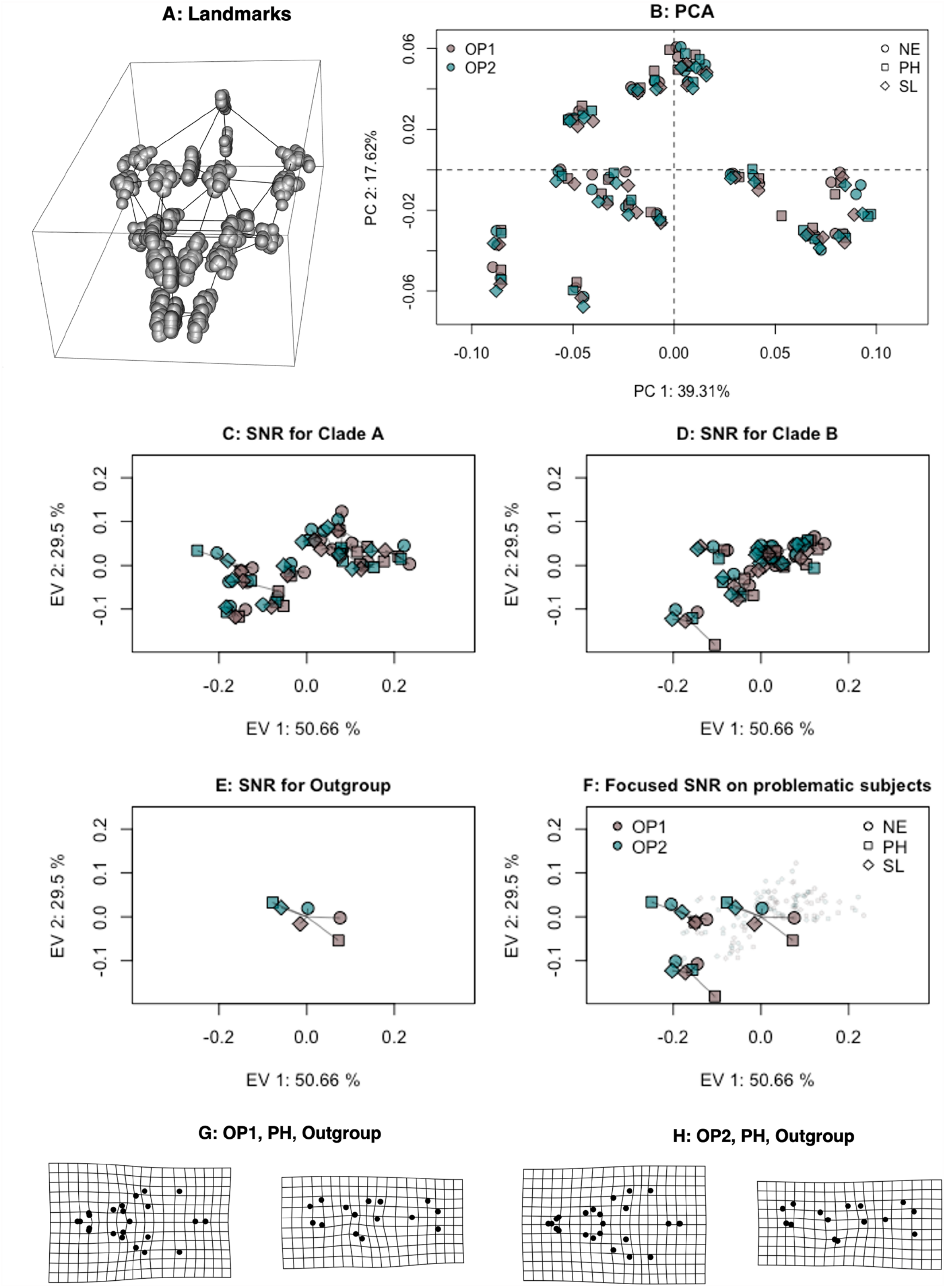

**Fig. 7** **A** Set of 138 Procrustes-aligned specimens, representing the skulls of 23 individuals whose landmarks were digitized by two different observers on each of three separate 3D scans. **B** Principal components plot of 138 shapes, colored by operator and with symbols representing different scanning devices. **C**–**F** *SNR* plots of systematic ME versus random ME, shown uniquely for different clades and focused on problematic specimens. The *SNR* plots are clade-centered, so the origin represents the clade mean. **G**, **H** Thin-plate spline (TPS) transformation grids (scaled 2× to facilitate interpretation) for one specimen, and one device (photogrammetry), but differing by operators in the two plots. Both dorsal and ventral grids are shown. The reference configuration is the clade-adjusted mean

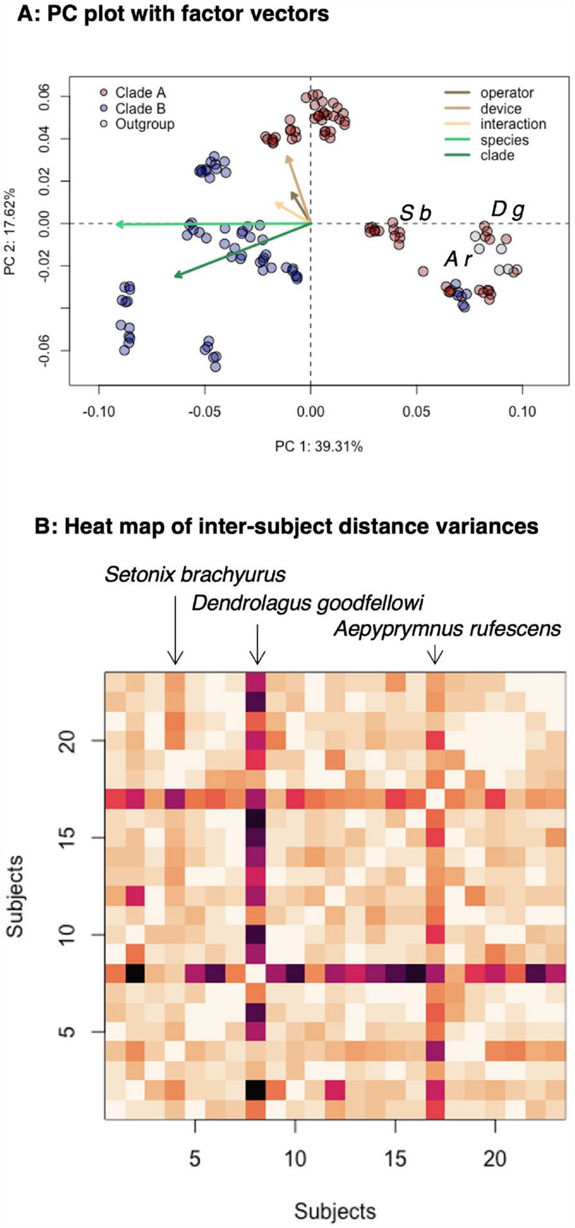

**Fig. 8** **A** The same PC plot as in the previous figure, however, color coded by clade, and with vectors illustrating principal eigenvectors of SSCP matrices for different effects. The vectors for operator, device, and interaction are appropriately scaled in a relative sense (longer vectors mean large effect). These vectors have also been scaled 10× with respect to species and clade vectors, to faciliate interpretation. Species with substantial measurement error are labeled with abbreviations: *Dendrolagus goodfellowi*, *Setonix brachyurus*, and *Aepyprymnus rufescens*. **B** A heat map showing the relative amount of variability (variance) for inter-species shapae differences, based on the six different replicates. Darker colors mean more variable estimates

not required (so long as it can be ascertained how a process produces a pattern). Therefore, that a consistent digitizing prejudice can produce varied amounts of systematic ME is not a worry, as much as one should be worried that subject-specific systematic ME can lead to spurious estimates of shape. Furthermore, relativizing ME, whether systematic or random, by subject variation can minimize concern for ME, and (3) relying on statistics that find a ratio of subject variation and within-subject replicate variation (like *ICC* statistics) should be avoided. Both our simulated and empirical results emphasized this. *ICC* statistics measure repeatability, and strong repeatability might seem to be associated with lack of ME, but such interpretations depend on the scale of subject variation. A researcher might be comforted to recognize that despite digitizing prejudices and potential (random) instrument ME, their ability to measure species differences in shape is substantial, as species are much more different in shape than replicated measurements on the same species. This line of thinking is probably okay, provided the data set does not comprise any similarly shaped species. Alternatively, if some species have recent evolutionary divergence and are more similar in shape, and these species are compared to other disparately shaped species with longer periods of divergence, it should be imperative to have precise estimates of shape differences between the similar species, especially if within-clade rates of evolutionary divergence could be measured. Reducing concern for ME in such cases based on a more global perspective of shape variation would be unfortunate.

Foremost, ME studies should be considered experimental. They might not sample from all specimens that would be used in broader study but understanding the impact of using different researchers, different cameras, different scanning devices, etc., would likely be an early-step, exploratory procedure (preliminary experiment) rather than a hopeful confirmation after all data have been collected, haphazardly. Therefore, with a careful, balanced design that employs all possible replicate measurements on the same set (or subset) of subjects, a concomitant analytical paradigm with the statistical power to detect subtle but meaningful sources of shape variation should be desired. The simulated and empirical results in this paper confirm that (4) large effect sizes can be measured for systematic ME, even if the amount of variation is small compared to subject variation. Furthermore, (5) *SNR* plots can help elucidate the localized problems that trigger large systematic ME effect sizes. The *SNR* plots are especially helpful, as they find eigenvectors that maximize systematic ME relative to random ME. Both simulated and empirical results illustrated how these plots can reveal patterns that might be missed with PCA, alone. If one wishes to identify potential sources of systematic ME rather than reassure themselves that it is not an issue, then the methods we presented appear to facilitate this goal.

One inadvertent suggestion we might have made is that a GM-ME experiment needs to be balanced. This implication is more so related to the calculation of *ICC* statistics that use the number of replicates in their calculation. Although imbalance of replicate sampling does not necessarily preclude *ICC* calculation, its value as an effect size would certainly be compromised without balanced replication. Alternatively, the RRPP strategy we have used does not require replicate balance. By restricting RRPP permutations within subjects, it is possible to generate distributions of statistics based on uneven replicate sampling within subject. (Even subjects with only one replicate could be technically included in the analysis, although any inference about systematic ME with regard to such subjects would not be possible.) For GM-ME studies, we do not recommend designs that are greatly imbalanced, as it would be difficult to rely on the eigenvectors produced for replicate effects if some replicates are poorly represented. However, provided all replicates are suitably sampled from most subjects, it would still be possible to make subject-specific evaluations in *SNR* plots, in spite of missing replicates. Further research would be required to develop a better understanding of how sampling problems could cause misinterpretations of systematic ME. With the methods we have developed here, such research should be possible to explore (in terms of statistical properties).

One outcome that we did not anticipate is that GPA can mitigate the systematic ME caused by a digitizing prejudice. This phenomenon was evidenced by the comparatively, substantially lower statistical power to detect general or group-specific systematic ME in simulation experiments that applied a digitizing prejudice to one group. By having simulation experiments where the general locations of landmarks were somewhat fixed because of invariance to translation and rotation (small random displacements, notwithstanding), we could perform ME analyses on raw landmarks. Furthermore, because type I error rates were appropriate, the larger statistical power associated with analysis on landmarks cannot be explained by random size, orientation, or location results of configurations. Rather, in the case of using landmarks, systematic ME was akin to a Pinocchio effect, and more evident by the change in location of just two landmarks between replicates. GPA mitigated this effect. This is an interesting result, as recent concerns whether GPA can induce spurious results in terms of variable covariances (e.g., Cardini, 2019) could lead one to be concerned whether GPA could induce systematic ME. Our results found no evidence of this, but just the opposite. A consistent digitizing prejudice that misplaces one or few landmarks might not be as profound for Procrustes coordinates as for the raw landmarks. Furthermore, GPA cannot induce spatial covariances of Procrustes coordinates within configurations that are different than the original configurations, unless a sliding algorithm is used for semilandmarks. GPA will necessarily alter the covariances among landmarks for a set of configurations. It remains possible that a digitizing prejudice applied to just one or few configurations could grossly alter the covariance structure of a set of Procrustes coordinates for many specimens, but for such a case in reality, an aberrant specimen in terms of shape or extreme systematic changes to landmarks only in a few specimens would likely be needed to provoke such results. The methods we have introduced would probably not be needed to identify the inherent problems with such data.

One practical issue we have not considered is what might be a plan of action, given results from an analysis of data from an GM-ME experiment. For example, with the empirical data collected by Fruciano et al. (2017) it could be decided that obtaining the means of the six replicates for each species is a safe endeavor for further analysis (see Arnqvist & Mårtensson, 1998). Alternatively, a research team might wish to revisit the operator-device combinations for the few exceptional species, especially to learn why photogrammetry produced disparate results. The analytical results and plots we produced indicate potential sources of problems but do not necessarily have to alarm researchers that these problems are substantial. By contrast, relying on *ICC* statistics could have the opposite problem of assuaging researchers' concerns when concerns are warranted. The especially useful tool of using points in *SNR* plots to generate thin-plate spline transformation grids can allow one to decide if shape changes associated with systematic ME are minor or major. We provided one example of such exploration of shape differences between replicates (Fig. 7G, H). Whether this warrants re-digitization is a decision the researcher can make. Alternatively, one might consider in the empirical example which operator and scanning device combinations tended to yield the most consistent results. (For example, the combination of operator 1 and Solutionix laser scanner tended to produce shape estimates nearest to the means of replicate measurements for most species, in the *SNR* plots.) The analytical paradigm we propose here makes such determinations possible.

Nevertheless, one motion we wish to make in this paper is that researchers should not assuage concern for ME by focusing strongly on subject variation. The large statistical power from our simulation experiments (Supplementary Material) is possible by having a statistical method that preserves subject variation across random permutations, allowing a precise, focused test of replicate variation, capable of discerning trends independent of and despite subject variation. This is important. It should be possible to detect these trends, even if a PC plot fails to reveal them (because the first few principal components are strongly associated with inter-subject shape variation). Fruciano et al. (2017) also observed significant variation in shape estimates based on scanning device but suggested using fewer principal components of the data alleviated these concerns. Naturally, using a subset of principal components that largely reveal trends in subject shape variation could eliminate concern for ME. But this is a biased statistical approach. Our results suggest, by contrast, that using a better method of inquiry and evaluation pinpoints the concerns that could be addressed rather than swept under the rug with data reduction. As a research tool, the results of this example indicate a path for addressing measurement error. The researchers can (1) identify which subjects are of concern, (2) visualize the shape difference associated with the first few *SNR* eigenvectors, (3) ascertain whether it is an operator or device digitizing prejudice that is a concern, or (4) whether it is an interaction of these prejudices that are a concern, and (5) identify whether systematic ME is localized to a portion of the sample shapes.

Naturally, there will be an inherent desire for researchers to reconcile whether ME (especially systematic ME, but random ME, as well) impedes their ability to test hypotheses that address biological questions. There might also be a natural inclination to wish to assuage fears about ME, if the amount of overall ME variation is small compared to subject variation. We have indicated that sampling from a diverse population of shapes can mitigate concerns for ME using the methods that have been traditionally employed to measure ME. We do not wish to suggest that sampling from a diverse population of shapes is bad idea; quite the contrary, we recommend it! However, if one wishes to evaluate whether ME is an attribute that can be disregarded, it is imperative that honest assessments of components of ME are made independent of subject variation. The analytical paradigm we present makes it possible to produce sampling distributions of statistics, found independent of the subject variation sampled, meaning one need not be concerned with how subject variation impacts interpretation of ME.

An interesting juxtaposition arises with these new methods. We could consider, for example, a research team that performs a GM-ME experiment with a small portion of the taxa they wish to examine in a full study, to investigate whether non-unique digitizing strategies could impact their results. Upon obtaining results, they decide to add a few more subjects, especially adding representation of more divergently shaped taxa, and re-evaluate the data. With traditional statistics like *ICC*, results seem to improve. With the ME test we introduce here, perhaps the systematic ME × groups effect size increases. How would one deal with this possible outcome? With the methods we introduce, it becomes possible with broad sampling to determine if digitizing prejudices can manifest as localized systematic ME. This has not been an easily achievable inference to attain with traditional methods. The biometer retains the capacity to decide if ME is negligible but now with methods that do not conflate subject and digitizer variation. More importantly, the biometer is not dissuaded from investigating possible sources of digitizing prejudices, even if subtle, unlike the false reassurance that might be found from simple descriptive statistics.

To the best of our knowledge, there has not been statistical development as rigorous as we have covered in this paper, for ME studies with GM data. Although we do not expect that the methods we present here represent the possible panoply of methods that could be developed on this subject, we believe the development of appropriate statistical methods (that test systematic ME, independent of subject variation) and graphical tools advance the scientific endeavor of measurement error analysis in GM studies considerably more than it has advanced in the last few decades. We suspect that a future research direction could be the development of better experimental designs for GM-ME experiments, another area that has not received strong consideration. Coupled with an appropriate and expandable method of analysis (in terms of factorial models), this development should be easily achievable.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11692-024-09627-6.

**Acknowledgements** We thank P.D. Polly and an anonymous reviewer for helpful comments on an earlier version of this paper. We also thank the MorphMet listserv—and A. Cardini in particular—for a discussion of measurement error, which made clear to us that current recommendations regarding measurement error in morphometrics were inadequate, and required a rethink. The present paper takes the nascent ideas we expressed in that thread, and converts them into fully developed analytical methods. This work was sponsored in part by National Science Foundation Grants DBI-1902694 and DEB-2146220 (to MLC), and DBI-1902511 and DEB-2140720 (to DCA). All analyses in this paper were performed in `R` (R Core Team, 2023), using the packages, `geomorph` (Adams et al., 2023; Baken et al., 2021) and `RRPP` (Collyer & Adams, 2018, 2023). The functions `measurement.error`, `plot.measurement.error`, `focusMEonSubjects`, `interSubVar` and `plot.interSubVar` in `RRPP`, and `gm.measurement.error` in `geomorph`, contain all new analytical approaches described in this paper.

**Author Contributions** M.L.C and D.C.A wrote the main manuscript. M.L.C wrote computer scripts, and prepared figures and tables. Both authors reviewed the manuscript.

**Data Availability** All data from simulattion experiments can be generated with scripts in the Supplementary material. Data for the empirical example can be found at the Dryad Digital Repository: https://doi.org/10.5061/dryad.t9888.

## Declarations

**Conflict of interest** The authors declare that they have no known conflicts of interest.

## References

Adams, D. C. (2014). A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. *Evolution, 68*, 2675–2688. https://doi.org/10.1111/evo.12463

Adams, D. C., & Collyer, M. L. (2018). Phylogenetic ANOVA: Group-clade aggregation, biological challenges, and a refined permutation procedure. *Evolution, 72*(6), 1204–1215.

Adams, D. C., & Collyer, M. L. (2019). Comparing the strength of modular signal, and evaluating alternative modular hypotheses, using covariance ratio effect sizes with morphometric data. *Evolution, 73*, 2352–2367. https://doi.org/10.1111/evo.13867

Adams, D. C., & Collyer, M. L. (2022). Consilience of methods for phylogenetic analysis of variance. *Evolution, 76*(7), 1406–1419.

Adams, D. C., Collyer, M. L., Kaliontzopoulou, A., & Baken, E. K. (2023). Geometric morphometric analyses of 2D and 3D landmark data, version 4.0.6. R Foundation for Statistical Computing. https://cran.r-project.org/package=geomorph

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix, 24*, 7–14.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32–46.

Anderson, M. J., & Walsh, D. C. (2013). PERMANOVA, ANOSIM, and the mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs, 83*(4), 557–574.

Arnqvist, G., & Mårtensson, 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.

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

Baken, E. K., Collyer, M. L., Kaliontzopoulou, A., & Adams, D. C. (2021). Geomorph 4.0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods in Ecology and Evolution, 12*, 2355–2363.

Barbeito-Andrés, J., Anzelmo, M., Ventrice, F., & Sardi, M. L. (2012). Measurement error of 3D cranial landmarks of an ontogenetic sample using computed tomography. *Journal of Oral Biology and Craniofacial Research, 2*, 77–82. https://doi.org/10.1016/j.jobcr.2012.05.005

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*, 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Bookstein, F. L. (1991). *Morphometric tools for landmark data: Geometry and biology*. Cambridge University Press.

Bookstein, F. L. (2015). Integration, disintegration, and self-similarity: Characterizing the scales of shape variation in landmark data. *Evolutionary Biology, 42*, 395–426. https://doi.org/10.1007/s11692-015-9317-8

Bookstein, F. L., Gunz, P., Mitterœcker, P., Prossinger, H., Schæfer, K., & Seidler, H. (2003). Cranial integration in homo: Singular warps analysis of the midsagittal plane in ontogeny and evolution. *Journal of Human Evolution, 44*(2), 167–187. https://doi.org/10.1016/s0047-2484(02)00201-4

Bookstein, F. L., & Mitterœcker, P. (2014). Comparing covariance matrices by relative eigenanalysis, with applications to organismal biology. *Evolutionary Biology, 41*, 336–350.

Cardini, A. (2019). Integration and modularity in procrustes shape data: Is there a risk of spurious results? *Evolutionary Biology, 46*(1), 90–105.

Collyer, M. L., & Adams, D. C. (2013). Phenotypic trajectory analysis: Comparison of shape change patterns in evolution and ecology. *Hystrix, the Italian Journal of Mammalogy, 24*, 75–83. https://doi.org/10.4404/hystrix-24.1-6298

Collyer, M. L., & Adams, D. C. (2018). RRPP: An R package for fitting linear models to high-dimensional data using residual randomization. *Methods in Ecology and Evolution, 9*, 1772–1779.

Collyer, M. L., & Adams, D. C. (2023). RRPP: Linear model evaluation with randomized residuals in a permutation procedure, version 1.3.2. R Foundation for Statistical Computing. https://cran.r-project.org/package=RRPP

Collyer, M. L., Baken, E. K., & Adams, D. C. (2022). A standardized effect size for evaluating and comparing the strength of phylogenetic signal. *Methods in Ecology and Evolution, 13*(2), 367–382.

Collyer, M. L., Sekora, D. J., & Adams, D. C. (2015). A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity, 115*(4), 357–365.

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics, 15*(2), 171–185.

Conaway, M. A., & Adams, D. C. (2022). An effect size for comparing the strength of morphological integration across studies. *Evolution, 76*, 2244–2259. https://doi.org/10.1111/evo.14595

Cramon-Taubadel, N., von Frazier, B. C., & Lahr, M. M. (2007). The problem of assessing landmark error in geometric morphometrics: Theory, methods, and modifications. *American Journal of Physical Anthropology, 134*, 24–35. https://doi.org/10.1002/ajpa.20616

Daboul, A., Ivanovska, T., Bülow, R., Biffar, R., & Cardini, A. (2018). Procrustes-based geometric morphometrics on MRI images: An example of inter-operator bias in 3D landmarks and its impact on big datasets. *PLoS ONE, 13*, e0197675. https://doi.org/10.1371/journal.pone.0197675

Fisher, R. A. (1950). *Statistical methods for research workers* (11th ed.). Oliver; Boyd.

Fleiss, J. L., & Shrout, P. E. (1977). The effects of measurement errors on some multivariate procedures. *American Journal of Public Health, 67*, 1188–1191.

Fox, N. S., Veneracion, J. J., & Blois, J. L. (2020). Are geometric morphometric analyses replicable? Evaluating landmark measurement error and its impact on extant and fossil Microtus classification. *Ecology and Evolution, 10*, 3260–3275. https://doi.org/10.1002/ece3.6063

Fruciano, C. (2016). Measurement error in geometric morphometrics. *Development Genes and Evolution, 226*, 139–158. https://doi.org/10.1007/s00427-016-0537-4

Fruciano, C., Celik, M. A., Butler, K., Dooley, T., Weisbecker, V., & Phillips, M. J. (2017). Sharing is caring? Measurement error and the issues arising from combining 3D morphometric datasets. *Ecology and Evolution, 7*, 7034–7046. https://doi.org/10.1002/ece3.3256

Galimberti, F., Sanvito, S., Vinesi, M. C., & Cardini, A. (2019). Nose-metrics of wild southern elephant seal *Mirounga leonina* males using image analysis and geometric morphometrics. *Journal of Zoological Systematics and Evolutionary Research, 57*, 710–720. https://doi.org/10.1111/jzs.12276

Giacomini, G., Scaravelli, D., Herrel, A., Veneziano, A., Russo, D., Brown, R. P., & Meloro, C. (2019). 3D photogrammetry of bat skulls: Perspectives for macro-evolutionary analyses. *Evolutionary Biology, 46*, 249–259. https://doi.org/10.1007/s11692-019-09478-6

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological), 53*(2), 285–321.

Gunz, P., Mitterœcker, P., & Bookstein, F. L. (2005). Semilandmarks in three dimensions. In *Developments in primatology: Progress and prospects* (pp. 73–98). Kluwer Academic Publishers-Plenum Publishers. https://doi.org/10.1007/0-387-27614-9_3

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance*. Dryden Press.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society Series A (Statistics in Society), 159*, 445–492. https://doi.org/10.2307/2983326

Houle, D., Pélabon, C., Wagner, G. P., & Hansen, T. F. (2011). Measurement and meaning in biology. *The Quarterly Review of Biology, 86*, 3–34. https://doi.org/10.1086/658408

Klingenberg, C. P. (2010). MorphoJ: An integrated software package for geometric morphometrics. *Molecular Ecology Resources, 11*, 353–357. https://doi.org/10.1111/j.1755-0998.2010.02924.x

Klingenberg, C. P. (2021). How exactly did the nose get that long? A critical rethinking of the Pinocchio effect and how shape changes relate to landmarks. *Evolutionary Biology, 48*(1), 115–127.

Klingenberg, C. P., Barluenga, M., & Meyer, A. (2002). Shape analysis of symmetric structures: Quantifying variation among individuals and asymmetry. *Evolution, 56*, 1909–1920. https://doi.org/10.1111/j.0014-3820.2002.tb00117.x

Klingenberg, C. P., & Gidaszewski, N. A. (2010). Testing and quantifying phylogenetic signals and homoplasy in morphometric data. *Systematic Biology, 59*, 245–261.

Klingenberg, C. P., & McIntyre, G. S. (1998). Geometric morphometrics of developmental instability: Analyzing patterns of fluctuating asymmetry with procrustes methods. *Evolution, 52*, 1363–1375. https://doi.org/10.1111/j.1558-5646.1998.tb02018.x

Konishi, S., Khatri, C. G., & Rao, C. R. (1991). Inferences on multivariate measures of interclass and intraclass correlations in familial data. *Journal of the Royal Statistical Society Series B (Methodological), 53*, 649–659.

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, volume i: Additive and polynomial representations*. Academic Press.

Kreutz, C., Raue, A., Kaschek, D., & Timmer, J. (2013). Profile likelihood in systems biology. *FEBS Journal, 280*, 2564–2571. https://doi.org/10.1111/febs.12276

Kyburg, H. (1984). *Theory and measurement*. Cambridge University Press.

Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation—A discussion and demonstration of basic features. *PLoS ONE, 14*, e0219854. https://doi.org/10.1371/journal.pone.0219854

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement, volume III: Representation, axiomatization, and invariance*. Academic Press.

Marcy, A. E., Fruciano, C., Phillips, M. J., 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. https://doi.org/10.7717/peerj.5032

Menéndez, L. P. (2016). Comparing methods to assess intraobserver measurement error of 3D craniofacial landmarks using geometric morphometrics through a digitizer arm. *Journal of Forensic Sciences, 62*, 741–746. https://doi.org/10.1111/1556-4029.13301

Mitterœcker, P., & Bookstein, F. L. (2009). The ontogenetic trajectory of the phenotypic covariance matrix, with examples from craniofacial shape in rats and humans. *Evolution, 63*, 727–737.

Mitterœcker, P., Gunz, P., Bernhard, M., Schæfer, K., & Bookstein, F. L. (2004). Comparison of cranial ontogenetic trajectories among great apes and humans. *Journal of Human Evolution, 46*, 679–698. https://doi.org/10.1016/j.jhevol.2004.03.006

Mitterœcker, P., & Schæfer, K. (2022). Thirty years of geometric morphometrics: Achievements, challenges, and the ongoing quest for biological meaningfulness. *American Journal of Biological Anthropology, 178*, 181–210. https://doi.org/10.1002/ajpa.24531

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rabinovich, S. G. (2005). *Measurement errors and uncertainties: Theory and practice* (3rd ed.). Springer Nature. https://www.ebook.de/de/product/3897875/semyon_g_rabinovich_measurement_errors_and_uncertainties_theory_and_practice.html

Robinson, C., & Terhune, C. E. (2017). Error in geometric morphometric data collection: Combining data from multiple sources. *American Journal of Physical Anthropology, 164*, 62–75. https://doi.org/10.1002/ajpa.23257

Rohlf, F. J., & Corti, M. (2000). Use of two-block partial least-squares to study covariation in shape. *Systematic Biology, 49*, 740–753. https://doi.org/10.1080/106351500750049806

Rohlf, F. J., & Slice, D. E. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology, 39*, 40–59.

Shearer, B. M., Cooke, S. B., Halenar, L. B., Reber, S. L., Plummer, J. E., Delson, E., & Tallman, M. (2017). Evaluating causes of error in landmark-based data collection using scanners. *PLoS ONE, 12*, e0187452. https://doi.org/10.1371/journal.pone.0187452

Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (1989). *Foundations of measurement, volume II: Geometrical, threshold, and probabilistic representations*. Academic Press.

Vrdoljak, J., Sanchez, K. I., Arreola-Ramos, R., Huesa, E. G. D., Villagra, A., Avila, L. J., & Morando, M. (2020). 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). *Biological Journal of the Linnean Society, 130*, 800–812. https://doi.org/10.1093/biolinnean/blaa079

Yezerinac, S. M., Lougheed, S. C., & Handford, P. (1992). Measurement error and morphometric studies: Statistical power and observer experience. *Systematic Biology, 41*, 471–482. https://doi.org/10.2307/2992588