Functional Ecology

BRITISH ECOLOGICAL SOCIETY

Functional Ecology 2017, **31**, 685–696 doi: 10.1111/1365-2435.12779

Metabolic rates, and not hormone levels, are a likely mediator of between-individual differences in behaviour: a meta-analysis

Benedikt Holtmann*[,1], Malgorzata Lagisz[1,2] and Shinichi Nakagawa[1,2]

[1]*Department of Zoology, University of Otago, 340 Great King Street, Dunedin 9016, New Zealand; and* [2]*Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia*

Summary

1. Consistent individual differences in hormone levels and metabolic rates have been proposed to be potential state variables underlying consistent individual differences in behaviour (i.e. animal personality). However, it remains unclear whether either one alone or both of these potential state variables could be an underlying driver of animal personality.
2. We address this question using meta-analyses of published data from bird species. We hypothesized that state variables that mediate individual differences in behaviour would display similar or higher repeatability estimates than behavioural traits. To test this hypothesis, we quantified repeatability estimates of hormone levels, metabolic rates and behavioural traits.
3. We found moderate to high mean repeatability estimates for both metabolic rates and behavioural traits, but low repeatability estimates for hormone levels. These findings indicate that metabolic rates likely represent an important mechanism for generating adaptive personality differences in behaviour.
4. We also show that: (i) for hormones and behaviour, repeatability decreased with increasing interval time between two measurements; (ii) males and females differed in repeatability for behavioural traits; (iii) stress-induced hormone levels were more repeatable than baseline levels.
5. Future studies are now required to determine the direction of the association between metabolic rates and behavioural traits. At the same time, these studies should try to investigate which of the proposed mechanisms is responsible for the relationship between state variable and state-dependent behaviour.
6. In addition, we encourage researchers to report the coefficient of variation for between-individual variance (CV_B) along with repeatability estimates because these two indices carry different information. We discuss how CV_B may better facilitate future comparative studies, including meta-analyses.

Key-words: adaptive variation, feedback loops, hormones, metabolism, physiological traits, repeatability, state-dependent personality, systematic review

Introduction

Over the last two decades, the study of consistent individual differences in behaviour (i.e. animal personalities) has become a major area of research in behavioural ecology (Sih *et al*. 2004; Réale *et al*. 2007, 2010a). Although a large number of empirical studies examined animal personalities in various taxonomic groups (Gosling 2001; Bell, Hankison & Laskowski 2009), it is still poorly understood which underlying mechanisms are responsible for the maintenance of variation in animal personality (e.g. Biro & Stamps 2008; Réale *et al*. 2010b; Wolf & Weissing 2010; Sih *et al*. 2015). There is growing awareness, however, that consistent behavioural differences may have evolved as an adaptive consequence of consistent interindividual differences in 'states' (i.e. state-dependent personality, reviewed

*Correspondence author. Division of Evolutionary Biology, Ludwig-Maximilians-University of Munich, Großhardener Str. 2, 82152 Planegg-Martinsried, Germany.
E-mail: benedikt.holtmann@gmail.com

in Dingemanse & Wolf 2010; Wolf & Weissing 2010; Sih *et al.* 2015). The 'state' of an individual refers to any feature, extrinsic (e.g. conspecifics or predators) or intrinsic (e.g. metabolism or individual experience), that has an effect on an individual's behaviour and fitness (Dingemanse & Wolf 2010; Wolf & Weissing 2010; Sih *et al.* 2015).

State-dependent personality models are based on the major assumption that individual differences in states lead to individual differences in behaviour (Wolf & Weissing 2010). More specifically, state-dependent behaviour will exhibit time consistency (i.e. is repeatable) only if the underlying state is also repeatable (Wolf & Weissing 2010). Therefore, quantifying and comparing repeatability estimates for both potential state variables and behavioural traits can provide valuable information about their relationships and also can be used to identify suitable state variables controlling the 'machinery' that shapes animal personality (Sih *et al.* 2015).

Two physiological traits that have been proposed to be key candidate state variables are hormone levels (Cockrem 2007; Williams 2008; Hau & Goymann 2015) and metabolic rates (Careau *et al.* 2008; Biro & Stamps 2010; Mathot & Dingemanse 2015). Accordingly, if differences in hormone levels and metabolic rates regulate consistent individual differences in behaviour, these two state variables should exhibit similar or higher temporal consistency (i.e. repeatability) compared to behavioural traits (Sih *et al.* 2015).

Although previous meta-analyses have quantified the repeatability of behavioural traits (Bell, Hankison & Laskowski 2009) and metabolic rates (Nespolo & Franco 2007; White, Schimpf & Cassey 2013; Auer *et al.* 2016), no meta-analysis has yet assessed the repeatability of hormone levels. Moreover, no meta-analysis has attempted to find out which of the two proposed physiological traits is a better candidate state variable. Thus, we performed systematic reviews and multilevel meta-analyses to synthesize the current literature and to quantitatively assess the repeatability of interindividual differences in hormone levels, metabolic rates and behavioural traits. We restricted our analyses to studies on avian species for two reasons. First, birds represent one of the best-studied taxa in animal personality research (Bell, Hankison & Laskowski 2009) and a substantial number of bird species have also been studied for both hormone levels (Cockrem *et al.* 2009) and metabolic rates (Careau *et al.* 2008). Secondly, the well-resolved phylogeny within the class *Aves* (e.g. Jetz *et al.* 2012) allows us to control for effects of phylogenetic history among species. Despite the potential importance of phylogeny (Chamberlain *et al.* 2012), earlier meta-analyses have not taken it into consideration (e.g. Nespolo & Franco 2007; Bell, Hankison & Laskowski 2009; White, Schimpf & Cassey 2013; Auer *et al.* 2016).

Under the assumption that hormone levels and metabolic rates are the underlying basis of state-dependent behaviour, we expect that these two physiological traits will exhibit similar or higher repeatability estimates than behavioural traits. In addition, we conduct meta-regression analyses to examine possible sources of variation in repeatability estimates of the three investigated traits. In a similar manner to two previous meta-analyses, one on metabolism (Nespolo & Franco 2007) and one on behaviour (Bell, Hankison & Laskowski 2009), we focus on the following five questions: (i) Does repeatability decrease as the interval between measurements increases? (ii) Does repeatability differ between males and females? (iii) Does repeatability of wild individuals deviate from repeatability of captive individuals? (iv) Does repeatability vary between breeding and non-breeding season? (v) Does repeatability differ among different types of hormones, metabolic rates or behaviours? With regard to hormone levels, we also address the question whether repeatability estimates are different between baseline and stress-induced (e.g. after being constrained in a bag) measurements.

Materials and methods

ARTICLE SEARCH AND DATA COLLECTION

We collected articles separately for each of the three data sets (hormones, metabolism and behaviour) used in our meta-analyses. We followed different search strategies when compiling data sets for the meta-analysis on repeatability of hormone levels and the two updates of meta-analyses on repeatability of metabolic rates (Nespolo & Franco 2007; White, Schimpf & Cassey 2013; see also Auer *et al.* 2016) and behavioural traits (Bell, Hankison & Laskowski 2009). The details of these search strategies are presented in Appendix S1, Supporting Information along with three flow diagrams (Figs S1–S3 in Appendix S4; often referred to as PRISMA flow charts – the Preferred Reporting Items in Systematic Reviews and Meta-Analyses; Moher *et al.* 2009; Nakagawa & Poulin 2012), corresponding to the three different meta-analytic data sets.

INCLUSION AND EXCLUSION CRITERIA

To obtain some degree of uniformity between the three data sets, we applied the following general inclusion criteria. First, studies had to report repeatability estimates in the form of intraclass correlation coefficients (ICC) or Spearman/Pearson correlation coefficients (r) of individual hormone levels, metabolic rates or behavioural traits. For the new data set on repeatability of hormone levels, we also contact authors of studies that repeatedly measured individual hormone levels, but did not report correlation coefficients (r or ICC), asking for repeatability estimates or raw data. Second, studies were conducted on laboratory or wild birds (we excluded studies that used individuals from domesticated populations or birds from zoological gardens). Third, we only included studies that did not use lines selected for specific hormone or behavioural traits (e.g. low stress response or explorative behaviour). Fourth, when an experiment or manipulation was carried out (e.g. increase in brood size or injection of any substances), we included only control groups, which did not undergo any treatment. Fifth, we restricted all data sets to measurements on adult birds only. We used this criterion because hormone levels are likely to fluctuate during early developmental stages as the hypothalamic–pituitary–adrenal axis and hypothalamic–pituitary–gonadal axis of chicks and juveniles are still under development (Ottinger, Wu & Pelican 2002; Rensel, Boughton & Schoech 2010; Schmidt *et al.* 2014). Thus, studies on chicks or juveniles would

have given incomparable hormone measurements with regard to the hormone levels of adults. For uniformity, we applied this criterion to all three data sets. Sixth, because we were interested if repeatability changes as the time between two measurements increases, studies had to report intervals between measurements or any other information from which the interval could be estimated (e.g. length of study). In addition to these general inclusion criteria, we applied data set-specific inclusion criteria, which are detailed in Appendix S1.

DATA EXTRACTION

Along with repeatability estimates (r or ICC) from each study, we extracted the type of hormones, metabolic rates or behavioural traits. We classified hormones, metabolic rates and behavioural traits into categories (for more details on the classification, see Appendix S1): three hormone categories (androgens, oestrogens and glucocorticoids); two metabolic rate categories (basal-metabolic rates, BMR and maximum metabolic rates, MMR); and 10 behavioural trait categories (activity, aggression, antipredator behaviour, exploration, foraging, habitat selection, migration, parental care, social interaction and other). Further, we collected information on the study species, sex of individuals measured, the number of individuals (n), and the number of observations per individual (k). For studies that did not state k but reported the total number of observations, we calculated k by dividing the number of observations by the number of individuals. We also collected the mean interval (days) between two repeated trait measurements. When the interval was not explicitly reported, we estimated the interval using other reported time spans (e.g. 'between 15 March and 15 June 2013') or time points (e.g. 'in August 2009 and 2010'). We recorded study identities and group identities to account for correlated structures of repeatability estimates coming from the same study or the same group of birds. In addition, we excluded repeatability estimates from joined groups (e.g. one estimate for both sexes), when data for these groups were also reported individually (e.g. one estimate for males and one estimate for females). Another potential source of substructure among repeatability estimates may arise from phylogenetic relatedness (Nakagawa & Santos 2012). We therefore constructed phylogenetic trees for each data set to include in the analyses (for details, see below). Finally, we collected three more variables, which are also likely to affect estimated repeatabilities. (i) Publication year to test for so called time-lag bias (Nakagawa & Santos 2012). Time-lag bias is a type of publication bias and occurs when initially published studies have larger effect sizes then follow-up studies (Trikalinos & Ioannidis 2005), or when studies with larger effect sizes get published faster (Jennions & Møller 2002). (ii) We noted whether the study was performed on captive or wild birds, and (iii) at which time of the reproductive cycle (i.e. breeding or non-breeding season) the studies were conducted. For the hormonal data set, we additionally extracted whether hormone levels were baseline or stress-induced measurements, and whether repeatabilities were time adjusted or not. Similarly, for the metabolic rates, we extracted whether repeatabilities were based on whole-animal or mass-adjusted metabolic rates.

STATISTICAL ANALYSES

Repeatability and effect size standardization

We ran all statistical analyses in the R environment version 3.2.1 (R Core Team 2015). To calculate (unadjusted) repeatabilities (Nakagawa & Schielzeth 2010) for the hormonal data that we received from the contacted authors, and for our own behavioural data, we employed linear mixed-effects models (LMMs) using the *lmer* function from the LME4 package (Bates, Maechler & Bolker 2015). Where necessary, response variables (i.e. hormone levels or behavioural measurements) were (natural) log or square root transformed prior to analysis to meet the assumption of normality. Repeatabilities were calculated as ICC by dividing the between-individual variance by the total sampling variance (Nakagawa & Schielzeth 2010). In addition, we calculated repeatabilities using ANOVA according to Lessells & Boag (1987) for LMM-based repeatabilities that were close to zero (<0·005). We used the latter approach, because repeatability estimates obtained from LMMs are constrained to be ≥0, and the ANOVA-based approach allows the estimates to be negative (Nakagawa & Schielzeth 2010). Therefore, calculating negative ANOVA-based repeatabilities, for otherwise constrained data points, will better comply with the normality assumptions in meta-analytic models. All calculated repeatability estimates were then added to the hormone or behaviour literature-extracted data sets.

For the meta-analyses and meta-regressions, we standardized all repeatability estimates (r and ICC) using Fisher's Z transformation (Table 1, eqns 1–4; cf. McGraw & Wong 1996). Along with the standardized effect size Fisher's Z (Z_r and Z_{ICC}; referred to as Z_{ICC} hereafter), we calculated the corresponding sampling variances (VarZ_r and VarZ_{ICC}; referred to as VarZ_{ICC} hereafter). For all graphs, we back-transformed model parameters to their original scale (i.e. ICC; Table 1, eqn 6) to make effect sizes comparable with previous meta-analyses. For the back-

Table 1. Fisher's Z transformation (Z_r) and its back-transformation for correlation-based repeatabilities (r; also referred to as interclass correlation), and intraclass correlation coefficients (ICC). n is the number of individuals, and k is the (average) number of repeated measures per individuals

Repeatability	Fisher's r to Z transformation		Sampling variance	
r	$Z_r = 0{\cdot}5 \ln \frac{1+r}{1-r}$	eqn 1	$\mathrm{Var}Z_r = \frac{1}{n-3}$	eqn 2
ICC	$Z_{ICC} = 0{\cdot}5 \ln \frac{1+(k-1)\mathrm{ICC}}{1-\mathrm{ICC}}$	eqn 3	$\mathrm{Var}Z_{ICC} = \frac{k}{2(n-2)(k-1)}$	eqn 4
Standardized effect size	**Back-transformation Z to r**			
Z_r	$r = \frac{\exp(2Z_r)-1}{\exp(2Z_r)+1}$	eqn 5		
Z_{ICC}	$\mathrm{ICC} = \frac{\exp(2Z_{ICC})-1}{\exp(2Z_{ICC})+\bar{k}-1}$	eqn 6		

transformation, we used ICC rather than r because ICC was the most commonly reported repeatability (≥90%) in all three data sets.

Meta-analyses and meta-regressions

We conducted multilevel meta-analyses for each of the three data sets (repeatabilities of hormone levels, metabolic rates and behavioural traits) separately. We fitted LMMs using the MCMCGLMM package in R (Hadfield 2010; Hadfield & Nakagawa 2010). In all models, we fitted the effect size Z_{ICC} as the response variable together with its relative sampling error variance $\mathrm{Var}Z_{ICC}$. As random effects, we included effect size identities, species and group identities. We used group identities instead of study identities, because many studies reported repeatability estimates for more than one group (e.g. females and males). Given that studies also reported multiple measurements that originated from groups sharing the same or a similar set of birds (e.g. short- and long-term repeatabilities), we also fitted a variance–covariance matrix in all models to account for correlation arising from these shared groups. We assumed that the correlations among shared groups were 0·5 (see sensitivity analysis for a more conservative approach; for a similar approach, see Booksmythe *et al.* 2015).

For each data set, we ran four different models. First, we fitted a standard meta-analytic model (intercept-only model or meta-analysis) to estimate the overall effect size mean. Second, we repeated the intercept-only model, but controlled for phylogenetic effects fitted as a covariance matrix (phylogenetic meta-analysis). We then performed a meta-regression (full model) using the intercept-only model with added moderators (i.e. fixed effects). Finally, we extended the meta-regression by adding phylogenetic information (phylogenetic meta-regression).

The choice of moderators for the two meta-regressions was based on two criteria. First, we included fixed effects that are relevant to our questions stated in the introduction (i.e. interval, season, captive or wild, sex, and for hormones, baseline or stress-induced measurements). Second, we ran univariate (multilevel) meta-regression models using the METAFOR package in R (Viechtbauer 2012) to independently test the significance of additional potential moderators (see Data Extraction above for a full list). The moderators that influenced repeatability estimates in the univariate models (i.e. statistically significant as 95% confidence/credible intervals (CI) did not span across zero) were then added to the full model or examined separately (e.g. types of hormones, metabolic rates and behavioural traits). Prior to analyses, we log-transformed the values of measurement interval and publication year. Also, for all data sets, we standardized continuous moderators resulting in a mean of 0 and standard deviation of 1 (i.e. z-transformation).

We fitted phylogenetic meta-analytic models using two different phylogenetic trees from Jetz *et al.* (2012). Phylogenetic trees were prepared on the basis of Ericson backbone (Ericson tree; Ericson *et al.* 2006) and Hackett backbone (Hackett tree; Hackett *et al.* 2008). The topology of both trees for each data set is shown in Figs S4–S6. In an exploratory analysis, we ran the two phylogenetic models independently with each tree, and the model results turned out to be similar (quantitatively and qualitatively). Therefore, we only present results for the Hackett tree.

Details on model and prior specifications can be found in Appendix S1. In brief, we conducted three parallel runs for each model. We used Gelman–Rubin diagnostics and the *autocor* function in the MCMCGLMM package to check the three obtained MCMC chains for convergence and mixing (Gelman & Rubin 1992). We extracted model results from the chain with the lowest deviance information criteria (DIC; Spiegelhalter *et al.* 2002); for that chain, we report mean posterior estimate and the 95% CI. Following Cohen (1988), we considered mean posterior estimates of 0·1, 0·3 and 0·5 as weak, moderate and strong effects. Overall effects, fixed effects or differences between levels of fixed effects were considered to be statistically significant when their CIs did not include zero.

We assessed consistency among effect sizes for the intercept-only models by calculating heterogeneities (I^2). Along with the overall heterogeneity ($I^2_{[total]}$, between-study variance divided by the total variance; Higgins & Thompson 2002), we quantified I^2 values for each level of the random factors following Nakagawa & Santos (2012). As specified by Higgins *et al.* (2003), we defined I^2 values of 25%, 50% and 75% to be low, moderate and high levels of heterogeneity, respectively.

ADDITIONAL AND SENSITIVITY ANALYSES

In addition to the analyses above, we conducted three separate meta-regressions to examine differences between the types of hormones, metabolic rates and behavioural traits. Along with the random effects used for the meta-analytic models, we fitted 'types' as fixed effects (without intercept). Further, we included fixed effects that turned out to be statistically significant in the meta-regression models. We performed these additional analyses to examine our implicit assumption (for our main meta-analyses) that within each data set different types of a given trait (e.g. for metabolic rates: BMR and MMR) do not dramatically differ between each other in terms of repeatability. To test the robustness of our main results, we also conducted three statistical analyses using different subsets of the three main data sets (i.e. sensitivity analyses); the details of these sensitivity analyses can be found in Appendix S1.

PUBLICATION BIAS

Studies with positive or statistically significant results are more likely to be considered for publication than studies with negative or non-significant results (so called file drawer problem; Rosenthal 1979). Such publication bias may lead to incorrect conclusions from a meta-analysis. Accordingly, we tested the three data sets for publication bias, visually by using funnel plots, and statistically by conducting Egger's regression (Egger *et al.* 1997), as well as via trim-and-fill analyses (Duval 2005). For the former, we plotted meta-analytic residuals against their corresponding precision values (the inverse of the square root of the sampling variance) and checked the resulting plots for funnel asymmetry. For Egger's regression, meta-analytic residuals (the sum of within-study effects and sampling error effects) were regressed against sampling errors following Nakagawa & Santos (2012); evidence for publication bias can be inferred when the regression intercept is significantly different from zero (Egger *et al.* 1997). Finally, we ran trim-and-fill analyses on meta-analytic residuals (Nakagawa & Santos 2012) employing the *trimfill* function in METAFOR (Viechtbauer 2012).

Results

SUMMARY OF THE DATA SETS

The outcomes of the comprehensive literature searches are summarized in Figs S1–S3. Moreover, sample sizes at different levels of the three final data sets are summarized in Table 2. The data set on behavioural traits was the largest of all three data sets, followed by hormones, and the smallest data set on metabolic rates ($N_{[effect\ size]} = 477$, $N_{[effect\ size]} = 145$ and $N_{[effect\ size]} = 51$, respectively; Table 2). The number of individuals per effect size

Table 2. Summary of the three data sets included in our analyses. $n_{[individuals]}$ represents the number of individuals per group, and $k_{[repeats]}$ represents the number of repeated measures per individual. Interval is the time in days between two repeated measurements. An interval of zero was assigned when repeated measures were taken within the same observation (i.e. aggression against a decoy, of which each attack was recorded as an individual measurement; Garamszegi *et al.* 2006)

Data	$N_{[studies]}$	$N_{[effect\ sizes]}$	$N_{[species]}$	$N_{[groups]}$	$n_{[individuals]}$		$k_{[repeats]}$		Interval		log Interval	
					Mean (SD)	Median (range)	Mean (SD)	Median (range)	Mean (SD)	Median (range)	Mean (SD)	Median (range)
Hormones	47	145	34	60	31·9 (29·0)	25·0 (6–202)	4·2 (4·5)	2·4 (1·5–24·6)	274·7 (575·9)	50·5 (0·04–2555)	4·0 (2·1)	3·9 (−3·2 to 7·8)
Metabolism	17	51	16	21	36·6 (32·1)	25·0 (6–142)	2·8 (2·0)	2·4 (1·3–12·0)	175·1 (267·5)	30·3 (0·08–1095)	4·0 (1·7)	3·4 (−2·5 to 7·0)
Behaviour	115	477	75	168	53·6 (107·4)	23·0 (3–2343)	3·0 (3·3)	2·2 (1·1–47·3)	154·2 (163·2)	60 (0·0–547·5)	3·9 (1·9)	4·1 (−0·7 to 6·3)

($n_{[individuals]}$) showed the greatest variation for behavioural traits (range: 3–1243 individuals; Table 2). Individual measurements of hormone levels were on average more often repeated than individual measurements of behavioural traits and metabolic rates ($k_{[repeats]}$ = 4·2, $k_{[repeats]}$ = 3·0, $k_{[repeats]}$ = 2·8, respectively; Table 2). Also, in the hormone data set, one study (Small & Schoech 2015) reported by far the largest of all intervals between two measurements – 2555 days (i.e. 6–8 years; Table 2; see Appendix S1 for a sensitivity analysis excluding this interval). When log-transformed, however, intervals did cover similar ranges for all three data sets (Table 2).

OVERALL REPEATABILITY ESTIMATES AND HETEROGENEITIES

Our main meta-analyses (intercept-only models) revealed moderate to high mean repeatability estimates for metabolic rates (meta-analysis: ICC = 0·451, 95% CI = 0·269–0·611; phylogenetic meta-analysis: ICC = 0·545, 95% CI = 0·218–0·842; Fig. 1) and behavioural traits (meta-analysis: ICC = 0·410, 95% CI = 0·348–0·463; phylogenetic meta-analysis: ICC = 0·422, 95% CI = 0·304–0·547; Fig. 1), but a small mean repeatability estimate for hormones (meta-analysis: ICC = 0·150, 95% CI = 0·098–0·205; phylogenetic meta-analysis: ICC = 0·164, 95% CI = 0·088–0·273; Fig. 1; see also Tables S4–S6 in Appendix S3). We detected high total heterogeneities for all meta-analytic models (with and without phylogeny $I^2_{[total]}$ > 69%; Table 3), which justified our meta-regression approach. For both hormone and behaviour data sets, the largest proportion of variance was associated with the residuals, suggesting high variability at the level of

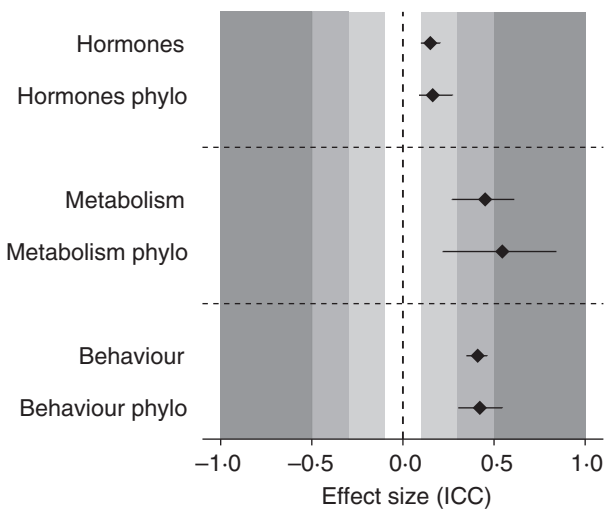

Fig. 1. Meta-analytic mean estimates of repeatability for hormone levels, metabolic rates and behavioural traits. We present posterior means and 95% credible intervals (CIs) of both meta-analyses and phylogenetic meta-analyses (phylo) obtained from linear mixed-effects models (LMMs). All estimates are back-transformed into intraclass correlation coefficients (ICCs). White and light grey, medium grey, and dark grey indicate small, medium and large effect sizes, following Cohen (1988).

Table 3. Deviance information criterion (DIC) and heterogeneities (I^2%) for meta-analytic models (without and with phylogeny) for each of the three investigated traits; hormone levels, metabolic rates and behavioural traits

Model	DIC	$I^2_{[group]}$	$I^2_{[species]}$	$I^2_{[phylo]}$	$I^2_{[residuals]}$	$I^2_{[total]}$
Hormones						
Meta-analysis	19·12	13·88	11·57	–	44·11	69·56
Meta-analysis phylo	19·09	10·67	7·19	14·05	40·25	72·17
Metabolism						
Meta-analysis	12·61	16·65	47·63	–	22·53	86·81
Meta-analysis phylo	12·48	9·59	25·23	36·96	18·05	89·84
Behaviour						
Meta-analysis	446·86	2·44	31·59	–	56·76	90·79
Meta-analysis phylo	446·48	2·21	26·55	10·5	52·49	91·75

effect sizes. With regard to the metabolism data set, species and phylogeny accounted for most of the variation (Table 3). For all data sets, however, the results of non-phylogenetic and phylogenetic models were similar (Fig. 1; Table 3), and corresponding models provided very similar DIC values (Table 3), suggesting both models are equally supported. Therefore, below, we only provide results from the non-phylogenetic models (meta-regression). Model results from the corresponding phylogenetic meta-analyses are reported in Fig. S8.

META-REGRESSION: EFFECTS OF MAIN MODERATORS

In our main meta-regression analyses, we investigated potential effects of five moderators: interval, sex, whether captive or wild individuals were used, season, differences among different types of hormones, metabolic rates or behavioural traits, and for hormones only whether hormones levels were baseline or stress-induced. For all data sets, repeatabilities decreased slightly with increasing interval between two measurements, although this effect was only statistically significant for hormones (ICC: $ß_{[interval]} = -0{\cdot}043$, 95% CI = −0·082 to −0·003) and behaviour (ICC: $ß_{[interval]} = -0{\cdot}055$, 95% CI = −0·084 to −0·025; Fig. 2; see also Fig. S7).

Additionally, the meta-regressions revealed two other notable effects. The first is a small, albeit statistically significant, difference between sexes in repeatability of behavioural traits (ICC: $ß_{[females - males\ diff.]} = 0{\cdot}129$, 95% CI = 0·009–0·266; Fig. 2c; see also Fig. S7c). Yet, this sex effect seemed to be due to the fact that one sex was mainly represented in some behavioural traits, which showed different magnitudes of repeatability. Specifically, for social interactions (e.g. mate preference), which were characterized by low repeatability compared to other behavioural traits (Fig. 3), females accounted for 74·29% of the effect sizes (males: 22·85%; both sexes: 2·86%). Moreover, for the relatively highly repeatable aggression-related behaviours, only 15·63% of the effect sizes came from females (males: 59·37%; both sexes: 25·00%). The second notable effect is a significant difference between repeatabilities of baseline and stress-induced hormone levels, with stress-induced levels being more repeatable than baseline levels (ICC: $ß_{[baseline - stress\ diff.]} = 0{\cdot}193$, 95% CI = 0·076–0·305; Fig. 2a; see also Fig. S7a).

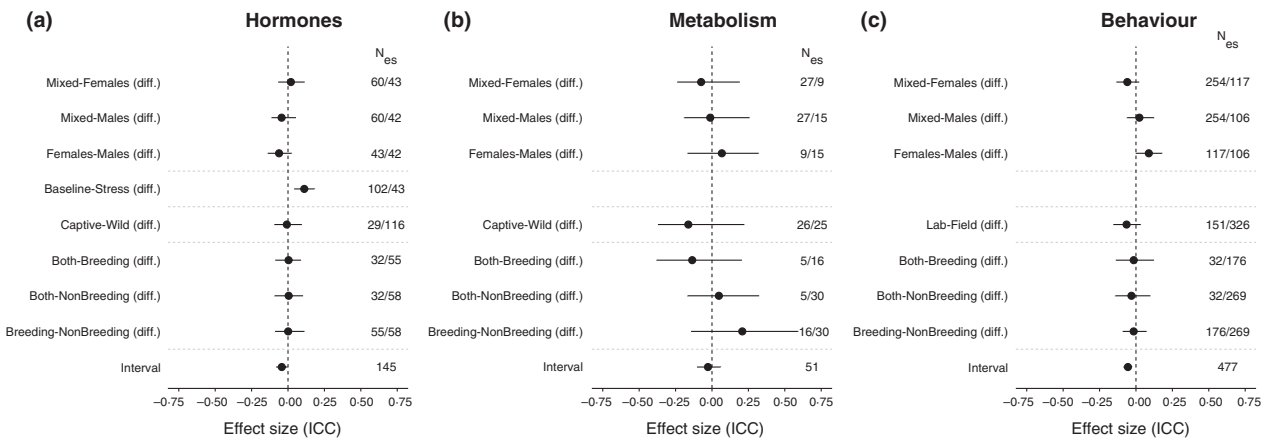

Fig. 2. Results of meta-regressions (full models) for (a) hormone levels, (b) metabolic rates and (c) behavioural traits. Meta-regressions were run assuming the correlation among shared groups to be 0·5. We present posterior means and 95% credible intervals (CIs) from linear mixed-effects models (LMMs). All estimates are back-transformed into intraclass correlation coefficients (ICCs). The number of effect sizes (N_{es}) is listed on the right side of each panel. Differences (diff.) between moderator levels can be inferred as statistically significant when 95% CIs do not overlap zero. Negative values for the continuous moderator 'Interval' indicate a decrease of repeatability with increasing periods of time between repeated measurements.

None of the other moderators showed significant influences on repeatability, either in our univariate models (including time adjusted hormone levels as well as mass-adjusted metabolic rates) or in our main meta-regressions (Fig. 2). Reflecting this result, we observed relatively small marginal R^2 values (variance explained by the fixed effects/moderators*; sensu* Nakagawa & Schielzeth 2013) in all the full meta-regression models ($R^2_{[hormones]} = 0{\cdot}196$; $R^2_{[metabolism]} = 0{\cdot}244$; $R^2_{[behaviour]} = 0{\cdot}053$).

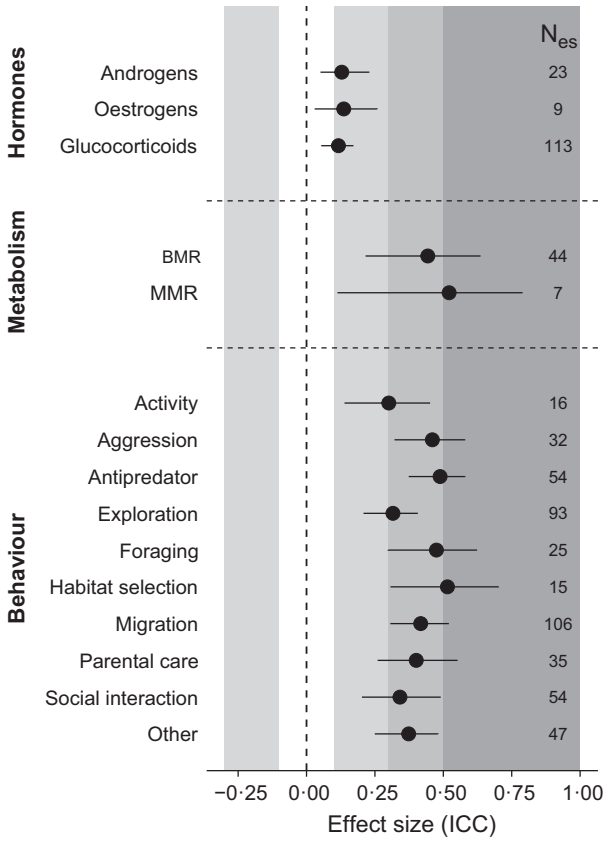

Fig. 3. Meta-analytic mean estimates of repeatability for different types of hormone levels (top), metabolic rates (middle) and behavioural traits (bottom). We present posterior means and 95% credible intervals (CIs) from linear mixed-effects models (LMMs). All estimates are back-transformed into intraclass correlation coefficients (ICCs). The number of effect sizes (N_{es}) is listed on the right side of each panel. White and light grey, medium grey and dark grey indicate small, medium and large effect sizes, following Cohen (1988).

ADDITIONAL ANALYSES AND PUBLICATION BIAS

We ran three additional meta-regression analyses to estimate differences between the types of hormones, metabolic rates and behavioural traits within each data set. This is because our meta-analyses were conducted under the assumption that these data set specific trait types show similar magnitudes in their effect sizes. For the hormone and metabolism data set, the results support this assumption (Fig. 3). In comparison, mean effect sizes for different behavioural traits varied to a greater extent (Fig. 3). However, many overlaps among the 95% CIs suggest that our assumption is still reasonable. As the results of the three sensitivity analyses did not alter the conclusions of our main models (i.e. the main model results are robust), we only present the results of the main models here. Results of the sensitivity analyses can be found in Figs S8–S10 (see also Tables S7–S10).

Finally, we did not identify any distinct asymmetry in the funnel plots by visual inspection (Fig. 4). This observation was confirmed by the results of Egger's regressions and trim-and-fill analyses (Table 4). Consequently, we found little evidence for publication bias. Thus, it is likely that our results are unbiased and potentially robust.

Discussion

In this study, we meta-analytically quantified overall repeatabilities of hormone levels, metabolic rates and behavioural traits to test which of the two proposed physiological state variables is a more likely candidate for generating consistent individual differences in behaviour (i.e. animal personality). Our meta-analyses have revealed overall moderate to high levels of repeatability estimates for

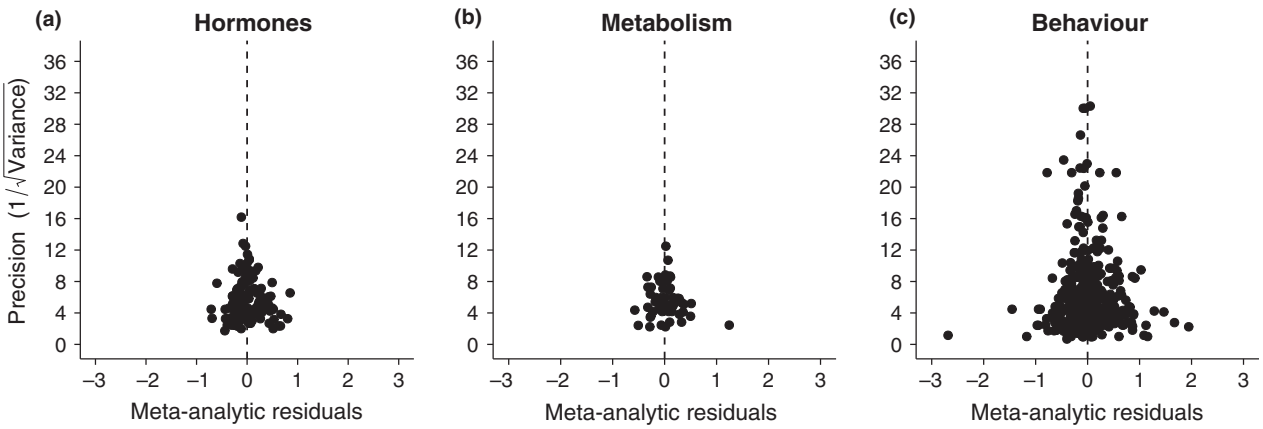

Fig. 4. Funnel plots of meta-analytic residuals plotted against their precision for (a) hormone levels, (b) metabolic rates and (c) behavioural traits. Dashed lines indicate zero.

Table 4. Results of Egger's regression and trim-and-fill analysis conducted on meta-analytic residuals for each of the three investigated traits (data sets); hormone levels, metabolic rates and behavioural traits

	Egger's regression			Trim-and-fill		
Model	*t*-Value	d.f.	*P*-value	Missing effect sizes	SE	*P*-value
Hormones						
Meta-analysis	1·301	143	0·195	1	2·0	0·25
Meta-analysis phylo	1·327	143	0·187	1	2·0	0·25
Metabolism						
Meta-analysis	0·895	49	0·375	0	1·4	0·50
Meta-analysis phylo	0·872	49	0·387	0	1·4	0·50
Behaviour						
Meta-analysis	1·310	475	0·191	0	1·4	0·50
Meta-analysis phylo	1·385	475	0·167	0	1·4	0·50

metabolism and behaviour (ICC = 0·459 and ICC = 0·413, respectively; Fig. 1), which are of similar magnitudes for both traits. In contrast, the mean repeatability estimate of hormones (ICC = 0·145; Fig. 1) was approximately three times lower, suggesting that differences in metabolic rates, rather than hormone levels, are likely to be one of the key drivers responsible for consistent between-individual variation in behaviour (Careau *et al.* 2008; Biro & Stamps 2010; Sih *et al.* 2015; but see below for other possible explanations of this relationship).

In addition, we addressed six questions (see Introduction) to determine moderator variables that could explain the heterogeneity observed in the data. However, only two of the hypothesized moderators accounted for significant amounts of the variation (Fig. 2). First, repeatability estimates declined with increasing time (intervals) between measurements of hormone levels and behavioural traits (but not of metabolism; Fig. 2). Second, repeatability estimates of stress-induced hormone levels were higher than repeatabilities of baseline hormone levels (Fig. 2). Below, we discuss the significance of our findings within the framework of state-dependent personality models (Dingemanse & Wolf 2010; Luttbeg & Sih 2010; Wolf & Weissing 2010; Sih *et al.* 2015).

METABOLISM AND BETWEEN-INDIVIDUAL VARIATION IN BEHAVIOUR

Theoretical work suggests that interindividual variation in metabolic rates should be repeatable and thus represents one of the main candidates for shaping consistent among-individual differences in behaviour (Careau *et al.* 2008; Biro & Stamps 2010; Mathot & Dingemanse 2015). Indeed, several empirical studies have shown that metabolic rates are repeatable over both short and long sampling periods (reviewed in Nespolo & Franco 2007; Versteegh *et al.* 2008; Biro & Stamps 2010). Our overall repeatability estimate of ICC = 0·451 for metabolic rates supports the general pattern of consistent metabolic rates, but was slightly lower than the repeatability estimate from an earlier meta-analysis (ICC = 0·57; Nespolo & Franco 2007; see also White, Schimpf & Cassey 2013; Auer *et al.* 2016). The discrepancy of repeatabilities between Nespolo & Franco's meta-analysis and our study is likely due to differences in study design. While Nespolo & Franco included repeatability estimates of whole-animal metabolic rates only, we included both whole-animal and mass-controlled estimates. We emphasize, however, that for our data set, repeatabilities of whole-animal and mass-controlled estimates did not significantly differ from each other (see Results).

Since the functional linkage between an individual's energetic state and its personality traits has been postulated (e.g. Careau *et al.* 2008; Biro & Stamps 2010; Wolf & Weissing 2010), several studies have investigated the relationship between metabolic rates and behavioural traits (reviewed in Biro & Stamps 2010; Careau & Garland 2012). In their review, Biro & Stamps (2010) implied that metabolic rates are predominantly associated with behaviours that likely influence an individual's energy budget (e.g. dominance, aggression, activity, boldness). In accordance with this assumption, our meta-analyses show that repeatability estimates of metabolic rates are most similar to repeatability estimates of behavioural traits that are likely associated with energy acquisition or expenditure (e.g. aggression, foraging and habitat selection; Fig. 3; but see activity; Biro & Stamps 2010; Mathot & Dingemanse 2015). Our results therefore have a bearing on the theory that the metabolic machinery is involved in generating between-individual differences of behavioural traits (e.g. Biro & Stamps 2010; Dingemanse & Wolf 2010; Sih *et al.* 2015).

HORMONE–BEHAVIOUR RELATIONSHIP AND WITHIN-INDIVIDUAL VARIANCE

Although endocrine mechanisms have been recognized as an important regulator of behaviour (Cockrem 2007; Koolhaas *et al.* 2010; Hau & Goymann 2015), surprisingly little is known about the consistency of circulating hormone concentrations and their association with variation in personality traits (Kempenaers, Peters & Foerster 2008; Lessells 2008; Williams 2008). In comparison with metabolic rates, our meta-analysis showed an overall small repeatability estimate for hormone levels (ICC = 0·145; Fig. 1). This lower repeatability estimate indicates that hormone levels are likely characterized by high within-individual variation relative to between-individual variation (Bell, Hankison & Laskowski 2009; Nakagawa & Schielzeth 2010). One widely accepted explanation for high within-individual variation in hormone concentrations is that the endocrine system acts as a mediator for short-term adjustment (i.e. daily or seasonal) and for rapid responses

to changing environmental and social conditions (Oliveira 2004; Kempenaers, Peters & Foerster 2008; Follett 2015). In accordance with this explanation, Laucht *et al.* (2011) showed that testosterone levels varied greatly within-individual male house sparrows (*Passer domesticus*) between the day and night time (see also Ouyang, Hau & Bonier 2011 for seasonal variation). Importantly, the authors also found that individuals differed in their amount of within-individual variation (i.e. predictability; Stamps, Briffa & Biro 2012; Cleasby, Nakagawa & Schielzeth 2015). Besides that hormone levels seem to be highly responsive to external factors, some part of the high within-individual variation is likely to be caused by measurement errors, which is less problematic in metabolic measurements (e.g. detection limits or assay variation; e.g. Ouyang, Hau & Bonier 2011; Hau *et al.* 2016).

Individual differences in within-individual variation are generally assumed to be the same for all individuals when repeatability is calculated (Stamps, Briffa & Biro 2012; Cleasby, Nakagawa & Schielzeth 2015; Westneat, Wright & Dingemanse 2015). However, incorporating these differences into animal personality research will be important to fully understand how environmental and social conditions affect the hormone–behaviour relationship.

EFFECTS OF MODERATORS ON REPEATABILITY

We showed that hormone levels and behavioural traits exhibited significantly lower repeatability estimates when the interval between repeated measures increased (Fig. 2a, d). These results are in agreement with findings of an earlier meta-analysis on the repeatability of behaviour (Bell, Hankison & Laskowski 2009). A similar effect of interval length has also been previously reported for the repeatability of metabolism (White, Schimpf & Cassey 2013; Auer *et al.* 2016). Although our results showed a similar direction for the effect of intervals for the metabolism data set, this effect was not statistically significant (Fig. 2b). This result might be explained by the lack of statistical power, as our analyses included only studies on birds, unlike the previous meta-analyses, which examined all species. Nevertheless, it is important to bear in mind that measurements repeatedly taken over a short interval are expected to produce higher repeatability estimates than measurements taken over long intervals (Bell, Hankison & Laskowski 2009; Boulton *et al.* 2014; Araya-Ajoy, Mathot & Dingemanse 2015). Particularly, high repeatability can occur when individuals are measured under different environmental conditions over a short period of time (e.g. low vs. high predation; Araya-Ajoy, Mathot & Dingemanse 2015; see also Westneat *et al.* 2011; Dingemanse & Dochtermann 2013 for the idea of pseudo-repeatability).

The second noteworthy finding is that baseline hormone levels showed a significantly lower repeatability than stress-induced hormone levels (Fig. 2). This result supports the theory of Hau & Goymann (2015) (see also Rensel & Schoech 2011), who argued that baseline levels have low repeatability because they are influenced by intrinsic as well as extrinsic factors, which an individual experienced prior to measurements (e.g. body condition or temperature). Stress-induced levels, on the other hand, are more repeatable because they represent the hormonal response to the same standardized stressor for all individuals.

Conclusions and future directions

Our meta-analyses on the repeatabilities of hormone levels, metabolic rates and behavioural traits provide evidence that between-individual differences in metabolism, rather than in hormone levels, are a likely underlying mechanism

Table 5. State-dependent personality models and their predictions with regard to repeatability and relationships between hormone levels, metabolic rates and consistent individual differences in behaviours (i.e. animal personality)

Mechanism	Description	Assumptions	References
State-dependent behaviour without feedback	Hormone levels or metabolic rates are the driver of behavioural traits	Physiological traits that mediate consistent individual differences in behaviour exhibit similar or higher repeatability than state-dependent behaviour	Dall, Houston & McNamara (2004), Dingemanse & Wolf (2010), Wolf & Weissing (2010)
State-dependent behaviour with feedback	Positive feedback mechanisms create a stabilizing effect between initial differences in hormone levels or metabolic rates and behavioural traits	Physiological and behavioural traits that are coupled due to stabilizing feedback loops display similar repeatability estimates, whereas behavioural traits and physiological mechanism that do not stabilize each other differ in their degrees of repeatability	Dall, Houston & McNamara (2004), Luttbeg & Sih (2010), Dingemanse & Wolf (2010), Wolf & Weissing (2010), Sih *et al.* (2015)
Co-evolution	Correlated differences of hormone levels or metabolic rates and behavioural traits co-evolve and selection acts on one or both of the correlated traits	Physiological and behavioural traits that co-evolved together have similar repeatability estimates, whereas traits that did not co-evolve together display different degrees of repeatability	Réale *et al.* (2010b), Wolf & McNamara (2012)

mediating consistent between-individual differences in behaviour. Nevertheless, caution should be exercised, since the found pattern of high repeatabilities in metabolic rates and behavioural traits could also have emerged if consistent behaviour promotes consistent metabolism, or these two traits feed back to each other (Careau *et al*. 2008; Biro & Stamps 2010). Indeed, high repeatabilities in both traits may be found even in the absence of any functional linkage. Thus, future investigations should first test whether there is a clear correlation between metabolic rates and personality traits. Subsequently, future studies should try to distinguish which of the proposed mechanism of state-dependent personality models (see Table 5) is the most likely candidate. For example, once a study has found a correlation between state variable (i.e. metabolism) and behavioural trait, quantitative genetic approaches could be used to determine the strength of genetic correlation between state (i.e. metabolism) and behavioural trait (Dingemanse & Wolf 2010; Careau *et al*. 2011). If no or little genetic correlation was found, a next step would be to determine whether or not state–behaviour feedback exists (Table 5; reviewed in Sih *et al*. 2015; see also Dingemanse & Wolf 2010). To test this, one could examine how changes in measurement intervals affect the repeatabilities of state variable and state-dependent behaviour. On the one hand, if feedback between state variable and behaviour exists, one would expect that changes in measurement interval would have comparable effects on the repeatability of both physiological state and behavioural trait. On the other hand, changes in interval length between measurements should have a larger effect on the repeatability of state-dependent behaviour than on the repeatability of the underlying state when feedback is absent.

Another aspect future studies should consider is that repeatability is a measure that is highly influenced by within-individual variance. Consequently, comparing repeatabilities across studies or populations may lead to incorrect conclusions with regard to consistent 'between-individual differences' (i.e. animal personality) when factors that influence within-individual variation differ dramatically. In the field of evolutionary quantitative genetics, it has been argued whether the measure of heritability (h^2, defined as the additive genetic variance divided by the total phenotypic variance) is a representative measure to compare the degree of 'evolutionary potential' or 'evolvability' between studies (Houle 1992; Hansen, Pélabon & Houle 2011; Garcia-Gonzalez *et al*. 2012). Houle (1992) proposed using the coefficient of additive genetic variation (CV_A), which is independent of other variance components (Garcia-Gonzalez *et al*. 2012). In line with the use of CV_A, we could utilize the coefficient of variation for between-individual variance (CV_B) in the field of animal personality. CV_B is an index, which is not confounded by within-individual effects and can easily be obtained as:

$$CV_B = \frac{\sqrt{V_B}}{\bar{x}}, \qquad \text{eqn 7}$$

where V_B is the between-individual variance and $\bar{x}$ the trait mean. [Correction added after online publication on 27 March 2017: equation 7 corrected]. Being standardized by its trait mean, CV_B allows us to directly compare the degree of among-individual variation between traits or populations (cf. Kruuk *et al*. 2000; Cleasby, Nakagawa & Schielzeth 2015; for examples of studies that used CV_B, see Cockrem 2007; Versteegh *et al*. 2008; van Dongen *et al*. 2010). CV_B is likely to represent a more suitable estimator for comparative studies, including meta-analyses (Nakagawa *et al*. 2015), because in a comparative context of animal personality within-individual variance can be considered as contextual noise. Thus, we encourage future studies to report CV_B together with repeatability estimates; these two dimensionless measures are likely to be useful under different questions and circumstances.

Acknowledgements

We are thankful to Michael Beaulieu, Eli Bridge, Nina Dehnhard, Pierre Deviche, Katharina Hirschenhauser, Jesse Krause, Meta Landys, Agnes Lewden, Katrin Ludynia, Melissa Mark, Amy Newman, Michael Romero, Scott Sakaluk, Baptiste Schmid, Alvin Setiawan, Roberto Victor Lacava e Silva, Douglas Tempel, Travis Wilcoxen and John Wingfield for sending us unpublished data or for providing additional details on their work. We thank Bruce Robertson, Sheri Johnson and two anonymous reviewers for valuable feedback and comments on previous versions of this manuscript. B.H. received a Doctoral Scholarship from the University of Otago. S.N. was supported by a Rutherford Discovery Fellowship (New Zealand) and a Future Fellowship (Australia) (FT130100268).

Data accessibility

All data (including data sets for hormone levels, metabolic rates and behavioural traits) used in this manuscript are present in the manuscript or can be found in its Supporting Information.

References

Araya-Ajoy, Y.G., Mathot, K.J. & Dingemanse, N.J. (2015) An approach to estimate short-term, long-term and reaction norm repeatability. *Methods in Ecology and Evolution*, **6**, 1462–1473.

Auer, S.K., Bassar, R.D., Salin, K. & Metcalfe, N.B. (2016) Repeatability of metabolic rate is lower for animals living under field versus laboratory conditions. *Journal of Experimental Biology*, **219**, 631–634.

Bates, D., Maechler, M. & Bolker, B. (2015) Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, **67**, 1–48.

Bell, A.M., Hankison, S.J. & Laskowski, K.L. (2009) The repeatability of behaviour: a meta-analysis. *Animal Behaviour*, **77**, 771–783.

Biro, P.A. & Stamps, J.A. (2008) Are animal personality traits linked to life-history productivity? *Trends in Ecology & Evolution*, **23**, 361–368.

Biro, P.A. & Stamps, J.A. (2010) Do consistent individual differences in metabolic rate promote consistent individual differences in behavior? *Trends in Ecology & Evolution*, **25**, 653–659.

Booksmythe, I., Mautz, B., Davis, J., Nakagawa, S. & Jennions, M.D. (2015) Facultative adjustment of the offspring sex ratio and male attractiveness: a systematic review and meta-analysis. *Biological Reviews*, doi: 10.1111/brv.12220.

Boulton, K., Grimmer, A., Rosenthal, G., Walling, C. & Wilson, A. (2014) How stable are personalities? A multivariate view of behavioural variation over long and short timescales in the sheepshead swordtail, *Xiphophorus birchmanni*. *Behavioral Ecology and Sociobiology*, **68**, 791–803.

Careau, V. & Garland, T. (2012) Performance, personality, and energetics: correlation, causation, and mechanism. *Physiological and Biochemical Zoology: Ecological and Evolutionary Approaches*, **85**, 543–571.

Careau, V., Thomas, D., Humphries, M.M. & Réale, D. (2008) Energy metabolism and animal personality. *Oikos*, **117**, 641–653.

Careau, V., Thomas, D., Pelletier, F., Turki, L., Landry, F., Garant, D. & Réale, D. (2011) Genetic correlation between resting metabolic rate and exploratory behaviour in deer mice (*Peromyscus maniculatus*). *Journal of Evolutionary Biology*, **24**, 2153–2163.

Chamberlain, S.A., Hovick, S.M., Dibble, C.J. *et al.* (2012) Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. *Ecology Letters*, **15**, 627–636.

Cleasby, I.R., Nakagawa, S. & Schielzeth, H. (2015) Quantifying the predictability of behaviour: statistical approaches for the study of between-individual variation in the within-individual variance. *Methods in Ecology and Evolution*, **6**, 27–37.

Cockrem, J.F. (2007) Stress, corticosterone responses and avian personalities. *Journal of Ornithology*, **148**, 169–178.

Cockrem, J.F., Barrett, D.P., Candy, E.J. & Potter, M.A. (2009) Corticosterone responses in birds: individual variation and repeatability in Adelie penguins (*Pygoscelis adeliae*) and other species, and the use of power analysis to determine sample sizes. *General and Comparative Endocrinology*, **163**, 158–168.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Sciences*, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, New Jersey, NJ, USA.

Dall, S.R.X., Houston, A.I. & McNamara, J.M. (2004) The behavioural ecology of personality: consistent individual differences from an adaptive perspective. *Ecology Letters*, **7**, 734–739.

Dingemanse, N.J. & Dochtermann, N.A. (2013) Quantifying individual variation in behaviour: mixed-effect modelling approaches. *Journal of Animal Ecology*, **82**, 39–54.

Dingemanse, N.J. & Wolf, M. (2010) Recent models for adaptive personality differences: a review. *Philosophical Transactions of the Royal Society of London. Sires B, Biological Sciences*, **365**, 3947–3958.

van Dongen, W.F.D., Maldonado, K., Sabat, P. & Vásquez, R.A. (2010) Geographic variation in the repeatability of a personality trait. *Behavioral Ecology*, **21**, 1243–1250.

Duval, S. (2005) The trim and fill method. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (eds H.R. Rothstein, A.J. Sutton & M. Borenstein), pp. 127–144. Wiley, Chichester, UK.

Egger, M., Smith, G.D., Schneider, M. & Minder, C. (1997) Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, **315**, 629–634.

Ericson, P.G.P., Anderson, C.L., Britton, T. *et al.* (2006) Diversification of Neoaves: integration of molecular sequence data and fossils. *Biology Letters*, **2**, 543–547.

Follett, B.K. (2015) "Seasonal changes in the neuroendocrine system": some reflections. *Frontiers in Neuroendocrinology*, **37**, 3–12.

Garamszegi, L.Z., Rosivall, B., Hegyi, G., Szoellosi, E., Toeroek, J. & Eens, M. (2006) Determinants of male territorial behavior in a Hungarian flycatcher population: plumage traits of residents and challengers. *Behavioral Ecology and Sociobiology*, **60**, 663–671.

Garcia-Gonzalez, F., Simmons, L.W., Tomkins, J.L., Kotiaho, J.S. & Evans, J.P. (2012) Comparing evolvabilities: common errors surrounding the calculation and use of coefficients of additive genetic variation. *Evolution*, **66**, 2341–2349.

Gelman, A. & Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Gosling, S.D. (2001) From mice to men: what can we learn about personality from animal research? *Psychological Bulletin*, **127**, 45–86.

Hackett, S.J., Kimball, R.T., Reddy, S. *et al.* (2008) A phylogenomic study of birds reveals their evolutionary history. *Science*, **320**, 1763–1768.

Hadfield, J.D. (2010) MCMC methods for multi-response generalized linear mixed models: the *MCMCglmm R* package. *Journal of Statistical Software*, **33**, 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**, 494–508.

Hansen, T.F., Pélabon, C. & Houle, D. (2011) Heritability is not evolvability. *Evolutionary Biology*, **38**, 258–277.

Hau, M. & Goymann, W. (2015) Endocrine mechanisms, behavioral phenotypes and plasticity: known relationships and open questions. *Frontiers in Zoology*, **12**, S7.

Hau, M., Casagrande, S., Ouyang, J.Q. & Baugh, A.T. (2016) Glucocorticoid-mediated phenotypes in vertebrates: multilevel variation and evolution. *Advances in the Study of Behavior*, **48**, 41–115.

Higgins, J.P.T. & Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, **21**, 1539–1558.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J. & Altman, D.G. (2003) Measuring inconsistency in meta-analyses. *British Medical Journal*, **327**, 557–560.

Houle, D. (1992) Comparing evolvability and variability of quantitative traits. *Genetics*, **130**, 195–204.

Jennions, M.D. & Møller, A.P. (2002) Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **269**, 43–48.

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.

Kempenaers, B., Peters, A. & Foerster, K. (2008) Sources of individual variation in plasma testosterone levels. *Philosophical Transactions of the Royal Society of London. Sires B, Biological Sciences*, **363**, 1711–1723.

Koolhaas, J.M., de Boer, S.F., Coppens, C.M. & Buwalda, B. (2010) Neuroendocrinology of coping styles: towards understanding the biology of individual variation. *Frontiers in Neuroendocrinology*, **31**, 307–321.

Kruuk, L.E.B., Clutton-Brock, T.H., Slate, J., Pemberton, J.M., Brotherstone, S. & Guinness, F.E. (2000) Heritability of fitness in a wild mammal population. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 698–703.

Laucht, S., Dale, J., Mutzel, A. & Kempenaers, B. (2011) Individual variation in plasma testosterone levels and its relation to badge size in house sparrows *Passer domesticus*: it's a night-and-day difference. *General and Comparative Endocrinology*, **170**, 501–508.

Lessells, C.M. (2008) Neuroendocrine control of life histories: what do we need to know to understand the evolution of phenotypic plasticity? *Philosophical Transactions of the Royal Society of London. Sires B, Biological Sciences*, **363**, 1589–1598.

Lessells, C.M. & Boag, P.T. (1987) Unrepeatable repeatabilities: a common mistake. *The Auk*, **104**, 116–121.

Luttbeg, B. & Sih, A. (2010) Risk, resources and state-dependent adaptive behavioural syndromes. *Philosophical Transactions of the Royal Society of London. Sires B, Biological Sciences*, **365**, 3977–3990.

Mathot, K.J. & Dingemanse, N.J. (2015) Energetics and behavior: unrequited needs and new directions. *Trends in Ecology & Evolution*, **30**, 199–206.

McGraw, K.O. & Wong, S.P. (1996) Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, **1**, 30–46.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. & PRISMA Group (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, **6**, e1000097.

Nakagawa, S. & Poulin, R. (2012) Meta-analytic insights into evolutionary ecology: an introduction and synthesis. *Evolutionary Ecology*, **26**, 1085–1099.

Nakagawa, S. & Santos, E.A. (2012) Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, **26**, 1253–1274.

Nakagawa, S. & Schielzeth, H. (2010) Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, **85**, 935–956.

Nakagawa, S. & Schielzeth, H. (2013) A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, 133–142.

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M. & Senior, A.M. (2015) Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, **6**, 143–152.

Nespolo, R.F. & Franco, M. (2007) Whole-animal metabolic rate is a repeatable trait: a meta-analysis. *Journal of Experimental Biology*, **210**, 2000–2005.

Oliveira, R.F. (2004) Social modulation of androgens in vertebrates: mechanisms and function. *Advances in the Study of Behavior*, **34**, 165–239.

Ottinger, M.A., Wu, J. & Pelican, K. (2002) Neuroendocrine regulation of reproduction in birds and clinical applications of GnRH analogues in birds and mammals. *Seminars in Avian and Exotic Pet Medicine*, **11**, 71–79.

Ouyang, J.Q., Hau, M. & Bonier, F. (2011) Within seasons and among years: when are corticosterone levels repeatable? *Hormones and Behavior*, **60**, 559–564.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Réale, D., Reader, S.M., Sol, D., McDougall, P.T. & Dingemanse, N.J. (2007) Integrating animal temperament within ecology and evolution. *Biological Reviews*, **82**, 291–318.

Réale, D., Dingemanse, N.J., Kazem, A.J.N. & Wright, J. (2010a) Evolutionary and ecological approaches to the study of personality. *Philosophical Transactions of the Royal Society of London. Sires B, Biological Sciences*, **365**, 3937–3946.

Réale, D., Garant, D., Humphries, M.M., Bergeron, P., Careau, V. & Montiglio, P.-O. (2010b) Personality and the emergence of the pace-of-life syndrome concept at the population level. *Philosophical Transactions of the Royal Society of London. Sires B, Biological Sciences*, **365**, 4051–4063.

Rensel, M.A., Boughton, R.K. & Schoech, S.J. (2010) Development of the adrenal stress response in the Florida scrub-jay (*Aphelocoma coerulescens*). *General and Comparative Endocrinology*, **165**, 255–261.

Rensel, M.A. & Schoech, S.J. (2011) Repeatability of baseline and stress-induced corticosterone levels across early life stages in the Florida scrub-jay (*Aphelocoma coerulescens*). *Hormones and Behavior*, **59**, 497–502.

Rosenthal, R. (1979) The file drawer problem and tolerance for null results. *Psychological Bulletin*, **86**, 638–641.

Schmidt, K.L., MacDougall-Shackleton, E.A., Soma, K.K. & MacDougall-Shackleton, S.A. (2014) Developmental programming of the HPA and HPG axes by early-life stress in male and female song sparrows. *General and Comparative Endocrinology*, **196**, 72–80.

Sih, A., Bell, A.M., Johnson, J.C. & Ziemba, R.E. (2004) Behavioral syndromes: an integrative overview. *The Quarterly Review of Biology*, **79**, 241–277.

Sih, A., Mathot, K.J., Moirón, M., Montiglio, P.-O., Wolf, M. & Dingemanse, N.J. (2015) Animal personality and state–behaviour feedbacks: a review and guide for empiricists. *Trends in Ecology & Evolution*, **30**, 50–60.

Small, T.W. & Schoech, S.J. (2015) Sex differences in the long-term repeatability of the acute stress response in long-lived, free-living Florida scrub-jays (*Aphelocoma coerulescens*). *Journal of Comparative Physiology B-Biochemical Systemic and Environmental Physiology*, **185**, 119–133.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Stamps, J.A., Briffa, M. & Biro, P.A. (2012) Unpredictable animals: individual differences in intraindividual variability (IIV). *Animal Behaviour*, **83**, 1325–1334.

Trikalinos, T.A. & Ioannidis, J.P. (2005) Assessing the evolution of effect sizes over time. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (eds H.R. Rothstein, A.J. Sutton & M. Borenstein), pp. 241–259. Wiley, Chichester, UK.

Versteegh, M.A., Helm, B., Dingemanse, N.J. & Tieleman, B.I. (2008) Repeatability and individual correlates of basal metabolic rate and total evaporative water loss in birds: a case study in European stonechats. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, **150**, 452–457.

Viechtbauer, W. (2012) Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, **36**, 1–48.

Westneat, D.F., Wright, J. & Dingemanse, N.J. (2015) The biology hidden inside residual within-individual phenotypic variation. *Biological Reviews*, **90**, 729–743.

Westneat, D.F., Hatch, M.I., Wetzel, D.P. & Ensminger, A.L. (2011) Individual variation in parental care reaction norms: integration of personality and plasticity. *The American Naturalist*, **178**, 652–667.

White, C.R., Schimpf, N.G. & Cassey, P. (2013) The repeatability of metabolic rate declines with time. *Journal of Experimental Biology*, **216**, 1763–1765.

Williams, T.D. (2008) Individual variation in endocrine systems: moving beyond the "tyranny of the Golden Mean". *Philosophical Transactions of the Royal Society of London. Sires B, Biological Sciences*, **363**, 1687–1698.

Wolf, M. & McNamara, J.M. (2012) On the evolution of personalities via frequency-dependent selection. *The American Naturalist*, **179**, 679–692.

Wolf, M. & Weissing, F.J. (2010) An explanatory framework for adaptive personality differences. *Philosophical Transactions of the Royal Society of London. Sires B, Biological Sciences*, **365**, 3959–3968.

Received 17 April 2016; accepted 20 September 2016
Handling Editor: Ignacio Moore

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Supplementary methods.

Appendix S2. Data used in this study. [Correction added after online publication on 27 March 2017: Appendix S2 file corrected].

Appendix S3. Supplementary tables

Appendix S4. Supplementary figures.