Evolutionary Biology
https://doi.org/10.1007/s11692-020-09527-5

TOOLS AND TECHNIQUES

Repeatability and Validity of Phenotypic Trait Measurements in Birds

Kalya Subasinghe[1,2,3] · Matthew R. E. Symonds[4] · Marta Vidal-García[1,5] · Timothée Bonnet[1] · Suzanne M. Prober[6] · Kristen J. Williams[2] · Janet L. Gardner[1]

Received: 15 June 2020 / Accepted: 11 December 2020

Abstract

Phenotypic trait data play a central role in ecology and evolutionary research. The quality of trait data, and the findings of subsequent analyses, depend on the quality of measurement. However, most studies overlook measurement accuracy in their study designs. We investigated the repeatability of five frequently used linear measurements of avian traits: wing length, tarsus length, bill length, bill depth and bill width and the validity of proxies for three traits: bill surface area, structural body size and tarsus size, using species from the infra-order Meliphagides (honeyeaters, fairy wrens and their allies). Repeatability varied between traits and across species for a given trait: traits larger than 13 mm showed high repeatability compared with smaller traits. By incorporating microCT technology, we showed that the formula for the surface area of a cone, a widely used proxy of bill surface area, accurately describes bill surface area within species. Surface measurement of tarsus and wing lengths were valid proxies for underlying osteology. We recommend preliminary estimation of repeatability should be undertaken for individual traits prior to data collection, in order to design suitable protocols that improve data quality, while optimizing costs involved, particularly for traits < 13 mm.

Keywords Measurement repeatability · Measurement validity · Bill surface area · Wing morphology · Tarsus length · Meliphagides

Introduction

Variation in the size of phenotypic traits is often estimated in ecological and evolutionary studies. For example, morphometric measures of traits have been widely used to establish phylogenetic relationships among taxa (Thiele 1993; Wiens 2004) and are increasingly applied to studies

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11692-020-09527-5.

✉ Kalya Subasinghe
kalya.subasinghe@anu.edu.au

Matthew R. E. Symonds
matthew.symonds@deakin.edu.au

Marta Vidal-García
marta.vidalga@gmail.com

Timothée Bonnet
timotheebonnetc@gmail.com

Suzanne M. Prober
Suzanne.Prober@csiro.au

Kristen J. Williams
Kristen.Williams@csiro.au

Janet L. Gardner
janet.gardner@anu.edu.au

[1] Research School of Biology, The Australian National University, Canberra, ACT 0200, Australia

[2] CSIRO Land and Water, GPO Box 1700, Canberra, ACT 2601, Australia

[3] Department of Zoology and Environmental Management, University of Kelaniya, Kelaniya 11600, Sri Lanka

[4] Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Burwood, VIC 3125, Australia

[5] Department of Cell Biology and Anatomy, University of Calgary, Calgary, AB T2N 4N1, Canada

[6] CSIRO Land and Water, Private Bag 5, Wembley, WA 6913, Australia

Published online: 03 January 2021

of adaptation to climate (McLean et al. 2014; Gardner et al. 2019). These studies are in turn used to inform nature conservation decisions, including prioritisation of investment in threatened species (phylogenetic conservation prioritisation, Mooers et al. 2008; Billionnet 2018), selection of provenancing strategies for climate-resilient ecosystem restoration (Prober et al. 2015), and other climate adaptation management applications (Mawdsley et al. 2009). Accurate estimation of phenotypic traits underpins the effectiveness of these investigations and applications. The accuracy of estimations partly depends on measurement quality, described by measurement accuracy and precision. Measurement accuracy is how close a measured value is to its true value, whereas precision is the closeness of agreement between repeated measures, affected by random errors (Hick and Emmerson 2014). Despite its importance, few studies explicitly consider measurement quality in their study designs (Gosler 1987; Martin and Pitocchelli 1991; Benítez-Díaz 1993; Harris and Smith 2009; Perktaş and Gosler 2010; Anderson et al. 2019).

Repeatability of Measurements

Ideally, repeated measures of a trait from a single individual should be identical, if the measurer used the same method and instrument for the measurement (Harper 1994). However, the value will vary between measures due to random errors, errors associated with both the measurer (e.g. variability in locating landmarks that define the trait) and instrument imprecisions. If the variation caused by measurement error represents a significant portion of the variation between individuals, then the measurements will not represent the true biological variation. Such measurements are not reliable and are of limited use in studies. The portion of variation that is attributed to true biological variation can be quantified using repeatability, a widely used index of measurement quality, also known as the intra-class correlation coefficient (ICC) (Nakagawa and Schielzeth 2010).

Some researchers have incorporated precautionary steps in their study designs to reduce potential measurement errors, which will ultimately improve measurement quality. These include using the same measuring instrument (e.g. digital callipers, rulers) (Benítez-Díaz 1993) and employing the same observer for data collection (Barrett et al. 1989). Despite these attempts, the errors associated with observer measurement inconsistencies such as variability in locating landmarks that define variables, remain (Harris and Smith 2009; Goodenough et al. 2010; Perktaş and Gosler 2010). This may be particularly problematic in cases where trait size, for example bill surface area, is estimated from multiple independent traits whereby the errors associated with each trait can accumulate, affecting the final estimate.

Some trait measurements are inherently less precise than others. A previous study of Short-tailed Shearwaters found higher measurement error for smaller characters including tarsus width (36%) and unguis width (32%) (Totterman 2016). Measurement error is also higher when measuring characters without clearly defined landmarks (e.g. 15% for total culmen and 20% for bill base width in Short-tailed Shearwaters) (Goodenough et al. 2010; Totterman 2016). In small mammals, right hind foot measurement shows a relatively high measurement error (Blackwell et al. 2006; Stephens et al. 2015). However, it is not clear whether the precision of these trait measurements is lower, in general, across species or if there is a size threshold below which the error associated with measurement increases sharply.

Measurement Validity

Validity is the degree to which a measurement represents what it is supposed to measure (Hick and Emmerson 2014). In some cases, phenotypic traits are difficult or impossible to measure directly from live individuals in the field. In these instances, it is a common practice to use an easily measurable character as an index or proxy for a particular trait. These proxies are valid only if they can satisfy the underlying assumptions.

Studies investigating the thermal properties of avian bills, their conformity with Allen's rule (the tendency for appendages to be larger in warmer climates) and adaptive potential to climate change increasingly use linear measurements (i.e. bill length, depth and width) to estimate bill surface area from the formula for the lateral surface area of a nearly circular elliptical cone (Greenberg et al. 2012; Luther and Greenberg 2014; Campbell-Tennant et al. 2015). However, bird bills are adapted for different foraging strategies and are highly diverse morphologically (Sulloway and Kleindorfer 2013; Temeles et al. 2009) so the formula for surface area based on a circular elliptical cone may not be reliable across species. To our knowledge, the applicability of the circular elliptical cone formula as a proxy for bill surface area has never been tested on any group of birds, though it has been used extensively.

Tarsus length, another common measurement, has been widely used to assess geographic variation in bird appendages and conformity with Allen's rule (Nudds and Oswald 2007; Symonds and Tattersall 2010). It has also been recommended as a proxy for structural body size by some authors (Senar and Pascual 1997; Freeman and Jackson 1990). Tarsus length is a measurement taken from the integument cover of tarsometatarsus, usually from the intertarsal joint, to the lower edge of the last undivided scale at the toe divergence (Salewski et al. 2014). There may be differences in integumentary cover of the tarsometatarsus (i.e. the layer of scales) in regard to positioning, degree of overlap, fusion or sizes of scales, among species (Stettenheim 2015). Hence, tarsus length measurements may not be a valid proxy for tarsus size in all species.

In ornithological research, wing length is used as a proxy for structural body size i.e. size of the skeletal frame which supports the soft tissues of an organism (e.g. Aldrich and James 1991; Gardner et al. 2014a). Wing length is commonly measured from the carpal joint to the tip of the longest primary feather, and thus reflects variation in the length of underlying wing bones (carpometacarpus and phalanges) as well as the length of the primary feathers. Flight feathers are continuously abraded and worn between successive moults (which occurs annually in most passerines: Kiat and Sapir 2017) causing changes in the feather length (Leverton 1989; Rising and Somers 1989; Flinks and Salewski 2012). In this context, it is important to assess if manual linear measures of wing length accurately represent structural size based on underlying osteology, and the proportion of measurement variation that comes from feather wear and abrasion.

We investigated the repeatability of commonly used linear body measurements of birds and the validity of linear measurements as proxies for trait size using museum specimens from the infraorder Meliphagides (honeyeaters, fairy wrens, thornbills and their allies). This group, which comprises the largest radiation of Australian passerines, is widely distributed across Australia inhabiting different climatic regions, and hence has been used to study trends in trait sizes across environmental and temporal gradients (Gardner et al. 2014a, b, 2016, 2019; Friedman et al. 2017). Here we (1) compare the repeatability of five common linear body measurements i.e. bill length, bill width, bill depth, wing length, tarsus length and (2) assess how repeatability changes with mean trait size across species. We assess the validity of linear measurements as estimators of trait size; (3) we measure bill surface area digitally and compare with surface area estimated using a mathematical equation based on manual linear measurements as well as the manual measurement of bill length; and (4) assess how bill curvature affects these associations. Finally, (5) we explore how well tarsus length and wing length measurements represent the size of underling bones.

Methodology

Study Species

Using museum specimens of Meliphagides radiation (Aves: Passeriformes), we examined the bills and tarsi of 31 species and the wings of 78 species for repeatability assessment. Additionally, ten species were chosen to represent a diversity of bill shapes and body sizes to assess measurement validity (Supplementary material Table S1, Fig. 1). Specimens were housed in the major museum collections in Australia. Metadata relevant to each specimen (the month of capture, year of capture and sex) were obtained from pre-existing database, to incorporate in models. We excluded damaged specimens and those preserved after 2015 to avoid specimen shrinkage issues (Totterman 2016).

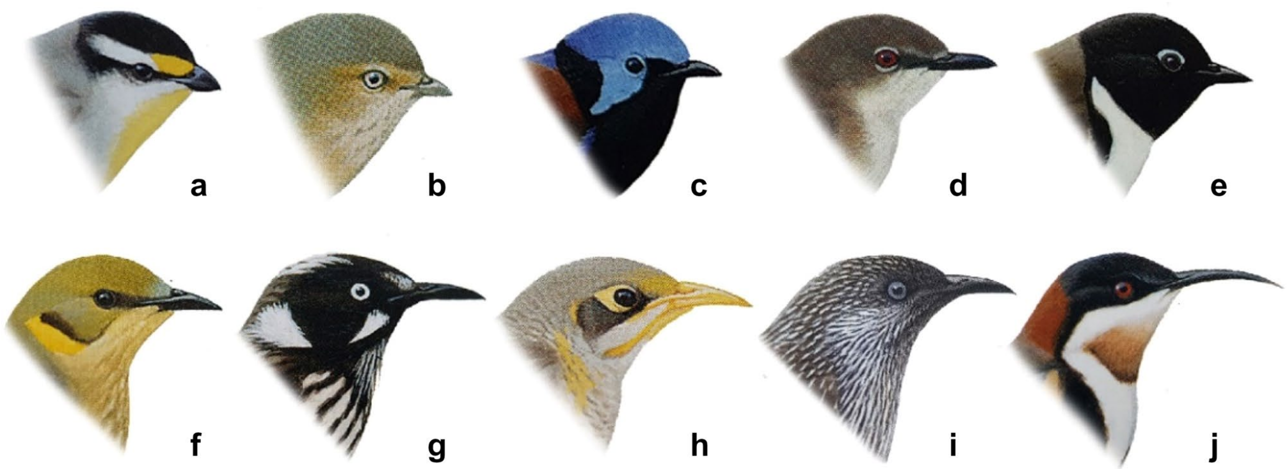

Fig. 1 Bill shape of **a** Striated Pardalote (*Pardalotus striatus*); **b** Weebill (*Smicrornis brevirostris*); **c** Variegated Fairy-wren (*Malurus lamberti*); **d** Large-billed Gerygone (*Gerygone magnirostris*); **e** Black-headed Honeyeater (*Melithreptus affinis*); **f** Grey-fronted Honeyeater (*Ptilotula plumula*); New Holland Honeyeater (*Phylidonyris novaehollandiae*); **h** Yellow-throated Miner (*Manorina flavigula*); **i** Little Wattlebird (*Anthochaera chrysoptera*); **j** Eastern Spinebill (*Acanthorhyncus tenuirostris*). ("Illustrations by Peter Marsack, reproduced from Menkhorst et al. (2017), The Australian Bird Guide, with permission from CSIRO Publishing")

Data Collection

Measurements for Repeatability Assessment

We took two measurements each for bill length, bill depth, bill width and tarsus length of 2062 specimens and wing length of 6709 specimens (Supplementary material Table S1). Bill length was measured from the feathering at the base of the upper mandible to the bill tip; bill width from the posterior edge of the nares on one side of the bill to the same on the other side; and bill depth from the upper mandible to the lower mandible at the posterior edge of the nares at right angles to the tomia (Baldwin et al. 1931; Gardner et al. 2016). Tarsus length was measured from the back of the intertarsal joint to the lower edge of the last undivided scale before the toes diverge (Baldwin et al. 1931; Salewski et al. 2014). Measurements for bill length, depth and width and tarsus length were taken with a pair of digital callipers (Mitutoyo) to an accuracy of 0.01 mm. For wing length, we measured the total length from the carpal joint to the tip of the longest primary of the flattened right wing chord using a butt-ended ruler to the nearest 0.5 mm.

Bill and tarsus measurements for each specimen were collected using the following protocol: one measurement of bill length followed by bill depth, bill width and tarsus length was taken, then a second set of measurements were taken in the same order. The same pair of digital callipers was used for all measurements, and the jaws of the calliper were closed between successive measurements. The repeat measurements of wing length were obtained in two successive handlings widely separated in time, the first in 2012 and the second in 2018. Three measurers were involved in data collection; bill and tarsus measurements of all species were taken by the same investigator, wing length measurements were collected by two investigators, but each species was measured by the same investigator.

Measurements for Validity Assessment

We gathered Three-Dimensional (3D) data of 100 specimens from ten species using X-ray micro-computed tomography (microCT). All specimens were CT scanned using the Quantum FX Micro CT scanner in the Imaging and Cytometry Facility, John Curtin School of Medical Research, Australian National University. Specimens for this work were obtained from the Australian National Wildlife Collection, Canberra. We excluded specimens mounted on metal rods because of the beam-hardening artefacts (e.g. streak artefacts) caused by imaging materials with vastly different X-ray opacity values (i.e. metal and bone). The microCT scanner settings used to scan each structure (bill, wing and tarsus) are given in Supplementary material Table S2. The source voltage and source current maintained was the same across all scans, with values of 90 kV and 200 μA respectively.

The resulting projections were reconstructed into a virtual stack of 2D cross section image slices using Quantum FX software interface. The reconstructed image stacks in DICOM format were converted into processed volume (*.pvl.nc format) using the Drishti importer utility in Drishti v2.6.4 (Limaye 2012). These processed volumes contained information on the voxel type and voxel size (equivalent of a 3D pixel), and were later used, rendered as 3D volumes to collect digital bill measurements and length measurements of underlying bones in the wing and tarsus.

We collated the bill data (bill length, bill width, and bill depth), wing length and tarsus length data manually for all specimens, following the method described at the beginning of this section. The surface area of bills was manually estimated using the following equation (Eq. 1) for the lateral surface area of a nearly circular elliptical cone (Greenberg et al. 2012; Luther and Greenberg 2014);

$$\text{Bill surface area} = \left(\frac{\text{BW} + \text{BD}}{4}\right) \times \text{BL} \times \pi \tag{1}$$

where BL is the bill length, BW is the bill width and BD is the depth.

Digital Bill Measurements

This involved two key steps: (1) creating a bill polygon from the processed volumes; (2) obtaining bill measurements from the polygon. We used Drishti renderer of Drishti v2.64 (Limaye 2012) to create the polygon and MeshLab v2016.12 (www.meshlab.net) to estimate surface area. In order to create the bill polygon, we first defined a clipping plane using three landmarks on the bird skull, instead of the bill surface, that were relatively stable and easy to find (Fig. 2a, b). We applied transfer functions for the clipped volume across different density thresholds, in order to have a full view of the bill (including the soft structures like the keratinous layers around the bill) without interference from feathers or from background on the bill surface. The tool ‘Mop carve’ in Drishti was used to remove regions of the volume that obscured the target region, as it was not possible to differentiate all constituent tissue types by only thresholding density values through transfer functions. Colour gradients in the transfer functions were adjusted to differentiate the keratinous layer around the nares from rest of the structure. The keratinous structure around the nares was worn to varying degrees, with differences among individuals. Therefore, this region was excluded from the surface area estimation as explained later. The bill polygon from the clipped volume, was generated using the ‘Mesh generator’ plugin in Drishti and exported as a 3D surface mesh in Polygon File Format (PLY). Default settings were used when generating meshes

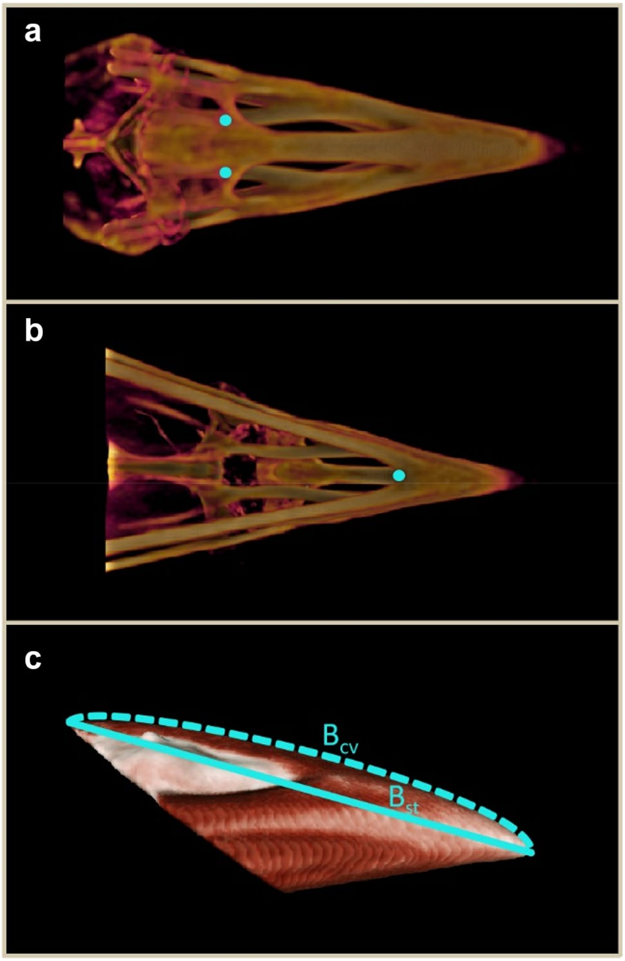

Fig. 2 Landmarks used to create the clipping plane on **a** the upper mandible and **b** lower mandible, **c** bill polygon created using Drishti software, also showing the two length measurements obtained; straight bill length (B_{st}) from dashed line and curved bill length (B_{cv}) from solid line

from the plugin, for all individuals. The default value for colour type 'VR Lut Color', was used, as it incorporates the colour gradient and opacities selected by the user to the bill polygon.

The PLY meshes were then imported to MeshLab (Cignoni et al. 2008) for surface area estimation. Here, we first coloured the surface of the polygon, excluding the region around the nares which is the dense white region of the bill polygon shown in Fig. 2c. Then the area of the coloured region was estimated as the digital surface area. Because our analysis showed that repeatability was high (ICC > 0.99) based on a subset of data (n = 46 individuals), we took only one measurement per specimen thereafter.

We collated two length measurements from the clipped bill volume i.e. straight bill length and curved bill length (taking the bill curvature into account) of the upper mandible to estimate an index of bill curvature (Fig. 2c). Here, the residuals of the simple linear model between straight length and curved length [using the function *lm()* in R] were used as the index of bill curvature.

Digital Measurements of Underlying Bones

We measured the total length of carpometacarpus and phalanges (W_{cp}) and the length of carpometacarpus (W_c) digitally from wing volume files and the length of tarsometatarsus from the processed volumes of tarsus as shown in Fig. 3, using Drishti renderer (Limaye 2012). We discarded all measurements from the CT scan data for individuals that were damaged internally (e.g. broken bones).

Statistical Analyses

Measurement Repeatability

Repeatability is calculated as: Eq. 2.

$$\mathrm{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} \tag{2}$$

where σ_α^2 is among-individual variance and σ_ε^2 is within-individual variance (Nakagawa and Schielzeth 2010). The values of repeatability range between 0 and 1, and are equal to one if measurements are without error. We calculated the repeatability for all trait measurements of all species using a linear mixed-effects model-based approach (Nakagawa and Schielzeth 2010). We used *MCMCglmm* in R (Hadfield 2010) to estimate among-individual variance (σ_α^2) and within-individual variance or the residual variance (σ_ε^2) of Eq. 2.

We performed separate models for each trait, by fitting sex, latitude, longitude, year of capture, season of collection (in bill trait models), feather wear (in wing length model) and order of handling as fixed effects (see explanation below), individual identity as a random effect with trait measurement as the dependent variable. Individual identity estimates variance between repeated measurements of individuals (σ_ε^2). In addition to variation between two measurements of the same individual, the variation between individuals (σ_α^2) can also directly affect repeatability. High repeatability values are expected from more heterogeneous groups than from homogenous groups. Therefore, factors such as sex, location (latitude, longitude), or year of capture which could increase the heterogeneity within the group (i.e. species) should be included as covariates to control for their effect and prevent false inferences on repeatability of trait measurements (Martin and Pitocchelli 1991; Wiklund 1996). Likewise, we included season of collection to control for seasonal differences in bill size due to foraging behaviours

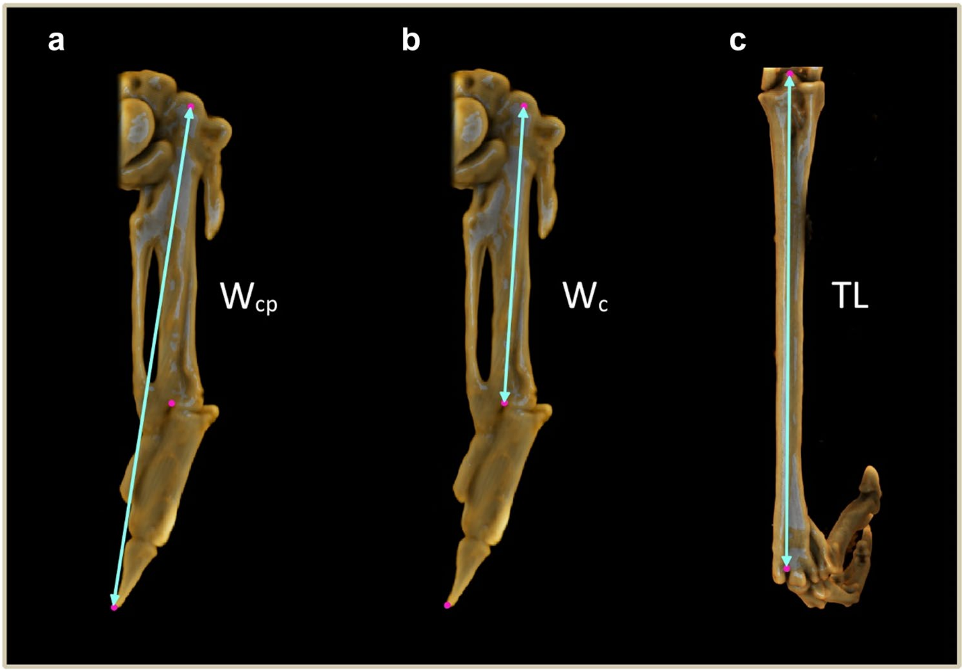

Fig. 3 **a** Total length of carpometacarpus and phalanges (W_{cp}); **b** length of carpometacarpus (W_c); **c** length of tarsometatarsus (TL)

and an index of feather wear, estimated as described in Gardner et al. 2014b to control for difference in wing length due to feather wear and abrasion (Matthysen 1989; Martin and Pitocchelli 1991; Gosler 1987). Finally, we included order of handling as a covariate, as we were interested to see if measurements taken during the first handling were consistently different to those taken during the second handling.

We carried out non-parametric Kruskal–Wallis tests followed by post-hoc Dunn tests with Bonferroni adjustment using the *dunn.test* package in R (Dinno 2017), to test for differences in repeatability between traits. For this analysis, we only used species for which we examined all traits (n = 23).

We used phylogenetic generalized linear mixed models using *MCMCglmm* in R (Hadfield 2010) to assess the linear associations between (1) repeatability, (2) among-individual variance and (3) within-individual variance with mean trait size, across all species (n = 78). Response and predictor variables were log transformed before model fitting to ensure normality. The phylogeny for this analysis was constructed, using data downloaded from the Global Phylogeny of Birds website (www.birdtree.org) (Jetz et al. 2012). We downloaded 1000 trees with the Hackett backbone (Hackett et al. 2008) and calculated a 50% majority-rule consensus phylogeny using the *consensus()* function of the *ape* packages in R (Paradis et al. 2004). We then added the phylogenetic component to the models using the function *inverseA()* from the *MCMCglmm* package (Hadfield and Nakagawa 2010). We included trait as a second random term along with species in all three models. We ran the model for 401,000 iterations using weakly informative priors, with a thinning interval of 400 and burn-in phase of 1000. We used default broad Gaussian priors for fixed effects and inverse-Wishart priors, with parameters V = 1 and nu = 0.002, for random effects. We visually examined the plots of parameter estimates to ensure model convergence.

Validity of Measurements

We first investigated the repeatability of digital measurement of bill surface area using *MCMCglmm*, as described in the previous section. We applied standardized major axis regression (SMA), also known as reduced major axis regression using the R package *smatr* (v. 3.4–8) (Warton et al. 2018) for comparisons relating to manual and digital measurements i.e. manual surface areas vs digital surface area and manual bill lengths vs digital surface area across each species. The slope test in *smatr* package was used to test whether the slopes of regression lines significantly deviated from 1. We used phylogenetic controlled generalized linear mixed model using *MCMCglmm* in R, adding bill curvature as an interaction term with the manual measurement (bill surface area or bill length), to test if associations are affected by bill curvature across species. All response and predictor variables were standardized before carrying out the analysis. We included species as a random term to account for differences in sample sizes between species.

We used the same mixed model approach, for comparisons between measurements of wing length and tarsus length with underlying bones. Here, we performed separate models for each bone length i.e. (1) total length of carpometacarpus and phalanges, (2) length of carpometacarpus, (3)

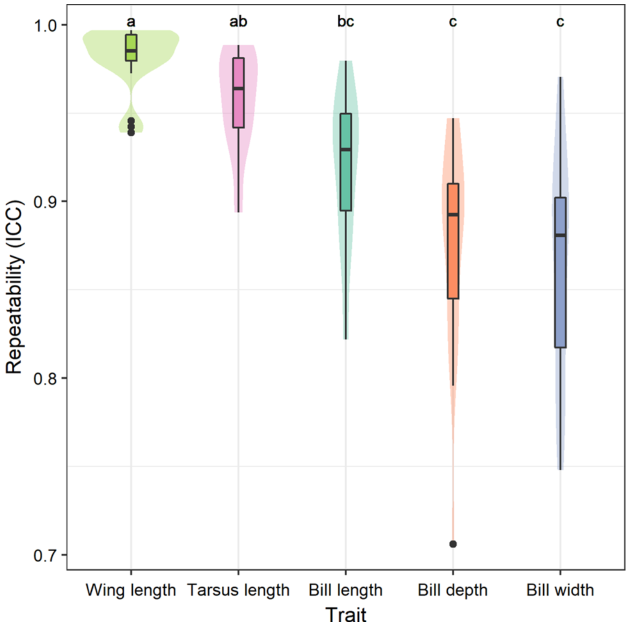

Fig. 4 Repeatability estimates (posterior mean) for five traits across 23 species; *a* wing length, *b* tarsus length; *c* bill length; *d* bill depth and *e* bill width, showing differences between traits. Shaded area shows the variation in the distribution of measurement repeatability for the trait (n = 23). Compact letter displays (*a*, *ab*, *bc*, and *c*) indicate significant differences between traits (*P* < 0.05; Dunn post-hoc test)

tarsometatarsus by fitting bone length as the response variable and surface measurement as the predictor. We included feather wear as an interaction term in wing length models, to account for effect of feather wear on wing length-bone associations. Feather wear may reduce the strength of association between wing length and wing bone.

Results

Repeatability Between Traits

The repeatability coefficients were different among the five traits (Kruskal–Wallis test: chi-squared = 69.155, df = 4, *P* ≤ 0.001; Fig. 4) and among species for a given trait (Fig. 4; Supplementary material Tables S3, S4, S5, S6, S7). The ICCs of all trait measurements are given in Supplementary material Table S3, S4, S5, S6, S7.

Wing measurements showed the highest mean ICC (0.987) of all five traits considered in this study; the values were > 0.9 for all species (Supplementary material Table S3). Mean ICC of tarsus measurements across species was also high, 0.955, with all species, except *Acanthiza ewingii* (ICC = 0.897) and *Anthochaera paradoxa* (ICC = 0.894) showing repeatability of > 0.9 (Supplementary material Table S4). Unlike wing and tarsus measurements, the repeatability of the three bill measurements were highly variable within the range of species considered (Fig. 4). Bill length showed high repeatability and was the highest (ICC = 0.919) of the three bill traits (Fig. 4; Supplementary material Table S5). The bill depth and width, the smallest traits, showed the lowest mean ICCs (Supplementary material Tables S6, S7) and were significantly different from wing and tarsus measurements (Fig. 4).

Repeatability Versus Mean Trait Size

The repeatability of measurements that were less than approximately 13 mm was highly variable, compared with measurements above 13 mm (Fig. 5a). Measurements above 13 mm were less variable and showed high repeatability with > 0.9. We found significant positive linear associations between repeatability, among-individual variance and within-individual variance with mean trait size (Table 1). The increase of within-individual variance was less pronounced, than the increase in among-individual variance (Fig. 5b). High repeatability, > 0.9, across larger traits (i.e. traits > 13 mm) are explained by higher among-individual variance, rather than within-individual variance.

Validity of Trait Measurements

All ten species showed a significant association between digital measurement and manual estimates of bill surface area. This association was strong in seven of the ten species i.e. *Acanthorhynchus tenuirostris*, *Anthochaera chrysoptera*, *Gerygone magnirostris*, *Manorina flavigula*, *Melithreptus affinis*, *Phylidonyris novaehollandiae*, *Ptilotula plumulus* (R^2 > 0.7, *P* < 0.05; Table 2). However, in two species, the slopes of the regression lines were significantly different from one.

Only two of ten species showed strong associations between manual measurements of bill length and digital surface area (R^2 > 0.7; Table 2). Slopes of one of the two species were significantly higher than 1 and over-represented bill surface area variation (Table 2). Species that showed a weak correlation (R^2 < 0.4) between manual and digital measurements of bill length (i.e. *Malurus lamberti* and *Pardalotus striatus*) displayed weak to moderate associations with both proxies (Table 2). There was no significant effect of bill curvature on associations between digital surface area and both proxies across species (Table 3). Curved bill length of all species was strongly associated with straight length in digital measurements (Table 4). Mean values of all digital and manual trait measurements of bills with standard errors are given in Supplementary material Table S8.

Almost all species showed a strong significant positive linear relationship between tarsus length measurements and

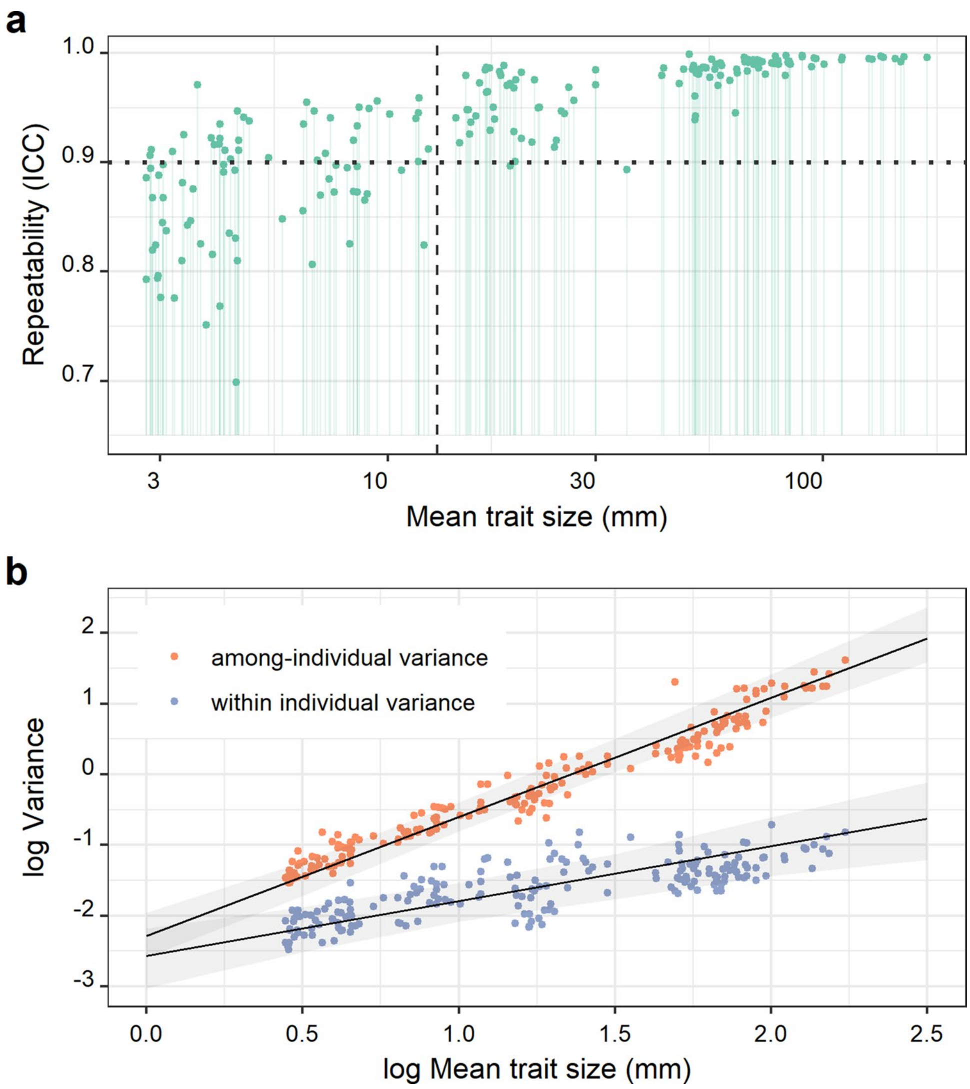

Fig. 5 Change in **a** measurement repeatability (ICC), **b** among-individual variance σ^2_α and within-individual variance σ^2_ε of traits across 78 species, with mean trait size. The scale of the x axis of **a** is log10 transformed. Dashed vertical line of **a** demarcates the traits above and below 13 mm and the dotted horizontal line demarcates the ICC above and below 0.9

Table 1 Phylogenetic generalized linear mixed model, testing associations between mean trait size and repeatability, among-individual variance, within-individual variance across five different traits of 78 species from the infraorder Meliphagides

Response	Parameter	Parameter estimate (β)	95% credibility interval	R^2
Log repeatability	Log mean trait size	**0.022**	0.001–0.057	0.222
	Intercept	**– 0.014**	– 0.122 to – 0.027	
Log among-individual variance	Log mean trait size	**1.692**	1.479–1.908	0.950
	Intercept	**– 2.289**	– 2.571 to – 1.998	
Log within-individual variance	Log mean trait size	**0.863**	0.578–1.1933	0.702
	Intercept	**– 2.652**	– 3.109 to – 2.236	

Shown are the posterior mean estimate, associated 95% credibility intervals and R^2 showing the variance explained by mean trait size. Parameter estimates that were significant ($P < 0.05$) are highlighted in bold

tarsus bone length measured digitally, across species with a slope close to 1 (Table 5). Wing length was also strongly associated with underlying wing bone lengths across the ten species (Table 5). However, the slope of wing bones on wing length was different from 1. We found no evidence for a significant effect of feather wear on wing length- bone association (Table 5).

Discussion

Of the five phenotypic traits included in this study, the measurements of larger traits, i.e. wing and tarsus length, showed consistently high repeatability (> 0.9) across species, while the repeatability of smaller traits (bill length, depth, width) varied greatly among species. Repeatability,

Table 2 The results of standardized major axis (SMA) regressions between manual estimates of bill surface area and bill lengths (predictor variable), with digital measurement of bill surface area (response variable)

	Species		n	Intercept	Slope	Lower CI	Upper CI	Adjusted R^2	Adjusted R^2 for association with BL_{st}
	Acanthorhyncus tenuirostris	SA	10	0.046	0.981	0.781	1.232	**0.920**	
		BL	10	0.172	1.415[a]	1.190	1.683	**0.954**	**0.960**
	Anthochaera chrysoptera	SA	10	0.728	0.662[a]	0.456	0.961	**0.781**	
		BL	10	0.788	1.110	0.688	1.793	**0.627**	**0.558**
	Gerygone magnirostris	SA	9	0.208	0.887	0.581	1.352	**0.763**	
		BL	9	0.669	1.057	0.640	1.747	**0.657**	**0.744**
	Malurus lamberti	SA	8	0.375	0.828	0.454	1.510	**0.592**	
		BL	8	0.372	1.342	0.576	3.125	0.096	0.102
	Manorina flavigula	SA	10	−0.152	1.025	0.701	1.499	**0.772**	
		BL	10	−0.030	1.715[a]	1.097	2.681	**0.680**	**0.748**
	Melithreptus affinis	SA	8	0.312	0.799	0.552	1.156	**0.857**	
		BL	8	0.326	1.408	0.904	2.193	**0.790**	**0.582**
	Pardalotus striatus	SA	10	−0.658	1.428	0.839	2.429	**0.533**	
		BL	10	0.760	1.099	0.523	2.311	0.005	0.449
	Phylidonyris novaehollandiae	SA	10	−0.014	1.018	0.748	1.386	**0.852**	
		BL	10	0.155	1.564	0.892	2.744	**0.472**	**0.715**
	Ptilotula plumulus	SA	9	0.533	0.695[a]	0.581	0.832	**0.959**	
		BL	9	0.701	1.052	0.655	1.692	**0.696**	**0.625**
	Smicrornis brevirostris	SA	9	0.136	0.931	0.579	1.497	**0.696**	
		BL	9	0.648	1.084	0.580	2.025	**0.444**	**0.646**

All variables were log transformed. SA: surface area estimated using mathematical formula. BL: bill length from bill tip to the feather base. Final column of the table shows the association between manual measurement of bill length (BL) and digital measurement of straight bill length (B_{st}). Associations that are significant ($P < 0.05$) are highlighted in bold

[a] Slopes significantly different from 1 at 0.05 level of significance

among- and within-individual variances were dependent on trait size. However, among-individual variance increased at a higher rate than within-individual variance, resulting in high repeatability coefficients for larger traits. Smaller traits were subject to relatively high within-individual variance that is associated with measurement precision, compared to among-individual variance of those traits. The surface area of bills estimated from manual measurements (length, depth and width) was strongly correlated with the digital estimates, although less so in the two species that had the smallest bill lengths. However, bill length alone was a relatively poor correlate of digital bill surface area. Wing and tarsus lengths were strongly associated with the lengths of underlying bones, indicating that they may be useful as indices of structural body size.

Table 3 The results of phylogenetic generalized linear mixed models showing the effect of bill curvature on surface area predictions using two proxies: mathematical formula based estimations (SA) and bill length (BL) across ten species of Meliphagides

Proxy	Parameter	Parameter estimate	95% credibility interval	*P* value
Formula based estimations	Intercept	−0.044	−0.601 to 0.436	0.854
	Surface area (manual)	0.871	0.775–0.964	**<0.001**
	Curvature	0.057	0.013–0.101	**0.020**
	Surface area × curvature	0.022	−0.011 to 0.058	0.192
Bill length	Intercept	−0.074	−0.714 to 0.543	0.824
	Bill length (manual)	0.869	0.753–0.992	**<0.001**
	Curvature	0.084	0.031–0.135	**0.002**
	Bill length × curvature	0.048	−0.002 to 0.094	0.066

Shown are the posterior mean estimate, associated 95% credibility intervals and *P* values. Significant values ($P > 0.05$) are highlighted in bold

Table 4 The results of standardized major axis (SMA) regression between straight length (BL_{st}) and curved lengths (BL_{cv}) of bill polygons for ten species of Meliphagides

Species	n	Lower CI	Upper CI	R^2
Acanthorhyncus tenuirostris	10	0.946	1.012	**0.998**
Anthochaera chrysoptera	10	0.869	1.028	**0.989**
Gerygone magnirostris	9	0.907	1.107	**0.988**
Malurus lamberti	8	0.826	1.673	**0.870**
Manorina flavigula	10	0.937	1.200	**0.977**
Melithreptus affinis	8	0.928	1.113	**0.992**
Pardalotus striatus	10	0.475	1.081	**0.731**
Phylidonyris novaehollandiae	10	0.980	1.041	**0.999**
Ptilotula plumulus	9	0.946	1.160	**0.987**
Smicrornis brevirostris	9	0.802	1.184	**0.952**

Significant associations are highlighted in bold ($P < 0.05$)

Repeatability of Measurements

Our findings of significantly high repeatability coefficients for measurements of larger traits across 24 species (i.e. wing length and tarsus), compared with smaller traits (i.e. bill length, width and depth), is in line with findings of previous studies that only use measurements of a single species to compare coefficients between traits (Lougheed et al. (1991) on American Coots, *Fulica American* and Totterman (2016) on Short-tailed Shearwaters, *Puffinus tenuirostris*). We demonstrate that the repeatability coefficients (estimated using a mixed model approach) vary not only between traits, but also between closely related species for a given trait which, to our knowledge, has not been previously tested using multiple species. Such differences, could have been partially caused by differences in the clarity of land-marks due to structural differences between species, affecting precision of the measurement. Further, differences in trait size may also have contributed to variation in repeatability between species for a particular trait. This suggests the importance of quantifying repeatability of a trait for each species under study (especially for small traits), even when the same measurer is employed to collect measurements from all individuals. The repeatability coefficients are highly variable for traits below 13 mm, whereas for traits above 13 mm they are close to 1.0. Therefore, caution is required when working with traits below 13 mm regardless of trait type or species. Muñoz-Muñoz and Perpiñán (2010) reported a decrease in

Table 5 Phylogenetic generalized linear mixed model, testing associations between manual measurement of wing length, tarsus length with underlying bones i.e. total length of carpometacarpus and phalanges (W_{cp}), length of carpometacarpus (W_c), tarsometatarsus (TL) across ten species from the infraorder Meliphagides

Response	Parameter	Parameter estimate	95% credibility interval	Marginal R^2
TL	Tarsus length	**0.966**	0.893–1.037	0.948
	Intercept	**2.141**	0.733–3.809	
W_{cp}	Wing length	**0.169**	0.141–0.198	0.793
	Feather wear	0.076	−0.075 to 0.225	
	Wing length × feather wear	0.000048	−0.0017 to 0.0018	
	Intercept	**3.210**	0.195–6.666	
W_c	Wing length	**0.098**	0.080–0.116	0.870
	Feather wear	0.025	−0.072 to 0.119	
	Wing length × feather wear	0.0002	−0.0009 to 0.001	
	Intercept	1.133	−0.471 to 2.852	

Shown are the posterior mean estimate, associated 95% credibility intervals and *P*-values. Significant values ($P > 0.05$) are highlighted in bold percentage measurement error (1-repetability) with increasing trait size, across 157 skeletal characters from a wide range of taxa. In their study, measurement error declined when characters reached a mean size of 10 mm.

Our results clearly indicate that the consistently high measurement repeatability for traits > 13 mm was not due to higher precision in measurement, but resulted from relatively higher among- individual variance. Both within- and among-individual variances showed a strong significant increase with mean trait size across 78 species of Meliphagides, but at a higher rate in the later, resulting in high repeatability coefficients for larger traits. Our findings are supported by Muñoz-Muñoz and Perpiñán (2010), who also showed that the decline in percentage measurement error associated with increasing trait size was the result of rapid increase in among-individual variation using skeletal characters. Hallgrímsson and Maiorana (1999) suggest three explanations for the positive relationship between mean trait size and among individual variation of body size (body mass) in species of mammals and birds; explanations were based on niche width, metabolic scaling and the scaling of body mass components. Although the latter two explanations are exclusive to the variability of body mass, niche width can inform variability—trait size relationships. Previous authors reported an increase in intra-specific trait variability with an increase in niche width (Van Valen 1965; Rothstein 1973). For example, Van Valen (1965) reported a two times higher variabilities for bill length and width for birds in more variable niches, suggesting that high trait variability is an adaptation for utilizing diverse niche space. While trait variability is linked to niche width, the size of the species may also be associated with niche width. For example, Hallgrimsson and Maiorana (1999) demonstrate an indirect association between trait variability and size. Smaller species should be more specialized with narrower niche widths, based on the right skewed frequency distribution of body size for birds (Blackburn and Gaston 1994; O'gorman and Hone 2012), perhaps caused by a reduction in inter-specific competition for niche space with increasing size. Narrower niche widths may underlie lower variability in trait size among small species. High trait variability among large species, on the other hand, may be caused by broader niche width elicited by stable population size (Hallgrimsson and Maiorana 1999).

The variability introduced through measurement error can have two types of consequences, depending on whether the variable measured with error is a response variable or a predictor. First, if the variable measured with error is the response variable of a model, increased measurement error tends to increase the variability in parameter estimates as well as measures of uncertainty (standard error, span of confidence interval, *P* values). Despite increased rates of false positive and false negative results, measurement error in the response produces no bias, assuming measurement error is symmetrical (Hyslop and Imbens 2001). On the other hand, if the variable measured with error is a predictor, increased measurement error tends to shrink parameter estimates toward zero, thus introducing bias (Fuller 1987; Hyslop and Imbens 2001; Carriquiry 2015). Measures of uncertainty tend to decrease too, but at a slower rate than the decrease in the absolute value of the parameter point estimate, meaning that the rate of false negative results tends to increase (Carriquiry 2015). For instance, estimates of the effect of climate change on changes in phenotypic traits should not be biased by measurement error in trait measurements, although the error decreases the precision of estimates and increases the risk of false positive or negative results. On the other hand, estimates of the effect of changes in phenotypic traits on demographic rates [reproduction or survival (Kruuk et al. 2002; O'Sullivan et al. 2019)] will tend to be biased downward by measurement error in trait measurements and false negative results will tend to be more common.

Whether the trait measured with error is used as response or predictor, the impact of measurement quality on the accuracy of estimations depends on how much variation exists between individuals in the study population, relative to the amount of measurement error. Therefore, measurement error, if not accounted for, will weaken statistical inference for predictors that explain a relatively small amount of variation in the response variable. In that case, even a small amount of variation introduced by measurement error may become critical. When substantial measurement error is identified, authors should consider correcting for measurement error, either by adjusting parameter estimates post-hoc, or preferably by modelling the measurement error directly in the analysis (Carriquiry 2015; Ponzi et al. 2018). This correction is especially important when the trait measured with error is used as a predictor. As a preventive measure, we suggest testing repeatability for a randomly selected sample of the individuals under study, before designing data collection protocols. Adapting sampling protocols to reduce expected measurement error, relative to between-individual differences, will not only improve the accuracy of parameter estimates, but also optimize costs involved.

Repeatability of a measurement can be improved by minimizing the within-individual variance via improving measurement precision. This can be done by averaging measurements of some within-individual replicates, using instruments with greater precision, employing experienced measurers or improving their skills prior to data collection. These strategies should improve repeatability of smaller traits, where overall repeatability is highly influenced by within-individual variance. However, little can be achieved for large traits, as most of the variance in measurements comprise among-individual variance. Yezerinac et al. (1992) showed that the reduction in variance of the mean, with each additional measurement (replicate), was relatively low for traits with low measurement error (or measurements with high repeatability). We caution that the repeatability of measurements may change depending on whether they are taken from live or dead individuals; measurement precision is likely to be lower for live individuals that struggle in hand.

Validity of Measurements

We found that the equation for the surface area of near elliptical cone was a good proxy for bill surface area and captured size variation within our sample of species. However, the quality of component trait measurements can affect estimates of bill surface area based on the equation. Our results for the species with the smallest bills (*Malurus lamberti*, *Pardalotus striatus* and *Smicrornis brevirostris*, with bill length range across individuals: 4.76–10.40 mm, bill width: 3.14–4.46 mm, bill depth: 2.5–4.16 mm) showed only a moderate association (R^2 between 0.4 and 0.7) between digital and manual estimates of bill surface area, perhaps due to high sensitivity to errors in manual measurements.

Bill length measurements of *Malurus lamberti* and *Pardalotus striatus* were inaccurate, evident from poor associations between manual and digital length measurements, and may have affected the surface area estimates. As discussed in previous studies, defining the precise location of the feather base for the bill length measurement is sometimes difficult, especially when feathers are worn (Winkler 1998). This might have affected the bill length measurements for these species. The accuracy of surface area estimates can be improved by reducing the errors in component manual measurements. Besides improving repeatability, researchers can consider alternative measurements of bill traits (e.g. bill length from bill tip to nares instead of feather base (Baldwin et al. 1931)) with high accuracy for problematic species.

In all species, but one (*Acanthorhynchus tenuirostris*), the mathematical formula based on multiple measurements (length, depth and width) was more accurate for predicting variation in bill surface area than the linear measurement of bill length alone. On average, there was a 23% reduction in R^2 when using bill length as the proxy to describe variation in bill surface area. Perktaş and Gosler (2010) indicated that multivariate approaches are better than univariate measures for size estimation. From the results of this study, none of the proxies of bill surface area i.e. mathematical formula based on estimations or the bill length, are impacted by curvature of the upper mandible. Curved bill length was almost fully explained by the equivalent linear measurement (straight length) in all species ($R^2 > 0.7$). Nevertheless, given that the diversity of bill shapes and sizes was relatively limited in our study, researchers need to consider the efficacy of the technique for species with different bill shapes.

Although accurate estimation of bill surface area is possible with microCT data following the procedure demonstrated in this study, its use will be limited due to high costs, tediousness of process as well as the availability of scanning facilities (Gould 2014; Openshaw et al. 2017). In particular, the approach is less feasible in large scale studies that examine several hundred specimens or for field-based studies. However, the approach will be extremely useful when testing the validity of low-cost alternatives for different species. Friedman et al. (2017) estimated bill surface area using a 2D image in their study. However, as pointed out in previous studies, flattening of 3D to a 2D image will lose some information related to shape variation in complex structures, in turn affecting surface area approximations (Buser et al. 2018; Andrea and Chiappelli 2019). Our study demonstrates the potential of using 3D imaging technology to assess the validity of trait measurements, another use of this technology for biologists (Semple et al. 2019).

We clearly showed that the manual tarsus length measurement was strongly associated with the underlying tarsometatarsus indicating that it accurately represents tarsus size. Some authors have suggested that the surface measurements of tarsus could be unreliable on museum specimens, due to a lack of clarity of landmarks, but this is not supported by our findings (Perktaş and Gosler 2010). Across species, wing length is a strong predictor of wing bone length, capturing a large proportion of variation ($R^2 = 0.793$ for W_{cp} and $R^2 = 0.870$ for W_c), but is less ideal than tarsometatarsus as a measure of tarsus length, because the proportion of variation captured is lower and the slopes of wing bones on wing length are less than one (0.169 for W_{cp} and 0.098 for W_c). Indeed, wing length does not depend only on bone length, but also on feather length, and the allometric relationship between wing bone and wing feather length may vary among species (Nudds 2007; Nudds et al. 2011). Findings suggests that wing length is a reasonable proxy for structural body size, for cross-species studies, assuming the wing bones themselves correlate with the size of overall skeletal frame, although testing this was outside the scope of this study. The ecological context in which these traits are used needs to be considered on a case by case basis. Feather wear and abrasion had no effect on the wing length, wing bone association in this study.

Acknowledgements We thank Jesse Smith and Roellen Little for undertaking all wing measurements, and the staff at The Australian National Wildlife Collection, Australian Museum, Museum Victoria, Queensland Museum, Queen Victoria Museum and Art Gallery, South Australian Museum, Western Australian Museum, Tasmanian Museum and Art gallery for providing access to specimens. We also thank Emma Sherratt for general advice on 3D imaging and analysis, Terry Neeman at ANU, for advice on statistical analyses, and Cathy Gillespie at the John Curtin School of Medical Research, ANU, for technical assistance in 3D imaging and Peter Marsack for allowing us to use his bird illustrations. We are also grateful to an anonymous reviewer, Lynda Sharpe and Heather McGinness for helpful comments on the draft manuscript.

Author Contributions KS and JG conceived the idea; KS, JG and MVG developed the methods; KS, TB and MRES designed the analyses; KS collected and analysed the data and wrote the manuscript; all authors read the manuscript and provided feedback.

Funding This work was partly funded by The Commonwealth Scientific and Industrial Research Organisation (PHD17-27) and the Recent Ecological Change in Australia project through the Department of Agriculture, Water and Environment (DAWE). KS was supported by ANU HDR Fee Merit Scholarship, University Research Scholarship and CSIRO research plus top-up scholarship (PHD17-27). MVG is supported by the University of Calgary's Banting Research Allowance and an ACHRI Postdoctoral Fellowship.

Data Availability Please contact corresponding author for data requests.

Code Availability Please contact corresponding author.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no competing interests.

References

Aldrich, J. W., & James, F. C. (1991). Ecogeographic variation in the American Robin (*Turdus migratorius*). *The Auk, 108*(2), 230–249.

Anderson, A. M., Friis, C., Gratto-Trevor, C. L., Morrison, R. I. G., Smith, P. A., & Nol, E. (2019). Consistent declines in wing lengths of Calidridine sandpipers suggest a rapid morphometric response to environmental change. *PLoS ONE, 14*(4), e0213930. https://doi.org/10.1371/journal.pone.0213930.

Andrea, C., & Chiappelli, M. (2019). How flat can a horse be? Exploring 2D approximations of 3D crania in equids. *bioRxiv*. https://doi.org/10.1101/772624.

Baldwin, S. P., Oberholser, H. C., & Worley, L. G. (1931). *Measurements of birds*. Cleveland: Cleveland Museum of Natural History.

Benítez-díaz, H. (1993). Geographic variation in morphology and coloration of the Acorn Woodpecker (*Melanerpes formicivorus*). *Condor, 95,* 63–71.

Billionnet, A. (2018). Phylogenetic conservation prioritization with uncertainty. *Biodiversity and Conservation, 27,* 3137–3153. https://doi.org/10.1007/s10531-018-1593-z.

Blackburn, T. M., & Gaston, K. J. (1994). Animal body size distributions: Patterns, mechanisms and implications. *Trends in Ecology and Evolution, 9,* 471–474. https://doi.org/10.1016/0169-5347(94)90311-5.

Blackwell, G. L., Bassett, S. M., & Dickman, C. R. (2006). Measurement error associated with external measurements commonly used in small-mammal studies. *Journal of Mammalogy, 87*(2), 216–223. https://doi.org/10.1644/05-mamm-a-215r1.1.

Barrett, R. T., Peterz, M., Furness, R. W., & Durinck, J. (1989). The variability of biometric measurements. *Ringing & Migration, 10,* 13–16. https://doi.org/10.1080/03078698.1989.9676001.

Buser, T. J., Sidlauskas, B. L., & Summers, A. P. (2018). 2D or not 2D? Testing the utility of 2D vs. 3D landmark data in geometric morphometrics of the sculpin subfamily *Oligocottinae* (Pisces; *Cottoidea*). *The Anatomical Record, 301*(5), 806–818. https://doi.org/10.1002/ar.23752.

Campbell-Tennant, D. J. E., Gardner, J. L., Kearney, M. R., & Symonds, M. R. E. (2015). Climate-related spatial and temporal variation in bill morphology over the past century in Australian parrots. *Journal of Biogeography, 42*(6), 1163–1175. https://doi.org/10.1111/jbi.12499.

Carriquiry, A. L. (2015). Measurement Error models. In J. D. Wright (Ed.), *International encyclopaedia of the social and behavioural sciences* (2nd ed., pp. 850–855). Oxford: Elsevier.

Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G. (2008). Meshlab: an open-source mesh processing tool. In *Proceedings of the Eurographics Italian Chapter Conference* (pp. 129–136). Salerno, Italy.

Dinno, A. (2017). dunn.test: Dunn's test of multiple comparisons using rank sums. R package version 1.3.5. http://CRAN.R-project.org/package=dunn.test.

Friedman, N. R., Harmáčková, L., Economo, E. P., & Remeš, V. (2017). Smaller beaks for colder winters: Thermoregulation drives beak size evolution in Australasian songbirds. *Evolution*. https://doi.org/10.1111/evo.13274.

Freeman, S., & Jackson, W. M. (1990). Univariate metrics are not adequate to measure avian body size. *The Auk, 107*(1), 69–74. https://doi.org/10.1093/auk/107.1.69.

Flinks, H., & Salewski, V. (2012). Quantifying the effect of feather abrasion on wing and tail lengths measurements. *Journal of Ornithology, 153*(4), 1053–1065. https://doi.org/10.1007/s10336-012-0834-2.

Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.

Gardner, J. L., Amano, T., Backwell, P. R. Y., Ikin, K., Sutherland, W. J., & Peters, A. (2014a). Temporal patterns of avian body size reflect linear size responses to broad scale environmental change over the last 50 years. *Journal of Avian Biology, 45*(6), 529–535. https://doi.org/10.1111/jav.00431.

Gardner, J. L., Amano, T., Mackey, B. G., Sutherland, W. J., Clayton, M., & Peters, A. (2014b). Dynamic size responses to climate change: Prevailing effects of rising temperature drive long-term body size increases in a semi-arid passerine. *Global Change Biology*. https://doi.org/10.1111/gcb.12507.

Gardner, J. L., Symonds, M. R. E., Joseph, L., Ikin, K., Stein, J., & Kruuk, L. E. B. (2016). Spatial variation in avian bill size is associated with humidity in summer among Australian passerines. *Climate Change Responses, 3*(1), 11. https://doi.org/10.1186/s40665-016-0026-z.

Gardner, J. L., Amano, T., Peters, A., Sutherland, W. J., Mackey, B., Joseph, L., et al. (2019). Australian songbird body size tracks climate variation: 82 species over 50 years. *Proceedings of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rspb.2019.2258.

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

Gosler, A. G. (1987). Pattern and process in the bill morphology of the Great Tit *Parus major*. *Ibis, 129*(s2), 451–476. https://doi.org/10.1111/j.1474-919X.1987.tb08234.x.

Gould, F. D. (2014). To 3D or not to 3D, that is the question: Do 3D surface analyses improve the ecomorphological power of the distal femur in placental mammals? *PLoS ONE, 9*(3), e91719. https://doi.org/10.1371/journal.pone.0091719.

Greenberg, R., Cadena, V., Danner, R. M., & Tattersall, G. J. (2012). Heat loss may explain bill size differences between birds occupying different habitats. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0040933.

Hackett, S. J., Kimball, R. T., Reddy, S., Bowie, R. C. K., Braun, E. L., et al. (2008). A phylogenomic study of birds reveals their evolutionary history. *Science, 320,* 1763–1768. https://doi.org/10.1126/science.1157704.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*(2), 1–22.

Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology, 23*(3), 494–508. https://doi.org/10.1111/j.1420-9101.2009.01915.x.

Hallgrímsson, B., & Maiorana, V. (1999). Variability and size in mammals and birds. *Biological Journal of the Linnean Society, 70*(4), 571–595. https://doi.org/10.1111/j.1095-8312.2000.tb00218.x.

Harper, D. G. C. (1994). Some comments on the repeatability of measurements. *Ringing and Migration, 15*(2), 84–90. https://doi.org/10.1080/03078698.1994.9674078.

Harris, E. F., & Smith, R. N. (2009). Accounting for measurement error: A critical but often overlooked process. *Archives of Oral Biology, 54,* S107–S117. https://doi.org/10.1016/j.archoralbio.2008.04.010.

Hick, J., & Emmerson, R. (2014). *RCGP AKT: Research, epidemiology and statistics*. Florida, Boca Raton: CRC Press.

Hyslop, R., & Imbens, G. (2001). Bias from classical and other forms of measurement error. *Journal of Business and Economic Statistics, 19*(4), 475–481.

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature, 491,* 444–448. https://doi.org/10.1038/nature11631.

Kiat, Y., & Sapir, N. (2017). Age-dependent modulation of songbird summer feather molt by temporal and functional constraints. *The American Naturalist, 189*(2), 184–195. https://doi.org/10.1086/690031.

Kruuk, L. E. B., Slate, J., Pemberton, J. M., Brotherstone, S., Guinness, F., & Clutton-Brock, T. (2002). Antler size in red deer: Heritability and selection but no evolution. *Evolution, 56*(8), 1683–1695. https://doi.org/10.1111/j.0014-3820.2002.tb01480.x.

Leverton, R. (1989). Wing length changes in individually-marked Blackbirds *Turdus merula* following moult. *Ringing and Migration, 10*(1), 17–25. https://doi.org/10.1080/03078698.1989.9676002.

Limaye, A. (2012). Drishti: a volume exploration and presentation tool. In: *Proceedings of SPIE 8506, Developments in X-ray Tomography VIII, 85060X*.

Lougheed, S. C., Arnold, T. W., & Bailey, R. C. (1991). Measurement error of external and skeletal variables in birds and its effect on principal components. *The Auk, 108,* 432–436.

Luther, D., & Greenberg, R. (2014). Habitat type and ambient temperature contribute to bill morphology. *Ecology and Evolution*. https://doi.org/10.1002/ece3.911.

Martin, J. L., & Pitocchelli, J. (1991). Relation of within-population phenotypic variation with sex, season, and geography in the blue tit. *The Auk, 108*(4), 833–841.

Matthysen, E. (1989). Seasonal variation in bill morphology of nuthatches *Sitta europaea*: dietary adaptations or consequences. *Ardea, 77,* 117–125.

Mawdsley, J. R., O'Malley, R., & Ojima, D. S. (2009). A review of climate-change adaptation strategies for wildlife management and biodiversity conservation. *Conservation Biology, 23*(5), 1080–1089. https://doi.org/10.1111/j.1523-1739.2009.01264.x.

Menkhorst, P., Rogers, D. I., Clarke, R., Davies, J. N., Marsack, P., Franklin, K., & CSIRO. (2017). *The Australian bird guide*. Clayton: CSIRO Publishing.

McLean, E. H., Prober, S. M., Stock, W. D., Steane, D. A., Potts, B. M., Vaillancourt, R. E., & Byrne, M. (2014). Plasticity of functional traits varies clinally along a rainfall gradient in *Eucalyptus tricarpa*. *Plant, Cell and Environment, 37*(6), 1440–1451. https://doi.org/10.1111/pce.12251.

Mooers, A. Ø., Faith, D. P., & Maddison, W. P. (2008). Converting endangered species categories to probabilities of extinction for phylogenetic conservation prioritization. *PLoS ONE, 3,* e3700. https://doi.org/10.1371/journal.pone.0003700.

Muñoz-Muñoz, F., & Perpiñán, D. (2010). Measurement error in morphometric studies: Comparison between manual and computerized methods. *Annales Zoologici Fennici, 47*(1), 46–56.

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews, 85*(4), 935–956. https://doi.org/10.1111/j.1469-185X.2010.00141.x.

Nudds, R. L. (2007). Wing-bone length allometry in birds. *Journal of Avian Biology*, *38*(4), 515–519.

Nudds, R. L., & Oswald, S. A. (2007). An interspecific test of Allen's rule: Evolutionary implications for endothermic species. *Evolution, 61*(12), 2839–2848. https://doi.org/10.1111/j.1558-5646.2007.00242.x.

Nudds, R. L., Kaiser, G. W., & Dyke, G. J. (2011). Scaling of avian primary feather length. *PLoS ONE, 6*(2), e15665–e15665.

O'gorman, E. J., & Hone, D. W. E. (2012). Body size distribution of the dinosaurs. *PLoS ONE, 7*(12), e51925. https://doi.org/10.1371/journal.pone.0051925.

Openshaw, G. H., D'Amore, D. C., Vidal-García, M., & Keogh, J. S. (2017). Combining geometric morphometric analyses of multiple 2D observation views improves interpretation of evolutionary allometry and shape diversification in monitor lizard (*Varanus*) crania. *Biological Journal of the Linnean Society, 120*(3), 539–552. https://doi.org/10.1111/bij.12899.

O'Sullivan, R. J., Aykanat, T., Johnston, S. E., Kane, A., Poole, R., Rogan, G., et al. (2019). Evolutionary stasis of a heritable morphological trait in a wild fish population despite apparent directional selection. *Ecology and Evolution, 9*(12), 7096–7111. https://doi.org/10.1002/ece3.5274.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics, 20,* 289–290. https://doi.org/10.1093/bioinformatics/btg412.

Perktaş, U., & Gosler, A. G. (2010). Measurement error revisited: Its importance for the analysis of size and shape of Birds. *Acta Ornithologica, 45*(2), 161–172. https://doi.org/10.3161/000164510X551309.

Ponzi, E., Keller, L. F., Bonnet, T., & Muff, S. (2018). Heritability, selection, and the response to selection in the presence of phenotypic measurement error: Effects, cures, and the role of repeated measurements. *Evolution, 72*(10), 1992–2004. https://doi.org/10.1111/evo.13573.

Prober, S. M., Byrne, M., McLean, E. H., Steane, D. A., Potts, B. M., Vaillancourt, R. E., & Stock, W. D. (2015). Climate-adjusted provenancing: A strategy for climate-resilient ecological restoration. *Frontiers in Ecology and Evolution*. https://doi.org/10.3389/fevo.2015.00065.

Rising, J. D., & Somers, K. M. (1989). The measurement of overall body size in birds. *The Auk, 106*(4), 666–674. https://doi.org/10.1093/auk/106.4.666.

Rothstein, S. I. (1973). The Niche-variation model-is it valid? *The American Naturalist, 107,* 598–620. https://doi.org/10.1086/282862.

Salewski, V., Siebenrock, K.-H., Hochachka, W. M., Woog, F., & Fiedler, W. (2014). Morphological change to birds over 120 years is not explained by thermal adaptation to climate change. *PLoS ONE, 9*(7), e101927–e101927. https://doi.org/10.1371/journal.pone.0101927.

Semple, T. L., Peakall, R., & Tatarnic, N. J. (2019). A comprehensive and user-friendly framework for 3D-data visualisation in invertebrates and other organisms. *Journal of Morphology, 280*(2), 223–231. https://doi.org/10.1002/jmor.20938.

Senar, J. C., & Pascual, J. (1997). Keel and tarsus length may provide a good predictor of avian body size. *Ardea, 85,* 269–274.

Stephens, R. B., Karau, K. H., Yahnke, C. J., Wendt, S. R., & Rowe, R. J. (2015). Dead mice can grow: Variation of standard external mammal measurements from live and three postmortem body states. *Journal of Mammalogy, 96*(1), 185–193. https://doi.org/10.1093/jmamma/gyu022.

Stettenheim, P. R. (2015). The integumentary morphology of modern birds: An overview. *Integrative and Comparative Biology, 40*(4), 461–477. https://doi.org/10.1093/icb/40.4.461.

Sulloway, F. J., & Kleindorfer, S. (2013). Adaptive divergence in a ground finch. *Biological Journal of the Linnean Society, 110,* 45–59. https://doi.org/10.1111/bij.12108.

Symonds, M. R. E., & Tattersall, G. J. (2010). Geographical variation in bill size across bird species provides evidence for Allen's rule. *The American Naturalist*. https://doi.org/10.1086/653666.

Temeles, E. J., Koulouris, C. R., Sander, S. E., & Kress, W. J. (2009). Effect of flower shape and size on foraging performance and trade-offs in a tropical hummingbird. *Ecology, 90,* 1147–1161. https://doi.org/10.1890/08-0695.1.

Thiele, K. R. (1993). The holy grail of the perfect character: The cladistic treatment of morphometric data. *Cladistics, 9,* 275–304. https://doi.org/10.1006/clad.1993.1020.

Totterman, S. L. (2016). Random measurement error and specimen shrinkage in short-tailed shearwaters *Puffinus tenuirostris*. *Marine Ornithology, 44,* 11–20.

Van Valen, L. (1965). Morphological variation and width of ecological niche. *The American Naturalist, 99,* 377–390.

Warton, D. I., Duursma, R. A., Falster, D. S., & Taskinen, S. (2018). smatr:(Standardised) Major Axis Estimation and Testing Routines. R package version 3.4-8. https://CRAN.R-project.org/package=smatr.

Wiens, J. J. (2004). The Role of Morphological Data in Phylogeny Reconstruction. *Systematic Biology, 53*(4), 653–661. https://doi.org/10.1080/10635150490472959.

Wiklund, C. G. (1996). Body length and wing length provide univariate estimates of overall body size in the merlin. *The Condor, 98*(3), 581–588. https://doi.org/10.2307/1369570.

Winkler, K. (1998). Suggestions for measuring external characters of birds. *Ornitologia Neotropical, 9,* 23–30.

Yezerinac, S. M., Lougheed, S. C., & Handford, P. (1992). Measurement error and morphometric studies: Statistical power and observer experience. *Systematic Biology, 41*(4), 471–482. https://doi.org/10.2307/2992588.