Received: 6 October 2016 | Accepted: 8 May 2017

DOI: 10.1111/2041-210X.12821

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

# Inter-laboratory variation in corticosterone measurement: Implications for comparative ecological and evolutionary studies

Kerry V. Fanson[1] | Zoltán Németh[2,3] | Marilyn Ramenofsky[1,2] | John C. Wingfield[1,2] | Katherine L. Buchanan[1]

[1]Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, Australia

[2]Department of Neurobiology, Physiology and Behavior, University of California Davis, Davis, CA, USA

[3]Department of Evolutionary Zoology, MTA-DE "Lendület" Behavioural Ecology Research Group, University of Debrecen, Debrecen, Hungary

**Correspondence**
Kerry V. Fanson
Email: kerry.fanson@deakin.edu.au

**Funding information**
Australian Research Council, Grant/Award Number: FT140100131; National Science Foundation, Grant/Award Number: ARC-1147289; National Research, Development and Innovation Office of Hungary, Grant/Award Number: NKFIH PD 121013; Deakin University, Grant/Award Number: RM28479

Handling editor: Erica Leder

## Abstract

1. Interspecific comparisons of endocrine data are useful for drawing broad conclusions concerning the role of ecological variables in the evolution of physiological pathways. However, comparisons of endocrine data generated by different research groups are problematic, due to inter-laboratory variation in measured hormone values. To date, we know of no study which has quantified the extent of inter-laboratory variation in the measurement of hormone levels, outside of biomedical studies.
2. To evaluate the extent to which laboratories differ in their measurement of hormones, we prepared seven samples of avian plasma with known concentrations of corticosterone and sent them for blind analyses to 19 laboratories and asked them to report the methods used and the values obtained.
3. Both absolute hormone concentrations and the ratios between samples were equally variable, up to an order of magnitude different for some concentrations. Laboratory identity accounted for more than 80% of the variation in reported corticosterone, but we could not identify any methodological factors that consistently contributed to this inter-laboratory variation. In addition, laboratory measurement error was significantly correlated with the latitude of the primary study species for each laboratory, suggesting that inter-laboratory variation has the potential to drive trends in corticosterone datasets.
4. Inter-laboratory variation in corticosterone measurement may have serious implications for quantitative comparisons of endocrine values across laboratories, although comparisons of qualitative patterns may be more robust because rank order of the samples was relatively consistent across laboratories. Ignoring laboratory effect and the non-independence of data may lead to an inflated rate of Type I error and spurious correlations.

KEYWORDS
comparative endocrinology, corticosterone, enzyme immunoassay, measurement error, meta-analysis, radioimmunoassay, stress

## 1 | INTRODUCTION

Hormones have the capacity to influence many aspects of behaviour, physiology and life history, and consequently there is much interest in the evolutionary forces that shape endocrine traits. The comparison of endocrine traits across species can reveal the costs and benefits of different endocrine strategies, informing our understanding of both the function and evolution of endocrine systems (Forlano, Schlinger, & Bass, 2006; Garland, Bennett, & Rezende, 2005; Thornton, Need, & Crews, 2003; Zera, Harshman, & Williams, 2007). Historically, comparative endocrinology studies have focussed on hormone structure or the underlying genetics (e.g. Forlano et al., 2006; Thornton et al., 2003). However, an increasing number of studies seek to compare hormone concentrations across species in an effort to understand how hormone expression may be shaped by factors such as life history, ecology, environment and behaviour (e.g. Baker, Gobush, & Vynne, 2013; Barron, Crespi, & Schwabl, 2015; Bókony et al., 2009; Eikenaar, Husak, Escallon, & Moore, 2012; Foo, Nakagawa, Rhodes, & Simmons, 2017; Garamszegi, Eens, Hurtrez-Boussès, & Møller, 2005; Garamszegi et al., 2008; Goymann, 2009; Goymann, Landys, & Wingfield, 2007; Goymann & Wingfield, 2014; Goymann et al., 2004; Hau, Ricklefs, Wikelski, Lee, & Brawn, 2010; Hirschenhauser & Oliveira, 2006; Hirschenhauser, Winkler, & Oliveira, 2003; Jessop, Woodford, & Symonds, 2013; Jessop et al., 2016; Lendvai, Bókony, Angelier, Chastel, & Sol, 2013; Moore, Shuker, & Dougherty, 2016; Oliveira, Hirschenhauser, Carneiro, & Canario, 2002; Roberts, Buchanan, & Evans, 2004; Romero, 2002; Swanson & Dantzer, 2014). The growth of these phylogenetic comparative studies in endocrinology has been fuelled by the rapid increase in species for which endocrine data are available. Efforts to assemble large, open-access endocrine databases have created rich datasets for comparative studies, and we expect the number of such studies will continue to increase.

A key assumption of these studies is that hormone values are comparable across laboratories. Most phylogenetic comparative endocrine studies do not even mention the potential importance of inter-laboratory variation affecting comparisons of hormone measurements across laboratories (see Table 1 for glucocorticoid studies). Some studies acknowledge that there may be inter-laboratory variation, but dismiss its importance by arguing that it would only increase random noise, thereby making it more difficult to detect an effect (Garamszegi et al., 2005; Goymann et al., 2004), or assume that laboratory differences cause minor error, if any (Garamszegi et al., 2008). However, the one study we found that directly tested for a laboratory effect reported a significant effect of laboratory, accounting for almost 40% of the variance in baseline and peak corticosterone (Bókony et al., 2009). That study could not completely disentangle laboratory measurement error from study species. Here, we directly quantify the variation among laboratories due to measurement error, thereby offering valuable information about inter-laboratory variation.

Understanding measurement error among laboratories is also an important step in addressing the “reproducibility crisis” that is facing many empirical fields (Forstmeier, Wagenmakers, & Parker, 2016; Nakagawa & Parker, 2015). In recent years, there has been growing recognition of the fact that many published findings cannot be replicated when a study is repeated (Nakagawa & Parker, 2015). Efforts to resolve this lack of reproducibility have largely focused on reporting issues (Parker et al., 2016) or statistical issues (Forstmeier et al., 2016). However, inter-laboratory variation, or laboratory transfer issues, is another important factor that could potentially contribute to this lack of reproducibility.

Protocols often vary among laboratories (e.g. extraction procedure, assay type, assay constituents), thereby leading to variation among laboratories in measured concentration (Wingfield, Hegner, Dufty, & Ball, 1990). Even if laboratories follow identical protocols, hormone measurements may vary among laboratories due to a range of factors, including water quality, pH and temperature (Feswick et al., 2014; Garde, Hansen, & Nikolajsen, 2003). These issues have been acknowledged for some time within the biomedical literature, where studies focus on humans as a single species, but can report vastly different measurements (Falk et al., 1999; Gail et al., 1996; Garde et al., 2003). Consequently, there is a concerted effort to standardize measurements across biomedical laboratories, primarily using recognized reference standards to calibrate assays (Myers, 2008; Vesper, Botelho, Shacklady, Smith, & Myers, 2008). However, this is not the case among ecological/wildlife laboratories.

The goal of this study was to (1) provide empirical quantification of the variance among laboratories in the measurement of plasma corticosterone values, (2) identify methodological factors that contribute to inter-laboratory variation and (3) consider the implications of inter-laboratory variation for comparative endocrine studies. A total of 19 laboratories that routinely measure corticosterone were asked to analyse aliquots of seven samples that varied in corticosterone concentration. The range of concentrations reflected actual corticosterone concentrations reported in birds. First, we examined which hormone measurements were most comparable among laboratories (absolute hormone concentrations or proportional response). Second, we tested for an effect of methodology (extraction method, assay type, number of replicates and loading volume) to see if methodological differences could account for any of the variation among laboratories. Finally, we examined the relationship between laboratory measurement error and traits of the primary study species for each laboratory to see if there is potential for inter-laboratory variation to introduce structured error in corticosterone datasets. Quantifying and understanding inter-laboratory variance is important for drawing biologically relevant conclusions from evolutionary endocrinology studies, and this study provides the first empirical assessment of this variance.

## 2 | MATERIALS AND METHODS

### 2.1 | Sample preparation

Chicken blood was collected at a commercial abattoir and stored at 4°C in large flask with sodium citrate (3 g/L in 10 ml phosphate-buffered saline) to prevent coagulation during transport. Blood was centrifuged and the plasma removed and mixed overnight with dextran-coated charcoal (Sigma C6241) to remove the native

**TABLE 1** Published meta-analyses on glucocorticoids and whether or not they considered laboratory effect. "Laboratory included" indicates whether they controlled for laboratory ID in their final model

| Reference | Taxa | Focus of study | Response variable(s) | "Laboratory" included? | Comments about laboratory effect | Covariates |
|---|---|---|---|---|---|---|
| Baker et al. (2013) | Fish + other vertebrates | Human disturbance | Qualitative change | No | No mention | |
| Bókony et al. (2009) | Birds | Parental care | Absolute conc<br>Relative magnitude of stress response<br>Relative difference (males vs. females) | No | Ran an initial model testing for the effect of "laboratory" and several methodological factors. Found a significant effect of laboratory (Baseline: $F_{14,177} = 7.81$, $p < .001$; Peak: $F_{9,89} = 5.79$, $p < .001$), but did not include it in final model | Brood value, sex differences in care, body mass, latitude |
| Dantzer et al. (2014) | Vertebrates | Human disturbance | Relative difference (control vs. disturbed) | No | No mention | |
| Eikenaar et al. (2012) | Amphibians & reptiles | Distribution, breeding season | Absolute conc | No | To check for laboratory effect, re-ran analyses for just the genus *Bufo* and found effect of latitude was similar to full amphibian dataset | |
| Hau et al. (2010) | Birds | Life history | Absolute conc | No | Initially ran analyses with just data from their laboratory, and then re-ran with values from the literature and found similar trends | Latitude, body mass, adult survival[a], breeding season length[a] |
| Jessop et al. (2013) | Reptiles & birds | Distribution, environment | Absolute conc | No | No mention | |
| Jessop et al. (2016) | Vertebrates | Temperature | Absolute conc<br>Relative difference (low vs. high temp) | No | No mention | |
| Lendvai et al. (2013) | Birds | Relative brain size | Absolute conc | No | No mention | Brood value, body mass, brain mass, latitude, migration[a] |
| Moore et al. (2016) | Vertebrates | Secondary sexual traits | Absolute conc | No | No mention | |
| Romero, (2002) | Vertebrates | Season | Qualitative change | No | No mention | |

"Covariates" are the variables extracted from studies with data on at least 75% of the species on our list.
[a]Indicate variables that were excluded from analysis due to too much missing data.

steroid hormones. The plasma was then centrifuged and filtered twice (Whatman no 1) and stored at −20°C. Seven samples were prepared from this pool of stripped chicken plasma. The samples were spiked with corticosterone (Sigma 27840) to create seven samples that ranged in concentration from 0.05 to 40 ng/ml (0.05, 0.25, 0.5, 2, 7, 18, 40 ng/ml). This range of concentrations covered nearly 80% of the reported corticosterone values in birds, both baseline and peak (Bókony et al., 2009). The samples were randomly ordered with regards to concentration and labelled only with the number 1–7. Participating laboratories had no knowledge of the concentration of these unknown samples.

## 2.2 | Sample distribution

We approached 19 research groups that routinely analyse avian plasma corticosterone and publish their work in the peer-reviewed literature. All laboratories that were approached agreed to participate and complete a questionnaire about sample handling procedures in their

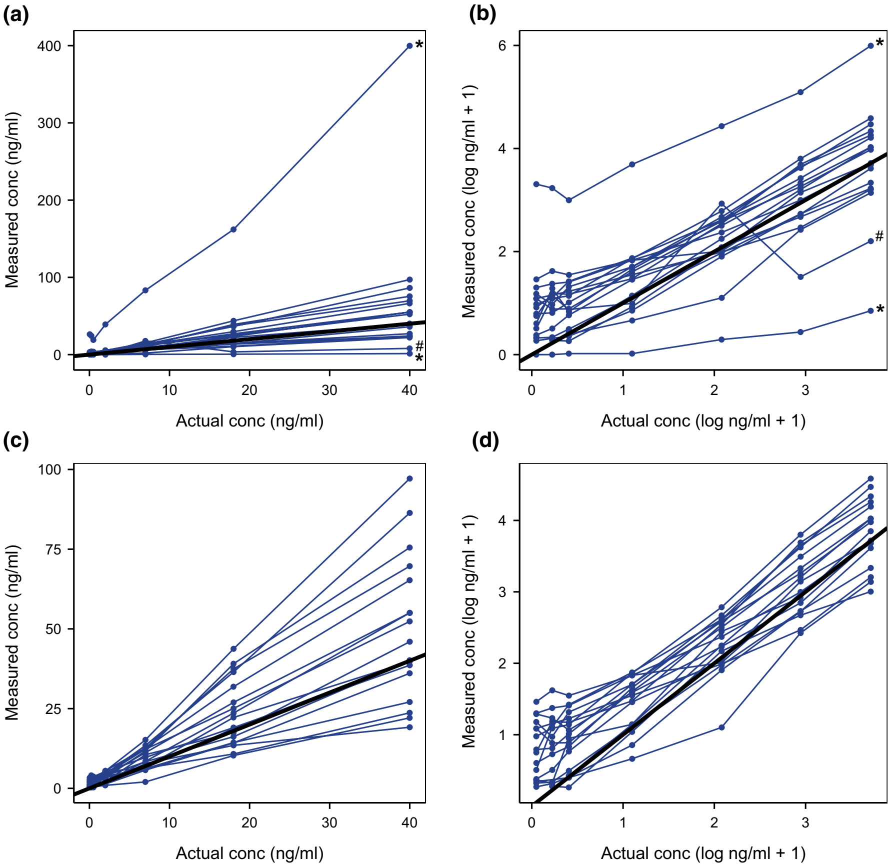

**FIGURE 1** Inter-laboratory variation in the measurement of seven samples ranging in concentration from 0.05 to 40 ng/ml. (a) Originally reported data, not transformed. (b) Originally reported data, log transformed. *indicates laboratories that initially miscalculated the reported concentration and calculations were corrected for final analyses. #indicates the laboratory that was excluded from analyses because many values fell off the standard curve. (c) Final dataset, not transformed (see text for explanation). (d) Final dataset, log transformed. Thin blue lines represent individual laboratories; thick black line represents slope of actual concentrations

laboratory (see Appendix S1). Laboratories were located in Australia, North America and Europe. Samples were shipped on dry ice using couriers that frequently topped-up dry ice levels and transported packages in climate-controlled vehicles, thereby ensuring that samples remained frozen during shipping.

## 2.3 | Statistical analysis

Data analysis was conducted in R (version 3.2, R core development). The original data as reported are shown in Figure 1a,b. Participants were asked to treat the samples and report results the same as they would published data (including samples that fell outside the range of the standard curve), and therefore we used all data that were reported to us. For seven laboratories, all values were detectable on the first assay. Five laboratories re-assayed samples ($n = 11$) that fell outside the detection limit of the standard curve using a different dilution factor and reported those values. Six laboratories reported that some sample values were at or near the edge of the detection limit ($n = 15$; 13 low concentration, 2 high concentration), but values were still reported (either extrapolated or assigned the min/max detectable value). Three laboratories reported sample values as “not detectable” and no values were provided ($n = 4$), so those were not included in analyses.

For final data analyses, the following modifications were made to the dataset (Figure 1c,d). Two laboratories initially reported values that were extremely different than the rest of the laboratories. Upon further inquiry, we discovered that the dilution factor was incorrectly accounted for, so we corrected the calculations and used the corrected values for all subsequent analyses. One laboratory was excluded because four of the seven samples fell outside the range of the standard curve and the remaining three samples did not show the expected trend. In cases where the laboratory ran at least four replicates of each sample, we excluded single replicates that fell further than 1.5 *SD* from the mean for that laboratory ($n = 9$ of 403 replicates), which was in line with comments that participants made about values they would consider outliers. We did not delete more than one replicate per sample per laboratory.

### 2.3.1 | Patterns of inter-laboratory variation

Our first goal was to identify where inter-laboratory variance might be introducing the most noise in meta-analytic studies. In other words, are some hormone measures more consistent across laboratories than others? Many meta-analyses use the absolute hormone concentration reported in the literature (Table 1; e.g. Jessop et al., 2013; Lendvai et al., 2013). To estimate the variability among laboratories for absolute measurements, we calculated the %*CV* on reported values.

Other meta-analytic studies use the proportional response, such as the ratio between peak and baseline or between two groups (Table 1; e.g. Jessop et al., 2016). It is often assumed that even if absolute hormone concentrations vary among laboratories, the relative increase or decrease measured should be fairly consistent across laboratories. To this end, we calculated the ratio between two pairs of samples: (1) 2–18 ng/ml, which

**TABLE 2** Patterns of variability among laboratories using absolute hormone concentrations and proportional response

| | # Laboratories | Mean (range) | %CV among laboratories |
|---|---|---|---|
| Individual samples | | | |
| 0.05 ng/ml | 16 | 1.39 ng/ml (0.31–3.31) | 69.91 |
| 0.25 ng/ml | 16 | 1.59 ng/ml (0.32–4.06) | 67.20 |
| 0.50 ng/ml | 18 | 1.72 ng/ml (0.30–3.70) | 59.71 |
| 2.00 ng/ml | 18 | 3.52 ng/ml (0.94–5.50) | 41.68 |
| 7.00 ng/ml | 18 | 9.29 ng/ml (2.01–15.20) | 36.70 |
| 18.00 ng/ml | 18 | 23.39 ng/ml (10.29–43.75) | 44.45 |
| 40.00 ng/ml | 18 | 49.39 ng/ml (19.16–97.15) | 45.16 |
| Ratio | | | |
| 18 ng/ml ÷ 2 ng/ml | 18 | 7.25 (2.91–11.63) | 36.48 |
| 40 ng/ml ÷ 0.5 ng/ml | 18 | 38.72 (9.83–103.83) | 60.62 |

both fell within the linear part of the standard curve for most laboratories, and (2) 0.5–40 ng/ml, which often fell closer to the ends of the curve but still had reported values for all laboratories. We then calculated %*CV* for the resulting ratios.

### 2.3.2 | Quantifying laboratory repeatability and identifying sources of variance

To examine potential sources of variation among laboratories and determine how much of the variance is due to laboratory effect, we ran a random slopes model (Package LME4, function lmer). Laboratory ID was modelled as a random intercept. Mean sample values for each laboratory were log transformed to standardize variance across the range of concentrations. This was modelled as a function of the actual sample concentration, which was log transformed for analyses and mean centred to calculate repeatability (Nakagawa & Schielzeth, 2010). The following fixed effects were included in the model: continent, extraction method, assay type, number of replicates and volume of sample loaded in the wells/tubes (See Appendix S1 for distribution). The proportion of the variance explained by laboratory ID (conditional repeatability) was calculated as the intra-class correlation coefficient (Nakagawa & Schielzeth, 2010). The 95% confidence interval was determined using parametric bootstrapping.

### 2.3.3 | Implications for comparative CORT studies

Our final aim was to consider the implications of this inter-laboratory variation for comparative endocrine studies. Specifically, we wanted to know whether within this dataset, laboratory measurement error correlated with traits of the study species, thereby introducing structure to corticosterone datasets and potentially leading to spurious results in meta-analyses (e.g. do laboratories with relatively high corticosterone measurements also tend to study longer-lived species?). To identify the primary study species for each laboratory, we conducted a Web of Science search with the laboratory leader's name and the keyword "corticosterone", and tallied the number of papers on each avian species. We only considered papers published from 2009 to 2015 because the results reported in this study would be most reflective of recently published corticosterone data. Review articles, book chapters and conference proceedings were not included. From these data, we identified the single species with the greatest number of publications for each laboratory leader. Two laboratories had the same primary study species.

To identify species traits, we looked at the supplementary material for all CORT meta-analyses (Table 1). Three papers included data for at least 75% of the primary study species identified above (Bókony et al., 2009; Hau et al., 2010; Lendvai et al., 2013). Missing trait values were obtained from the literature where possible. The final list of traits included body mass, brain mass, relative brain size, brood value, sex differences in parental care (negative scores indicate female-biased investment in incubation, chick feeding and brooding and positive scores represent male-biased investment; Bókony et al., 2009) and latitude (Table 1).

The deviation for each laboratory was quantified using the random intercept for each laboratory obtained from the random slopes model (as described above; R function ranef). We ran this analysis two ways; first with the corticosterone data left centred to reflect inter-laboratory variation in "baseline" measurements of corticosterone, and second with the data right centred to reflect variation in "peak" measurements. We ran a linear regression (R package lm) to test the relationship between laboratory deviation (intercept) and each covariate listed above. The analysis included 16 of the 19 participating laboratories (2 participants were early career researchers who had not yet published data from their own laboratory and one participant had not published a paper on bird corticosterone during the specified timeframe).

## 3 | RESULTS

### 3.1 | Patterns of inter-laboratory variation

Absolute hormone concentrations were highly variable among laboratories, with the %*CV* ranging from 37 to 70%, depending on the concentration (Table 2). The %*CV* generally decreased as sample concentration increased. Interestingly, the proportional response (ratio between two concentrations) was similarly variable among laboratories (36–60%*CV*; Table 2). It is often assumed that relative measures may be more consistent across laboratories (i.e. lower %*CV*). However, our data suggest that relative measures of endocrine function may not be any more comparable among laboratories than absolute values. Nonetheless, rank order of the top five samples (≥0.5 ng/ml) was consistently identified across laboratories. However, only 10 of 18

**TABLE 3** Effect of laboratory methods on measured hormone concentration. The intra-class correlation coefficient for Laboratory ID is 86.18%

| Variable | Num DF | Den DF | *F*-value | *p*-value |
|---|---|---|---|---|
| Fixed effects | | | | |
| Sample concentration | 1 | 16.89 | 551.77 | <.001 |
| Continent | 2 | 8.01 | 0.26 | .78 |
| Extraction method | 3 | 7.98 | 0.46 | .72 |
| Assay type | 1 | 7.99 | 0.23 | .64 |
| # replicates | 2 | 7.98 | 0.31 | .75 |
| Loading volume | 1 | 7.97 | 0.73 | .42 |
| Random effects | Variance | | | |
| Laboratory intercept | 0.17 | | | |
| Laboratory slope | 0.02 | | | |
| Error | 0.03 | | | |

laboratories correctly identified the rank order of the bottom three samples (≤0.5 ng/ml; Figure 1).

## 3.2 | Quantifying laboratory repeatability and identifying sources of variance

Laboratory ID accounted for over 80% of the variance in the data (repeatability = 0.86; 95% CI = 0.63–0.94; Table 3), which indicates that there were consistent differences among laboratories across the range of measured concentrations. However, we did not identify any methodological differences that accounted for these differences. None of the fixed effects included in the model had a significant effect on measured concentration (Table 3).

## 3.3 | Implications for comparative CORT studies

We did not find a relationship between laboratory measurement error of "baseline" corticosterone and any of the species traits (body mass, brain mass, relative brain size, brood value, sex differences in parental care, latitude; Table 4; Figure 2a,c). However, there was a significant positive relationship between laboratory deviance in "peak" corticosterone and latitude of study species (Table 4; Figure 2b). There was also a trend towards a negative relationship between "peak" corticosterone measurement error and relative brain size (Table 4; Figure 2d).

**TABLE 4** Relationship between the measurement of corticosterone concentration ("Laboratory Intercept") and covariates that have been used in meta-analyses

| | Baseline | Peak |
|---|---|---|
| log(Body mass) | $F_{1,14} = 0.0008$<br>$p = .98$ | $F_{1,14} = 2.99$<br>$p = .11$ |
| log(Brain mass) | $F_{1,14} = 0.06$<br>$p = .81$ | $F_{1,14} = 1.57$<br>$p = .23$ |
| Relative brain size | $F_{1,14} = 0.02$<br>$p = .90$ | $\mathbf{F_{1,14} = 3.73}$<br>$\mathbf{p = .07}$ |
| Brood value | $F_{1,10} = 0.0004$<br>$p = .98$ | $F_{1,10} = 0.72$<br>$p = .42$ |
| Sex difference in parental care | $F_{1,9} = 0.30$<br>$p = .60$ | $F_{1,9} = 1.42$<br>$p = .26$ |
| Latitude | $F_{1,14} = 0.16$<br>$p = .69$ | $\mathbf{F_{1,14} = 6.30}$<br>$\mathbf{p = .025}$ |

"Baseline" indicates random intercept for laboratory was obtained from a model in which corticosterone data were left centred; "Peak" indicates data were right centred.
Bold indicates $p < .1$.

# 4 | DISCUSSION

The number of phylogenetic comparative studies in endocrinology is rapidly increasing. The substantial body of literature which now exists on hormone levels from non-model organisms allows for fascinating questions to be posed concerning the role of the endocrine system in mediating evolutionary change in terms of life history (Barron et al., 2015; Bókony et al., 2009; Eikenaar et al., 2012; Goymann & Wingfield, 2014; Goymann et al., 2004; Hau et al., 2010; Romero, 2002; Swanson & Dantzer, 2014); reproductive strategy/Challenge Hypothesis (Garamszegi et al., 2005; Goymann, 2009; Goymann et al., 2007; Hirschenhauser & Oliveira, 2006; Hirschenhauser et al., 2003; Moore et al., 2016; Oliveira et al., 2002); immune function (Foo et al., 2017; Roberts et al., 2004); distribution/environment (Eikenaar et al., 2012; Garamszegi et al., 2008; Goymann et al., 2004; Jessop et al., 2013, 2016); human disturbance (Baker et al., 2013; Dantzer, Fletcher, Boonstra, & Sheriff, 2014) and cognition (Lendvai et al., 2013). These data also potentially allow identification of the constraints on the evolution of endocrine traits. However, these broad comparative approaches rely on comparisons of endocrine data across laboratories which may vary in several methodological (e.g. extraction procedure, assay type) and non-methodological factors (e.g. water chemistry, temperature). Here, we have attempted to quantify this variation and the implications for the conclusions drawn from such studies.

## 4.1 | How much variation is there in corticosterone measurement among laboratories?

Our results identified substantial variation among laboratories in their measurement of standardized endocrine samples. In this controlled

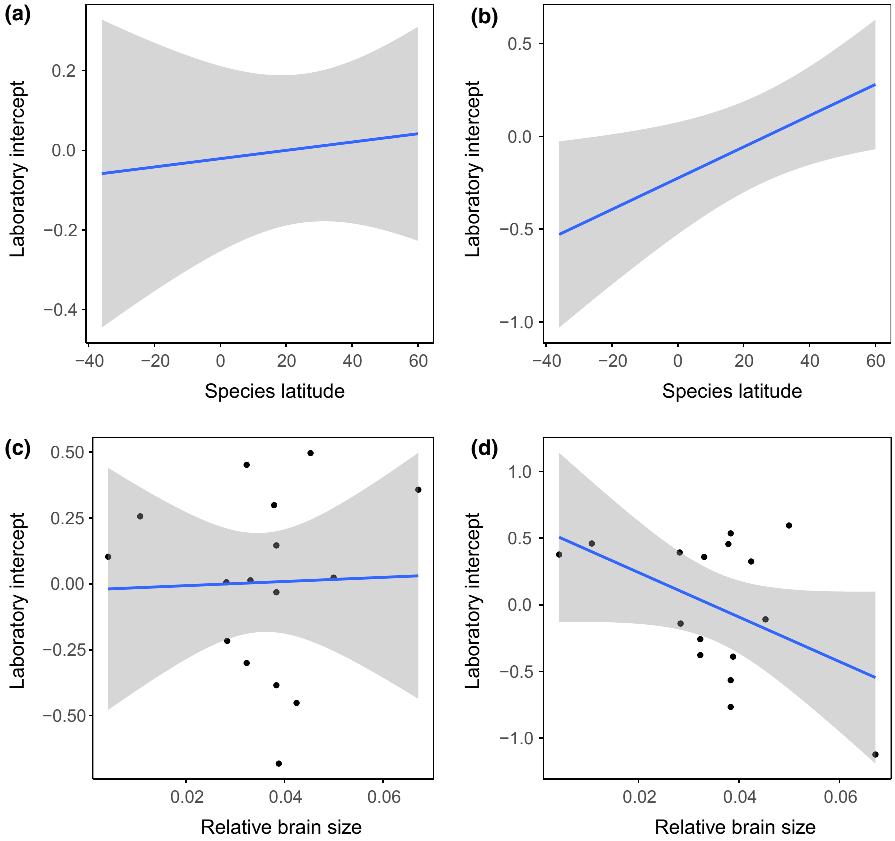

**FIGURE 2** Relationship between laboratory deviation in the measurement of corticosterone concentration ("Laboratory Intercept") and species traits of each lab's primary study species: (a, b) latitude of species' distribution; (c, d) relative brain size. Laboratory measurement error was estimated for both "baseline" (a, c) and "peak" (b, d) corticosterone concentrations using the random intercept for laboratory obtained from a model in which corticosterone data were left centred or right centred, respectively. For latitude, actual points are not shown to preserve the anonymity of participating laboratories

study, laboratory ID accounted for more than 80% of the variance in the data. Even in datasets that include data from multiple species, laboratory was found to account for almost 40% of the variance in corticosterone (Bókony et al., 2009). The reported measurements varied 4- to 13-fold depending on the concentration (Table 2) For example, reported values for one sample ranged from 0.30 to 3.7 ng/ml—an order of magnitude. The variance across laboratories for absolute hormone concentration ranged from 37 to 70%*CV*, depending on sample concentration (Table 2). It is worth noting that in comparison, the accepted level of within-study measurement error (intra- or inter-assay variation) is generally <15%*CV* to ensure meaningful comparisons.

Our study represents a conservative estimate of the actual inter-laboratory variation that may be present in the literature. One laboratory was excluded from analyses because several samples fell below the minimum detectable limit and the remaining samples did not show the expected trend, although the participant did not flag any concerns about their data. In addition, two laboratories (10% of participants) did not properly calculate final concentrations, leading to initially reported values that were off by an order of magnitude (Figure 1a,b). Although we could identify and correct this mistake, the reality is that these errors will likely exist in the published literature. Consequently, some number of published corticosterone values (possibly as many as 1 in 10 based on our experience) may be wrong by an order of magnitude.

Our study also conservatively estimates the error introduced from cross-reaction with other steroids hormones. We used plasma samples that had been stripped (thereby removing all steroids) and spiked with pure corticosterone. In reality, blood samples contain several different steroid hormones, which may cross-react differently with different antibodies and increase inter-laboratory variation. Also, our study does not factor in differences in collection methods (how animals are caught, time to sample collection) or variation in sample composition (lipid content, binding globulins). Given the already large inter-laboratory variance present in our dataset, where we attempted to control for as many factors as possible, we can project that the actual inter-laboratory variation present in the literature is not trivial.

Although we can clearly state that there are differences among laboratories, we do not know if these differences are maintained over time. Assay performance can shift over time, which may lead to changes in the rank order among laboratories. If laboratory effects degrade quickly over time, then the laboratory effect will not be confounded with study species, and will just add noise to the data. However, laboratories often use controls to maintain performance of their assays over time, so it is plausible that laboratory differences may be maintained. If this is the case, then this results in non-independence of measurements from a laboratory, and must be accounted for in statistical models (see below).

## 4.2 | Are there methodological factors that contribute to this variation?

While our study identified substantial inter-laboratory variation in the measurement of corticosterone, we were not able to identify any methodological predictors that explained this variation. This corresponds with previously published findings that laboratory ID had a significant effect on measured corticosterone values, but none of the individual methodological factors had a significant effect (assay sensitivity, intra-assay *CV*, inter-assay *CV*, assay recovery, chromatographic separation, sample size; Bókony et al., 2009). Therefore, at this stage we cannot identify any methodological factors that should be controlled for in comparative studies. It is worth noting that we did not find any difference between samples analysed via RIA or EIA, alleviating some of the concern about the transition from RIAs to EIAs. Similarly, extraction did not have a significant effect. However, extraction may be much more important for actual samples, where interfering substances may vary across states (reproductive status, stage of migration, body condition, etc.). Therefore, we cannot definitely conclude that differences in extraction protocol do not produce significant variance in the estimated hormone concentration, only that with this dataset we could not detect any systematic effect.

## 4.3 | Are some endocrine measurements more comparable across laboratories?

All laboratories included in the analyses correctly identified the rank order of the top five samples (≥0.5 ng/ml). However, only about half of the laboratories (10 of 18 included in analysis) correctly identified the rank order of the bottom three samples (≤0.5 ng/ml; Figure 1). This indicates that qualitative trends are largely robust across laboratories at higher concentrations, but may not be reliable at low concentrations. Therefore, qualitative patterns offer useful comparisons for comparative studies (e.g. Baker et al., 2013; Romero, 2002).

Although *qualitative* trends are comparable across laboratories, *quantitative* measurements (absolute hormone concentrations and ratios) are not. It is sometimes argued that relative endocrine measures are more comparable among laboratories. However, we found ratios to be almost as variable as absolute hormone values (36–60% *CV*). Therefore, we suggest the best practice for comparative endocrine studies is to use qualitative patterns rather than quantitative measurements of endocrine function.

## 4.4 | How does the magnitude of inter-laboratory variation compare to effects reported in published meta-analyses?

We found that reported values often varied ~10-fold (Table 2). In comparison, many meta-analyses report significant effects of much smaller magnitude. For example, Bókony et al. (2009) found that baseline corticosterone values only increased 1.5-fold as brood value increased. Eikenaar et al. (2012) reported that baseline corticosterone values increased 1.5-fold and 2.2-fold across latitude for amphibians and reptiles, respectively. This demonstrates that inter-laboratory variation may be quite large relative to effects reported in comparative studies.

## 4.5 | Implications of inter-laboratory variation for comparative endocrine studies

This study empirically demonstrates that there is a considerable variation in the measurement of corticosterone among laboratories. However, the question remains whether this has any implications for studies that seek to compare hormone measures among laboratories. Most comparative studies do not even mention the issue of laboratory differences, and we did not find any that controlled for the effect of laboratory in their main model (Table 1). Some studies acknowledge that there may be inter-laboratory variation, but dismiss its importance (Garamszegi et al., 2005, 2008; Goymann et al., 2004). A handful of studies acknowledge inter-laboratory variation and make some effort to test for it (see Table 1 for CORT studies). Often, these are indirect tests in which a subset of the data is analysed separately (Eikenaar et al., 2012; Hau et al., 2010), or species repeatability is used to gauge the presence of a laboratory effect (Garamszegi et al., 2005); Table 1). We were only able to find one comparative study on glucocorticoids that directly tested for a laboratory effect and found that it was significant, accounting for almost 40% of the variance in baseline and peak corticosterone (Bókony et al., 2009). Despite this finding, their main model did not control for laboratory.

To illustrate the potential of inter-laboratory differences to introduce structured error into hormone datasets, we examined the relationship between laboratory measurement error and six traits that are commonly considered in comparative studies. We found that measurement error of “peak” corticosterone was strongly correlated with latitude of the study species and marginally related to relative brain size (Table 4; Figure 2). Although this latter relationship is somewhat influenced by one point, these data are represented in the literature and in published meta-analyses. Inter-laboratory variation may not always introduce systematic error in corticosterone datasets, but these results highlight the potential for this type of error to generate spurious results in studies that compare endocrine data across laboratories.

## 4.6 | Recognizing the non-independence of values from the same laboratory

One major flaw with ignoring inter-laboratory variation is that estimates from one laboratory cannot be considered independent. A fundamental assumption of statistics is that the observations are independent. If related points are treated as independent, this will increase the rate of Type I error in which the null hypothesis is incorrectly rejected, thereby generating more false positives (Garland et al., 2005). Laboratories almost always publish on more than one species, and consequently meta-analyses often include multiple estimates from the same laboratory. Especially given the magnitude of the inter-laboratory variation that we found, measurements from one laboratory are expected to be more closely related than estimates coming from different laboratories. This issue of non-independence is further exacerbated by the fact that laboratories tend to study species that share certain phylogenetic, morphological, ecological or life-history characteristics. As discussed above, the importance of this depends on how rapidly the laboratory effect degrades over time, which remains unknown at this point.

The importance of controlling for non-independence is widely recognized in ecology and evolution when there are repeated measurements from individuals, populations or phylogenetic groups. Indeed, all of the comparative CORT studies we found controlled for phylogenetic non-independence, and some even controlled for the non-independence that arose from having repeated observations from the same species or the same study (Moore et al., 2016). It is therefore slightly surprising that the issue of non-independence with regard to repeated measures coming from the same laboratory has been largely overlooked.

### 4.7 | Recommendations for comparative endocrine studies

Based on our findings and a review of the literature, we recommend the following suggestions for comparative studies of endocrine measures across laboratories:

1. Whenever possible, it is better to use *qualitative* rather than *quantitative* measures of endocrine function (e.g. Baker et al., 2013, Romero, 2002). Both absolute hormone measures and ratios were highly variable among laboratories. Standardized effect sizes, which assess the strength of the signal relative to the variance, may also be useful for comparative studies, but we could not assess this with our dataset.
2. Studies should control for the non-independence of measurements from the same laboratory by including laboratory ID as a random effect. Just as it has become routine to control for phylogenetic non-independence in comparative studies, laboratory should also be accounted for. Not only will this reduce the occurrence of false positives, but it will also account for more of the variance in the dataset and lead to more robust estimates.
3. If using quantitative measurements, data should be log transformed to standardize the error introduced by laboratory effect across concentrations
4. We did not identify any key methodological factors that accounted for inter-laboratory variation, so we cannot suggest any methodological correlates that comparative studies need to consider.
5. If a key goal of the field is to be able to compare absolute hormone measures across laboratories, then we strongly suggest implementing practices to standardize measurements across laboratories, similar to the medical field. This includes establishing recognized reference standards that can be used to calibrate assays and formalizing reporting requirements when publishing (Myers, 2008; Vesper et al., 2008).

## 5 | CONCLUSIONS

The aim of our study is not to halt the growth of interest in comparative studies that may offer fascinating insights into the evolution of endocrine traits. Our goal is to raise awareness of sources of error in endocrine datasets and the potential implications of this error for comparative studies. Using measures that are comparable among laboratories and properly accounting for non-independence of observations from the same laboratory will help minimize the occurrence of spurious results and should help identify actual relationships. This should lead to more robust findings and help advance the field of comparative endocrinology.

## ACKNOWLEDGEMENTS

We would like to thank the research group leaders and their associated laboratory staff who contributed data to this study, without which this paper would not be possible: Greg Ball, Creagh Breuner, Kristal Cain, Ondi Crino, Pierre Deviche, Caitlin Gabor, Wolfgang Goymann, Michaela Hau, Ellen Ketterson, Alexander Kitaysky, Jesse S. Krause, Lynn Martin, Anna Miltiadous, Ignacio Moore, Michael Romero, Steve Schoech, Hubert Schwabl, Kiran Soma, Nina Svedin and Maren Vitousek. Thanks to B. Fanson for statistical advice. This project was supported by ARC Future Fellowship FT140100131 to K.L.B., National Science Foundation ARC-1147289 to M.R., National Research, Development and Innovation Office of Hungary fellowship (NKFIH PD 121013) to Z.N. and Deakin University Thinkers in Residence Fellowship RM28479 to M.R. and J.C.W. The authors do not declare any conflict of interest.

## AUTHORS' CONTRIBUTIONS

K.F. and K.B. conceived of the project and designed the methodology; K.B., M.R. and Z.N. led data collection; K.F. analysed the data and led the writing of the manuscript. All authors contributed critically to discussion of ideas, revision of the manuscript and gave final approval for publication.

## DATA ACCESSIBILITY

All data used in this manuscript are present in the manuscript and its supporting information (Table S1).

## REFERENCES

Baker, M. R., Gobush, K. S., & Vynne, C. H. (2013). Review of factors influencing stress hormones in fish and wildlife. *Journal for Nature Conservation*, *21*, 309–318.

Barron, D., Crespi, E., & Schwabl, H. (2015). Meta-analytical evaluation of the cort-fitness hypothesis. *Integrative and Comparative Biology*, *55*, E10.

Bókony, V., Lendvai, Á. Z., Liker, A., Angelier, F., Wingfield, J. C., & Chastel, O. (2009). Stress response and the value of reproduction: Are birds prudent parents? *The American Naturalist*, *173*, 589–598.

Dantzer, B., Fletcher, Q. E., Boonstra, R., & Sheriff, M. J. (2014). Measures of physiological stress: A transparent or opaque window into the status, management and conservation of species? *Conservation Physiology*, *2*, 1–18.

Eikenaar, C., Husak, J., Escallon, C., & Moore, I. T. (2012). Variation in testosterone and corticosterone in amphibians and reptiles: Relationships with latitude, elevation, and breeding season length. *The American Naturalist*, *180*, 642–654.

Falk, R. T., Gail, M. H., Fears, T. R., Rossi, S. C., Stanczyk, F., Adlercreutz, H., … Ziegler, R. G. (1999). Reproducibility and validity of radioimmunoassays for urinary hormones and metabolites in pre- and postmenopausal women. *Cancer Epidemiology Biomarkers & Prevention*, *8*, 567–577.

Feswick, A., Ankley, G. T., Denslow, N., Ellestad, L. E., Fuzzen, M., Jensen, K. M., … Munkittrick, K. R. (2014). An inter-laboratory study on the variability in measured concentrations of 17 beta-estradiol, testosterone, and 11-ketotestosterone in white sucker: Implications and recommendations. *Environmental Toxicology and Chemistry*, *33*, 847–857.

Foo, Y. Z., Nakagawa, S., Rhodes, G., & Simmons, L. W. (2017). The effects of sex hormones on immune function: A meta-analysis. *Biological Reviews*, *92*, 551–571.

Forlano, P. M., Schlinger, B. A., & Bass, A. H. (2006). Brain aromatase: New lessons from non-mammalian model systems. *Frontiers in Neuroendocrinology*, *27*, 247–274.

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2016). Detecting and avoiding likely false-positive findings – A practical guide. *Biological Reviews*.

Gail, M. H., Fears, T. R., Hoover, R. N., Chandler, D. W., Donaldson, J. L., Hyer, M. B., … Ziegler, R. G. (1996). Reproducibility studies and interlaboratory concordance for assays of serum hormone levels: Estrone, estradiol, estrone sulfate, and progesterone. *Cancer Epidemiology Biomarkers & Prevention*, *5*, 835–844.

Garamszegi, L. Z., Eens, M., Hurtrez-Boussès, S., & Møller, A. P. (2005). Testosterone, testes size, and mating success in birds: A comparative study. *Hormones and Behavior*, *47*, 389–409.

Garamszegi, L. Z., Hirschenhauser, K., Bokony, V., Eens, M., Hurtrez-Bousses, S., Moller, A. P., … Wingfield, J. C. (2008). Latitudinal distribution, migration, and testosterone levels in birds. *The American Naturalist*, *172*, 533–546.

Garde, A. H., Hansen, Å. M., & Nikolajsen, T. B. (2003). An inter-laboratory comparison for determination of cortisol in saliva. *Accreditation and quality assurance*, *8*, 16–20.

Garland, T., Bennett, A. F., & Rezende, E. L. (2005). Phylogenetic approaches in comparative physiology. *Journal of Experimental Biology*, *208*, 3015–3035.

Goymann, W. (2009). Social modulation of androgens in male birds. *General and Comparative Endocrinology*, *163*, 149–157.

Goymann, W., Landys, M. M., & Wingfield, J. C. (2007). Distinguishing seasonal androgen responses from male-male androgen responsiveness - revisiting the challenge hypothesis. *Hormones and Behavior*, *51*, 463–476.

Goymann, W., Moore, I. T., Scheuerlein, A., Hirschenhauser, K., Grafen, A., & Wingfield, J. (2004). Testosterone in tropical birds: Effects of environmental and social factors. *The American Naturalist*, *164*, 327–334.

Goymann, W., & Wingfield, J. C. (2014). Male-to-female testosterone ratios, dimorphism, and life history – What does it really tell us? *Behavioral Ecology*, *25*:685–699.

Hau, M., Ricklefs, R. E., Wikelski, M., Lee, K. A., & Brawn, J. D. (2010). Corticosterone, testosterone and life-history strategies of birds. *Proceedings of the Royal Society B: Biological Sciences*, *277*, 3203–3212.

Hirschenhauser, K., & Oliveira, R. F. (2006). Social modulation of androgens in male vertebrates: Meta-analyses of the challenge hypothesis. *Animal Behaviour*, *71*, 265–277.

Hirschenhauser, K., Winkler, H., & Oliveira, R. F. (2003). Comparative analysis of male androgen responsiveness to social environment in birds: The effects of mating system and paternal incubation. *Hormones and Behavior*, *43*, 508–519.

Jessop, T. S., Lane, M. L., Teasdale, L., Stuart-Fox, D., Wilson, R. S., Careau, V., & Moore, I. T. (2016). Multiscale evaluation of thermal dependence in the glucocorticoid response of vertebrates. *The American Naturalist*, *188*, 342–356.

Jessop, T. S., Woodford, R., & Symonds, M. R. E. (2013). Macrostress: Do large-scale ecological patterns exist in the glucocorticoid stress response of vertebrates? *Functional Ecology*, *27*, 120–130.

Lendvai, Á. Z., Bókony, V., Angelier, F., Chastel, O., & Sol, D. (2013). *Do smart birds stress less? An interspecific relationship between brain size and corticosterone levels* (p. 280). Biological Sciences: Proceedings of the Royal Society B.

Moore, F. R., Shuker, D. M., & Dougherty, L. (2016). Stress and sexual signaling: A systematic review and meta-analysis. *Behavioral Ecology.*, *27*, 363–371.

Myers, G. L. (2008). Introduction to standardization of laboratory results. *Steroids*, *73*, 1293–1296.

Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*, *13*, 88.

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

Oliveira, R. F., Hirschenhauser, K., Carneiro, L. A., & Canario, A. V. M. (2002). Social modulation of androgen levels in male teleost fish. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *132*, 203–215.

Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., … Nakagawa, S. (2016). Transparency in ecology and evolution: Real problems, real solutions. *Trends in Ecology & Evolution*, *31*, 711–719.

Roberts, M. L., Buchanan, K. L., & Evans, M. R. (2004). Testing the immunocompetence handicap hypothesis: A review of the evidence. *Animal Behaviour*, *68*, 227–239.

Romero, L. M. (2002). Seasonal changes in plasma glucocorticoid concentrations in free-living vertebrates. *General and Comparative Endocrinology*, *128*, 1–24.

Swanson, E. M., & Dantzer, B. (2014). *Insulin-like growth factor-1 is associated with life-history variation across Mammalia* (p. 281). Biological Sciences: Proceedings of the Royal Society B.

Thornton, J. W., Need, E., & Crews, D. (2003). Resurrecting the ancestral steroid receptor: Ancient origin of estrogen signaling. *Science*, *301*, 1714–1717.

Vesper, H. W., Botelho, J. C., Shacklady, C., Smith, A., & Myers, G. L. (2008). CDC project on standardizing steroid hormone measurements. *Steroids*, *73*, 1286–1292.

Wingfield, J. C., Hegner, R. E., Dufty, A. M., & Ball, G. F. (1990). The “Challenge Hypothesis”: Theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. *The American Naturalist*, *136*, 829–846.

Zera, A. J., Harshman, L. G., & Williams, T. D. (2007). Evolutionary endocrinology: The developing synthesis between endocrinology and evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*, *38*, 793–817.

## SUPPORTING INFORMATION

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

**How to cite this article:** Fanson KV, Németh Z, Ramenofsky M, Wingfield JC, Buchanan KL. Inter-laboratory variation in corticosterone measurement: Implications for comparative ecological and evolutionary studies. *Methods Ecol Evol*. 2017;8:1745–1754. https://doi.org/10.1111/2041-210X.12821