General and Comparative Endocrinology 270 (2019) 1–9

Contents lists available at ScienceDirect

General and Comparative Endocrinology

journal homepage: www.elsevier.com/locate/ygcen

Profile repeatability: A new method for evaluating repeatability of individual hormone response profiles

J. Michael Reed*, David R. Harris[1], L. Michael Romero

Department of Biology, Tufts University, Medford, MA, USA

ARTICLE INFO

Keywords:
Boldness
Behavioral phenotype
Personality
Reaction norm
Response pattern
Stress response

ABSTRACT

There is broad interest in determining repeatability of individual responses. Current methods calculate repeatability of individual points (initial and/or peak), time to peak value, or a single measure of the integrated total response (area under the curve), rather than the shape of the response profile. Repeatability estimates of response profiles using linear mixed models (LMM) generate an average repeatability for an aggregate of individuals, rather than an estimate of individual repeatability. Here we use a novel *ad hoc* method to calculate repeatability of individual response profiles and demonstrate the need for a more rigorous assessment protocol. Response profile repeatability has not been defined at the individual level. We do this using a new metric, Profile Repeatability (PR), which incorporates components of variance and the degree to which response profiles cross each other in a time series. Values range from 0 (no repeatability) to 1 (complete repeatability). We created synthetic data to represent a range of apparent time series repeatability, and 20 independent observers visually ranked those data sets by degree of repeatability. We also applied the method to real data on stress responses of European starlings *Sturnus vulgaris*. We then computed PR scores for the synthetic data and for real data from European starling corticosterone responses over time, and contrast the results to those from LMM. Finally, we assessed the sensitivity of PR to reductions in the number of time points in the corticosterone response, as well as reductions in the number of replicates per individual. We found the average PR scores for a group of individuals to be somewhat robust to reductions in points in the time series; however, the ranks of individuals (PR values relative to one another) could change substantially with reduction in the number of values in a time series. PR showed threshold sensitivity to losing replicate time series between 6 and 4 replicates. Surprisingly, human observers fell into two disparate groups when ranking repeatability of the synthetic data, and the PR score indicated that human observers may underestimate repeatability of data where replicates cross each other. In contrast to the average profile repeatability estimated using LMMs, our approach calculates individual repeatability. From our perspective, LMM does not provide a definitive idea of repeatability at the individual level; in essence, it concludes that suites of time series with low within-individual variance has high repeatability, regardless of replicate trajectories. LMM and PR have non-linear relationships between 0 and 1, but PR has greater discrimination for mid-values of repeatability. Consistent average group repeatability can be associated with substantial differences in individual ranks suggests that estimating individual repeatability is critical. The PR score should be useful in comparing repeatability of any type of nonlinear, including non-monotonic, response profiles over time, which are common in both physiology and behavior, and it demonstrates the specific needs for future improvements of a profile repeatability metric.

1. Introduction

One foundation of evolution is trait heritability. To quantify heritability of labile traits, investigators must have an idea of consistency or repeatability of their measurements of the trait under study. Labile traits, such as behavioral and physiological responses, change over time. Consequently, there has been a burgeoning interest in how to assess consistency or repeatability of labile traits. Repeatability is thus becoming an important metric for studies in evolution and ecology (van Oers et al., 2004). Examples of using repeatability metrics on labile traits include studies of mating behavior (Bell et al., 2009; Boake, 1989), parental care (Schwagmeyer and Mock 2003), exploratory

* Corresponding author.
E-mail address: michael.reed@tufts.edu (J.M. Reed).
[1] Present address: 115 Colburn Road, Milford, NH 03055, USA.

https://doi.org/10.1016/j.ygcen.2018.09.015
Received 1 June 2018; Received in revised form 21 September 2018; Accepted 25 September 2018
Available online 28 September 2018

behavior (Dingemanse et al., 2002), animal temperament (Reale et al., 2007), boldness or novelty seeking (Carere and van Oers, 2004; Fidler et al., 2007), animal personality (Carter et al., 2013; Smith and Blumstein, 2007), and natal dispersal (Dingemanse et al., 2003).

Physiological responses are also labile traits in which demonstrating repeatability is important to understanding species evolution and ecology (Hau et al., 2016; Romero and Wingfield, 2016). A number of studies have used a repeatability metric to compare physiological responses, including studies of metabolic rate (Biro and Stamps, 2010; Holtmann et al., 2017; Woods et al., 2010) and hormone titers (Hau et al., 2016; Koolhaas et al., 1999). The most common repeatability metric is an index popularized by Lessells and Boag (1987). This metric has been used to compare single measurements, such as initial or peak hormone titers (e.g,. Romero and Reed, 2008) and total amount of secreted hormone over time (Cockrem 2013), as well as comparing multiple values of the same measurement, such as flight initiation distance (Keyel et al., 2012). Many labile traits, however, consist of a response profile, i.e., the pattern of hormone secretion, such as the change in glucocorticoid hormone titers during a stress response, that might not be monotonic (Hau and Goymann, 2015; Romero and Wingfield, 2016).

There are five main aspects of a stress response: the initial (or baseline) titers, the peak titers, the time to reach peak titers, the total amount secreted over a set time period, and the shape of the overall response. Until recently, attempts to assess repeatability of glucocorticoid responses, i.e., the stress response, have analyzed either each time point separately or a single integrated measure of total glucocorticoid output over time (e.g. Cockrem, 2013; Cockrem and Silverin, 2002; Cook et al., 2011). Recent data, however, suggest that it is the shape, i.e., profile, of glucocorticoid release over time that is important for fitness, not single points along that response profile (e.g. Gesquiere et al., 2011; Jones et al., 2016). For example, the strength of negative feedback, not initial or maximal glucocorticoid titers, predicted which Galapagos marine iguanas *Amblyrhynchus cristatus* were more likely to survive famine (Romero and Wikelski, 2010). A growing number of studies have assessed repeatability of glucocorticoid titers as a way to address the evolution of hormone systems (Baugh et al., 2014; Cockrem et al., 2009; Narayan et al., 2013; Ouyang et al., 2011; Small and Schoech, 2015; Sparkman et al., 2014; Wada and Breuner, 2008). In general, the response profile itself is ignored; instead, single time periods are assessed, such as multiple samples of baseline glucocorticoids being evaluated for repeatability. Given the importance of the shape of the response for survival (including negative feedback), it is important to determine whether that entire response was repeatable within a single animal.

One might imagine that the area under the curve of a glucocorticoid response might provide a metric that one could use to determine response profile repeatability. However, there are many different response profiles that could result in exactly the same area under the curve. For example, a rapid increase (e.g., slope = 1), and rapid decline over time (e.g., slope = −1) have the same area under the curve, and are yet entirely different responses. Furthermore, they could have the same area under the curve as a flat, moderate response – but the three profiles have low repeatability. Consequently, area under the curve and response profile represent two different aspects of the stress response.

One approach that has recently been proposed for repeatability of repeated samples (but not the response profile itself) is an expanded version of Lessels and Boag's that uses linear mixed models (LMM) to decompose variance components of repeated time samples (Baugh et al., 2014; Dingemanse and Dochtermann, 2013; Hau et al., 2016). Several approaches are possible. For example, one could use values at each of several time points as fixed effects, individual clusters as random effects, and rank individuals using y-intercept order. Rather than evaluate the response profile itself, this method examines the components of variation at each time point. Another approach is to define repeatability using proportions of variance between and among individuals to estimate repeatability of the group, but this approach cannot estimate repeatability within an individual. Consequently, this approach is an analysis of variance rather than an analysis of the response profile itself. Although this approach is interesting, we believe it is insufficient to capture the nuances of repeatability; in particular, it captures the spread of values but not the profiles of responses. We can level similar criticisms about functional data analysis (e.g. Hegarty et al. 2016). In addition, the LMM approach also requires complete data records; incomplete records are often dropped from analyses, or filled via interpolation, which will inflate repeatability. One offered solution to capturing individual response profiles that uses mixed models is to evaluate the residuals from the average response (i.e., residual intra-individual variability) (Araya-Ajoy et al., 2015). We think this approach is also insufficient because the results will change with the average group response, and the approach still treats points in a time line individually rather than in the context of a response profile. To understand the response profile itself using this approach, we think the residuals would need to be understood within the context of the individual's sequence of residuals. Finally, a promising avenue for estimated individual response profiles has been proposed by Cleasby et al., (2015) using double-hierarchical generalized linear models (DHGLM) (see applications by Mitchell et al., 2016; Mitchell and Biro, 2017). We note, however, that the method is data hungry; under very good conditions, 20 replicates from 20 individuals will suffice for a stable estimate of repeatability, but often larger samples are needed (Cleasby et al., 2015). This far exceeds the amount of data gathered in almost every study of individual repeatability, particularly for physiological responses.

Our goal was to explore what is actually meant by response profile repeatability, and to present an *ad hoc* statistical method for assessing repeatability of a response profile, including those that are not monotonic. *Ad hoc* approaches are not ideal because of their potential lack of generality, but we were unable to find a published statistical approach that described individual response profile reliability to our satisfaction. Our approach was to focus both on variances within each time period (e.g. baseline titers) and in rank consistency across time periods (i.e. how often responses cross one another). Key features of our metric that are not present in LMM is that it takes into account response profiles that are not monotonic (for example, a zig-zag profile) and that cross one another, both of which are common in physiological responses. We also note that the DHGLM approach of Cleasby et al. (2015) does not address how to select meaningful fixed and random effect covariates; in this paper our evaluations will contribute to addressing this. We used the example of glucocorticoid responses over time during a stress response to create a statistical test. As part of our evaluations, we determine the sensitivity of profile repeatability to the number of points in the profile (time series), the number of replicates, and to missing data. Finally, we compared our approach with the LMM approach described above. We hope that introduction of our deliberately *ad hoc* approach will spark interest in statisticians to create a more general solution that addresses all aspects of repeatability presented here.

2. Methods

2.1. Profile repeatability

Glucocorticoids typically occur at baseline concentrations, increase in response to a stressor, and decrease in response to negative feedback, thereby creating a nonlinear series of glucocorticoid titers over time (henceforth, a Cort Series) (Romero, 2004). Given replicate Cort Series from the same individual, we want to determine the degree to which values and profiles (i.e., shape of the response) are repeatable with replication. To do this, we created a new metric, Profile Repeatability (PR), that ranges between 0 (no repeatability) to 1 (perfect repeatability). Repeatability can be somewhat subjective, and we created this metric (PR) from first principles of what we consider to be profile

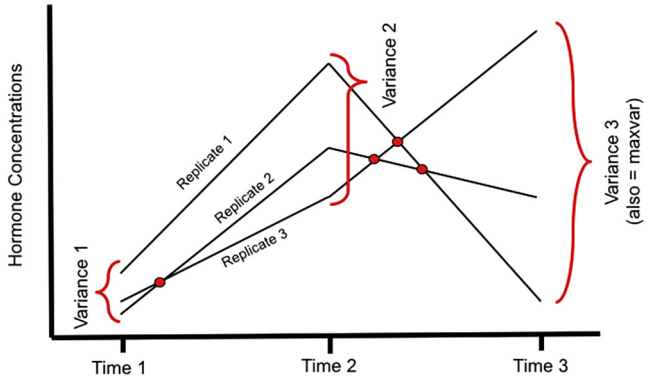

Fig. 1. Schematic of a single individual with three Cort Series replicates, each containing a three-point time series. Also depicted are the components used to compute PR (Profile Repeatability): variance at each time point, which are averaged; the maximum variance; and the number of line crossings (depicted by red dots), which contributes to the line-crossing score. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

repeatability. We created our PR metric to reflect variance across replicates and consistency over time of the rank order of the replicates.

First, if results are consistent across replicates, we assume variance of replicate values at each time point in the Cort Series will be small; i.e. small variance = more repeatable. Hence, we calculated sample variance in glucocorticoid value across replicates at each time point of the Cort Series for an individual, recorded the maximum of these variances (*maxvar*), and calculated the average of the variances across time periods (*avevar*). Second, we included a factor based on the interlacing of Cort Series, i.e. a function of the frequency with which the trajectories of each replicate of an individual Cort Series cross other replicate lines. For a visual, see Fig. 1. Specifically, we created a metric for use in the repeatability statistic based on the number of actual crossovers (*ncrossovers*) relative to the total number of possible crossovers (*npotential*). We used this approach to scale for the dependence of the number of crossovers upon the number of replicates and time periods in the Cort Series. We then compute

$$crossings = \text{int}(10 * ncrossovers / npotential)$$

to give us an integer between 0 and 10. *Int* refers to truncating to the nearest integer in the calculation, and

$$npotential = [nrep * (nrep - -1) * (ntime - -1)]/2,$$

where *nrep* is the number of replicates and *ntime* is the number of time points in the Cort Series. The final step in determining the crossover score is to scale this value so that it is commensurate with the maximum variance and average variance components of the consistency score. This gives us the final *crossover_score:*

$$crossover\_score = maxvar * crossings/5$$

In this equation, 5 is a scaling factor that we chose so that if half the crossings that could occur do occur, then the crossover score would be weighted the same as *maxvar* (because *crossings*/5 = 1). One can change relative weights by changing this value.

Next, the three values are combined into a *Base* value for each individual:

$$Base = \left(\frac{maxvar + avevar + crossover\_score}{100}\right) - 5$$

The '−5' spreads out values for PR when *Base* is very small, and it has little effect when *Base* is large. To normalize PR scores between 0 and 1, we convert the *Base* value using a logit transformation:

$$PR = 1 - \left(\frac{1}{1 + e^{-Base}}\right)$$

When *Base* is small (i.e., < 5 0 0) the '−5' correction results in a sigmoid value less than 0.5, and hence a PR metric greater than 0.5. Similarly, a large *Base* gives PR < 0.5.

If there are missing values in any of the Cort Series, PR can be modified. With missing values, there is less potential for crossovers, but the effect on variance is unpredictable. Deleting a Cort Series with missing data could lead to an inadequate set of replicates. There are two kinds of problems: missing values within a Cort Series, and an entire missing Cort Series (replicate) for an individual. One way to correct for missing values is through imputation, but this can artificially reduce variance and could decrease the number of crossovers in our replicates. Consequently, we penalrf our *Base* calculation: for each missing value, we added 0.5 to *ncrossovers*, and for a missing replicate we added 2.0 to *ncrossovers*. See Appendix S1 for R code to calculate PR, including correction for missing values.

We then evaluated the performance of PR using three data sets, one synthetic data set where we had control over the degree of repeatability, and two were data from a published study on European starlings *Sturnus vulgaris* (Romero and Remage-Healey, 2000).

2.2. Analysis of synthetic data

The synthetic data were 4 replicates per individual of a 4-point Cort series for each of 11 individuals (Fig. 2; Appendix S2). The first internal check was to rank the degree of repeatability across the individuals by eye. We did this using 20 people ranking the graphs independently; all rankers had training in biology, but they differed in experience (undergraduates, graduate students, research faculty, and a physician). The only instruction rankers were given was to rank the individuals by degree of repeatability, and that ties (assigning multiple graphs to equal ranks) were allowed. We then calculated PR for each of the synthetic individuals, and compared PR ranks to those done by eye. We wanted to determine the effect of reducing the number of samples in the Cort Series on estimated repeatability because many stress studies collect different numbers of samples within each Cort Series (typically 2–4). Consequently, we dropped the fourth sample points and recalculated PR, and then dropped the second sample points, allowing us to compare 4 vs. 3 vs. 2 sample points per replicate. We used Kendall's tau for between-group comparisons of PR consistency.

2.3. Analysis of starling data

European starling data came from captive individuals held under three conditions – short days, long days, and short days during molt (hereafter molt). Each condition had 18 birds (9 males and 9 females), but not every bird was sampled in every condition. Consequently, the data set represents 24 starlings with data for at least 2 conditions and one under only one condition (Appendices S3 and S4). Each bird was sampled at four time points: under 3 min (referred to as time "0", Romero and Reed, 2005), 15, 30, and 45 min. This Cort Series (4 sample points) is one replicate. The Cort Series was then repeated at 4 times of the day (02:00, 08:00, 14:00, 20:00) for a total of 4 replicates per bird per condition.

We first determined PR for each of 18 individuals during molt; this included 286 corticosterone (the avian glucocorticoid) measurements (2 missing values). We also determined the effect of reducing the number of samples within the Cort Series on PR consistency by dropping the 45-min sample points and recalculating PR, and then dropped the 15-min sample points. This allowed us to compare 4 vs. 3 vs. 2 sample points in the Cort Series (maintaining 4 replicates per individual).

Finally, using short-day and long-day conditions, we determined the effect of number of replicates per individual on PR estimates. For this, we analyzed 11 individuals for which we had 8 replicates each, 4 from short-days and 4 from long-days, for the same Cort Series described above (0, 15, 30, 45 min) yielding 348 corticosterone measurements (4

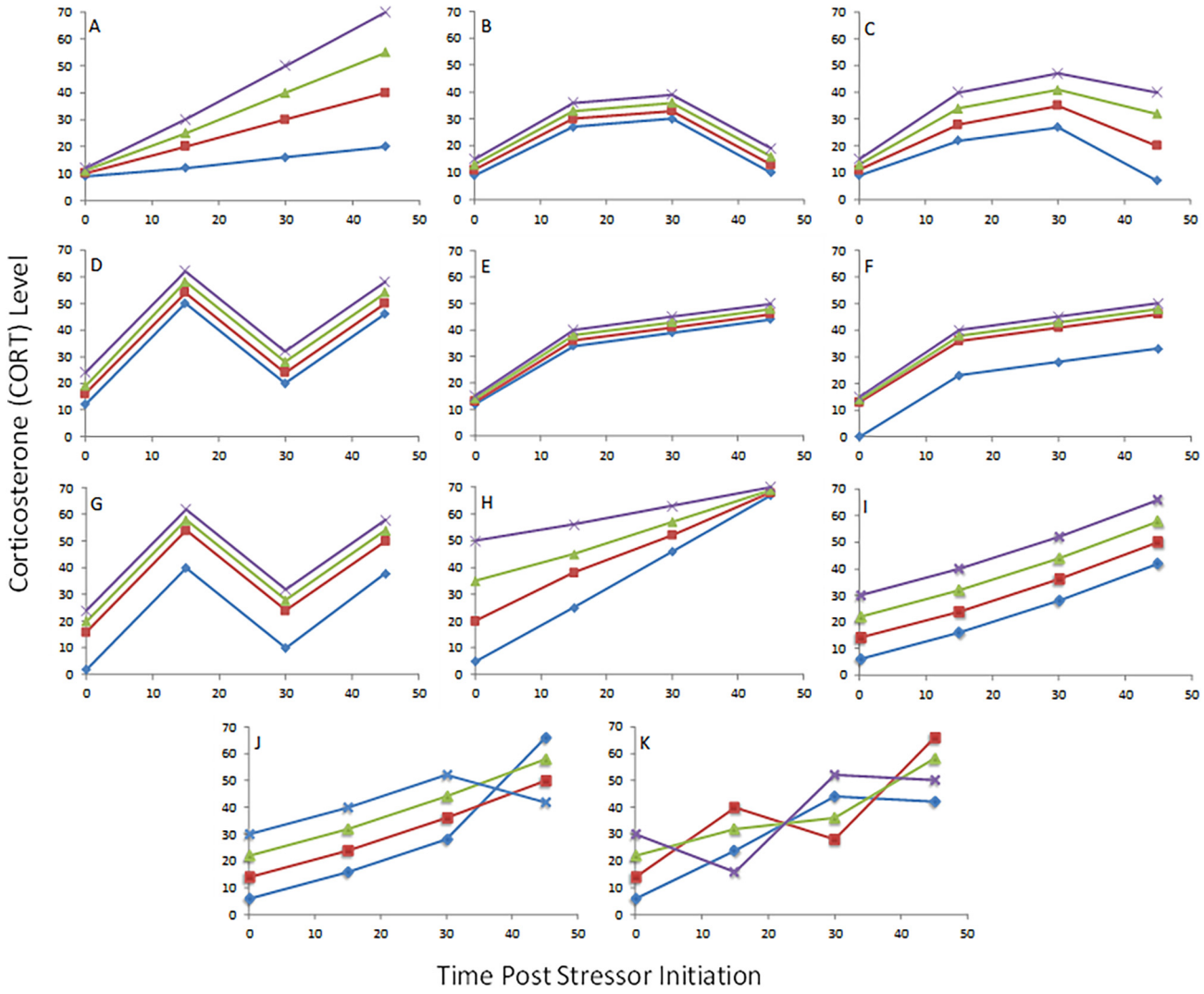

Fig. 2. A set of synthetic data where each graph represents four replicates of a four-time point Cort Series. Graphs were constructed to depict a range of possible relationships that would differ in their perceived repeatability.

Table 1
Synthetic Data ranked by the majority and minority of evaluators (by eye), and by Profile Repeatability (PR) score; sorted by Majority assessment.

Graph ID from Fig. 2	Majority Rank (n = 15)	Minority Rank (n = 5)	PR Score	PR Rank
E	1	8 = 9	0.992	1
B	2	2	0.991	2
D	3	6	0.989	3
F	4	4 = 5	0.979	4
G	5	3	0.961	5
I	6	1	0.946	6
C	7	4 = 5	0.886	8
H	8	7	0.437	10
A	9	8 = 9	0.199	11
J	10	10	0.903	7
K	11	11	0.761	9

Ties indicated by equated ranks (e.g. 8 = 9 means graphs A and E were ranked equivalently as the 8th and 9th most repeatable).

missing values). We calculated PR for 8, 6, 4, and 2 replicates by randomly removing one long-day and one short-day replicate for each replicate reduction.

2.4. Comparing PR to LMM

Because LMM (linear mixed models) is the current standard for analyzing repeatability, we compared the relationships between repeatability metrics. For LMM, we estimated repeatability for a range of values for between- and within-individual variance. For PR (profile repeatability), we used a range of values for average variance and crossover score. The respective range of values was selected to show the range of possible outcomes for repeatability from near zero to near 1 (the maximum range of both methods). The particular LMM used came from Baugh et al. (2014, p. 158).

3. Results

3.1. Synthetic data

The 20 people who ranked the synthetic data assorted the individuals figures in two distinct fashions. The majority (n = 15) gave highly consistent within-group ranks ($r^2 = 0.88$ when plotting the average rank from all 15 people vs. the rank assignments from each individual person from Fig. 2; Appendix S5) with Graph E deemed the most repeatable (Table 1). The other 5 people gave a different overall ranking that was also consistent ($r^2 = 0.56$; see Appendix S5). For

Table 2
Profile Repeatability (PR) scores and ranks from the synthetic data.

Graph ID (Fig. 2)	PR Score (Profile Repeatability)			PR Rank		
	4 Time Points	3 Time Points	2 Time Points	4 Time Points	3 Time Points	2 Time Points
E	0.992	0.992	0.993	1	1	1
B	0.991	0.991	0.991	2	2	2
D	0.989	0.989	0.987	3	3	3
F	0.979	0.979	0.980	4	4	4 = 5
G	0.961	0.959	0.959	5	6	6
I	0.946	0.946	0.946	6	7 = 8	7 = 8
J	0.902	0.946	0.946	7	7 = 8	7 = 8
C	0.886	0.978	0.980	8	5	4 = 5
K	0.761	0.761	0.903	9	10	9
H	0.437	0.322	0.290	10	11	11
A	0.199	0.879	0.861	11	9	10

Graph IDs are sorted by 4-time point scores (Note: PR Score with 4 time points is repeated from Table 1). PR Scores were then recalculated, and reranked, with reduction of Time Points. Ties are indicated by equated ranks (e.g. 4 = 5 means graphs C and F were ranked equivalently as the 4th and 5th most repeatable).

example, they ranked Graph I as the most repeatable and Graph E as not particularly repeatable with a rank of 8.5. As a consequence, the two group's rankings were poorly correlated with each other ($r^2 = 0.3$). The PR score was high for most of the synthetic individuals (Table 1). However, the PR score was able to successfully distinguish between the 11 individuals, albeit sometimes only at the third decimal point. When the PR score was converted to rank, the PR rank order was identical to the majority rank for the first 6 individuals (Table 1). The PR and majority ranks differed for the 5 lowest ranked individuals.

The PR score was relatively insensitive to the number of samples (time points) in the Cort Series (Table 2). The PR mean score increased slightly from 0.822 to 0.894 as the number of samples in the Cort Series was reduced, but the overall range decreased. PR ranks were fairly consistent (Table 2), regardless of the number of samples in the Cort Series (Kendall Tau: 4 vs 3 samples = 0.81; 4 vs 2 samples = 0.83; 3 vs. 2 samples = 0.95).

3.2. Starling data

3.2.1. Reduced number of time points

PR scores of the 18 molting starlings with 4 time points (4 replicates per bird), ranged from 0.980 to 0.474 (Table 3), with an average of 0.841 ± 0.155. As the number of time points was reduced, mean PR increased: for 3 time points, mean PR increased to 0.886 ± 0.137; for 2 points mean PR was 0.905 ± 0.151. Although the PR increased with decreasing time points, the standard deviation remained relatively constant. Fig. 3 depicts four examples of individual birds, representing those with high, medium, and low relative PR scores, and how those ranks change as the number of time points is reduced. Importantly, the reduction in time points had a strong effect on the relative ranks of bird PR. Although there was a tendency for low PRs to stay low and high PRs to stay high (Table 3), the correlation between PR for 4 vs. 3 time points was 0.761. Going to 2 time points, however, drastically changed the rank orders; the correlation between PR scores was 0.420 for 4 vs. 2 time points, and 0.454 for 3 vs. 2 time points.

3.2.2. Reduction in number of individual Cort Series (replicates)

Similar to reducing the number of time points, reducing the number of replicate Cort Series had little effect on mean PR scores (Table 4). When all 8 Cort Series were included, mean PR score was 0.419 ± 0.389 (range 0.960–0.000). Fig. 4 shows three examples of individual birds representing those with high, medium, and low relative PR scores. We included a graph of the second-to-lowest ranked individual rather than the lowest ranked individual because inspection of the lowest ranked individual indicated its PR score, and therefore its rank, was heavily influenced by a single Cort Series (suggesting the replicate was an outlier), that by random selection remained throughout the reduction process. Reduction to 6 to 4 to 2 Cort Series resulted in means of 0.402 ± 0.392, 0.376 ± 0.410, and 0.392 ± 0.436, respectively. In contrast, the medians changed notably for the same reduction in the number of Cort Series: 0.437, 0.436, 0.149, 0.158, for 8, 6, 4, 2, respectively, suggesting a threshold change in the underlying distribution of the data. Similar to reducing the number of time points, reducing the number of Cort Series resulted in notable changes in individual rankings (Table 4; Fig. 4). When reducing from 8 to 6 Cort Series, the correlation (Kendall's Tau) was 0.734, reducing from 6 to 4 was 0.755, and reducing from 4 to 2 was 0.641. Consequently, going all the way from 8 to 2 Cort Series resulted in a correlation of only 0.342.

Table 3
Profile Repeatability (PR) scores and ranks from molting starlings, sorted by 4-time point scores; profiles in time point reduction depicted in Fig. 3 for individuals marked in bold.

PR Score			PR Rank		
4 Time Points	3 Points	2 Points	4 Time Points	3 Points	2 Points
0.980	0.977	0.972	**1**	**4**	**8**
0.975	0.988	0.990	**2**	**1**	**1**
0.962	0.982	0.987	3	3	4
0.961	0.963	0.976	4	5	7
0.958	0.955	0.944	5	6	10
0.947	0.953	0.925	6	7	13
0.946	0.985	0.985	7	2	5
0.932	0.940	0.988	**8**	**10**	**3**
0.917	0.925	0.966	9	11	9
0.913	0.953	0.936	10	8	11
0.900	0.909	0.879	11	12	15
0.875	0.903	0.746	12	13	17
0.753	0.843	0.990	13	15	2
0.742	0.945	0.927	14	9	12
0.694	0.743	0.982	15	16	6
0.632	0.576	0.841	16	17	16
0.584	0.519	0.358	17	18	18
0.474	0.895	0.896	**18**	**14**	**14**

3.3. Comparing PR to LMM

As can be seen in Fig. 5, the PR metric has its greatest discrimination ability (i.e., steepest rate of change) in the middle of the distribution, while the LMM metric has its greatest discrimination at high values of repeatability.

4. Discussion

4.1. Synthetic data

The existence of two distinct ranking orders from the panel of 20 people for the synthetic data was unexpected. We anticipated that everyone would be highly consistent in their rankings, and this was the case for 75% of the panelists. In discussions with the panelists, it became clear that there were differences in how to assess profile repeatability. The majority gave weight to consistency of values at each time point (Graph E was rated highest; graphs refer to Fig. 1), whereas the minority gave consistency of the overall response shape (i.e., small differences in slope, also sometimes referred to as a reaction norm) greater weight than variances (Graph I was rated highest). Rankings were consistent within each group, suggesting that there are two distinct ways of judging response profile repeatability; that is, there appear to be two major components to profile repeatability. It is unclear how to address this issue, but it points to a potential problem if researchers do not agree on what is meant by 'response profile repeatability'. We

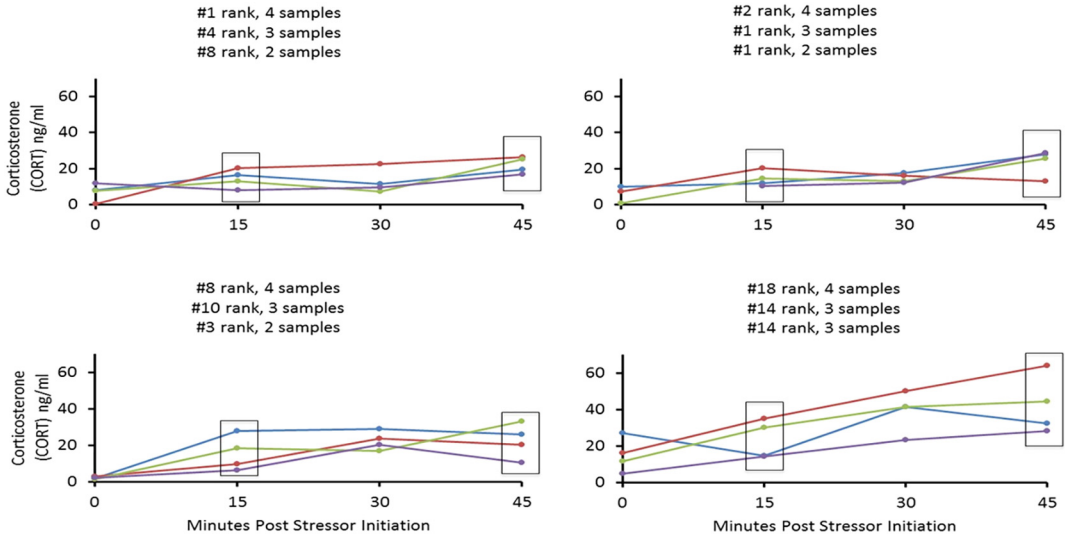

Fig. 3. Examples of individual molting starlings representing those with high, medium, and low relative ranked PR scores and how ranks changes as the number of time points was reduced (see Table 3). Reduction was done by removing first the 45-min time points, then by removing the 15-min time points (indicated by boxes). Note that in the upper right graph, one Cort Series is missing the 0 time point.

emphasize that our PR rank was consistent with the majority rank.

The PR metric was able to distinguish between all 11 synthetic individuals. Although it is striking that the PR perfectly correlates with the ranks of the 6 most-highly ranked individuals, this was expected because the synthetic data were created specifically to provide a template for assessing the efficacy of the PR. What is more interesting is that the PR flips the order of the 5 worst ranked individuals compared to majority assessment. In fact, PR may have exposed a human bias in profile recognition. Although the multiple crossing lines made all of the human evaluators rank individuals J and K as the least repeatable, these individuals have consistent average increases. Individuals A and H, on the other hand, show very different rates of increase for each replicate, even though the lines never cross. The *maxvar* component of PR was downgraded in importance by the human evaluators but was a major factor in the PR for the lower ranking of these individuals. Individuals J and K are certainly less repeatable than those ranked more highly, but a hypothetical 5th replicate would be more predictable than would be a 5th replicate for individuals A and H. In retrospect, individuals J and K are probably more consistent than are A and H. Consequently, PR is likely to be more accurate than the "eyeball" test used by human evaluators when repeatability is low. This has interesting implications because real data are often far less consistent than these synthetic data.

In the synthetic data, there was little change in PR values and ranks as the number of samples in the Cort Series was reduced. However, this is likely due to the nature of the synthetic data. The data were intentionally constructed to emphasize certain differences and had much less variation than a natural Cort series would have. Even in this situation, however, PR scores generally decreased as more data became available.

4.2. Starling data

Similar to the synthetic data, reducing the number of time points for molting starlings caused only a slight increase in PR value. It thus appears that PR is relatively insensitive to changes in the number of time points in a Cort series. However, we are interested in interchangeability of the ranks, rather than statistical significance per se (as explained by Romero and Reed, 2008). That is, do individuals with high PR score remain of high rank compared to other individuals as the amount of information decreases? This is an important feature in anticipating the potential for evolutionary response to selective pressure. The answer is clearly no. Restricting analyses to 2 time points had a drastic impact on individual ranks, so it appears that 2 data points are insufficient to give an accurate estimate of real series repeatability. This is consistent with low estimates of rank repeatability of single-point samples in captive house sparrows *Passer domesticus* (Romero and Reed, 2008).

Table 4
Profile Repeatability (PR) scores and ranks from starlings with data on both short and long days, sorted by 8-time point scores; profiles depicted in Fig. 4 for individuals marked in bold.

PR Score				PR Rank			
8 replicates	6 replicates	4 replicates	2 replicates	8 replicates	6 replicates	4 replicates	2 replicates
0.960	0.961	0.945	0.621	**1**	**2**	**2**	**5**
0.902	0.980	0.963	0.972	2	1	1	2
0.884	0.834	0.548	0.000	3	3	5	7–11
0.663	0.436	0.149	0.888	4	6	6	3
0.522	0.499	0.048	0.158	5	5	7	6
0.437	0.099	0.743	0.993	**6**	**7**	**3**	**1**
0.110	0.077	0.004	0.000	7	8	8	7–11
0.083	0.007	0.000	0.000	8	9	9 = 10 = 11	7–11
0.047	0.534	0.734	0.678	9	4	4	4
0.001	0.000	0.000	0.000	**10**	**10 = 11**	**9 = 10 = 11**	**7–11**
0.000	0.000	0.000	0.000	11	10 = 11	9 = 10 = 11	7–11

PR Scores recalculated, and reranked, with reduction of replicates (Cort Series). Ties are indicated by equated ranks (e.g. 10 = 11 indicates individuals with equal ranks, 7–11 indicates 7 = 8 = 9 = 10 = 11).

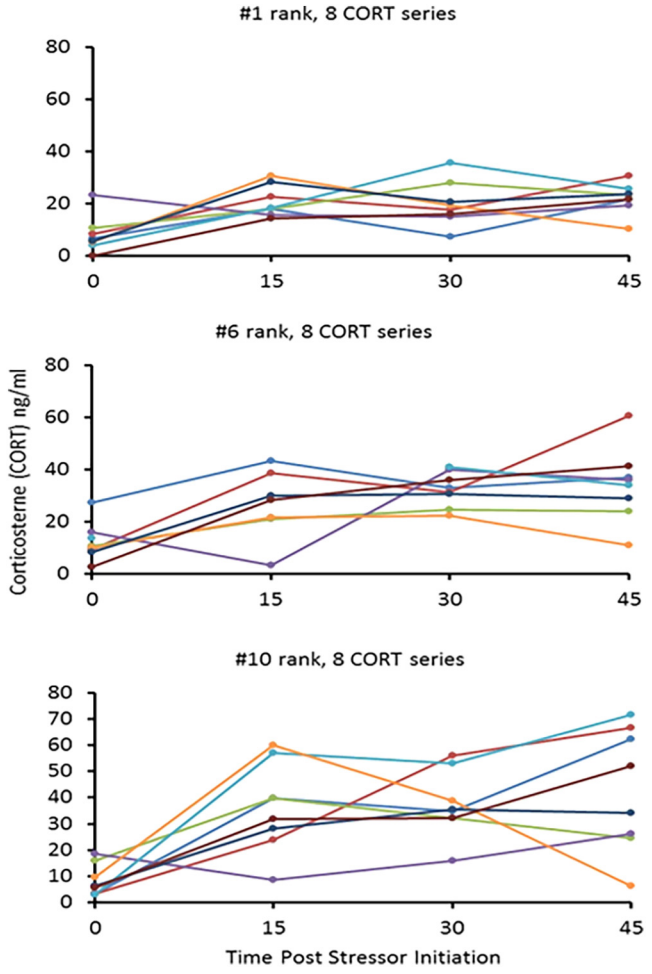

Fig. 4. Three examples of individual starling trajectories representing those with high, medium, and low relative ranked PR scores (see Table 4). Note that in the middle graph that one Cort Series is missing the 15-min time point.

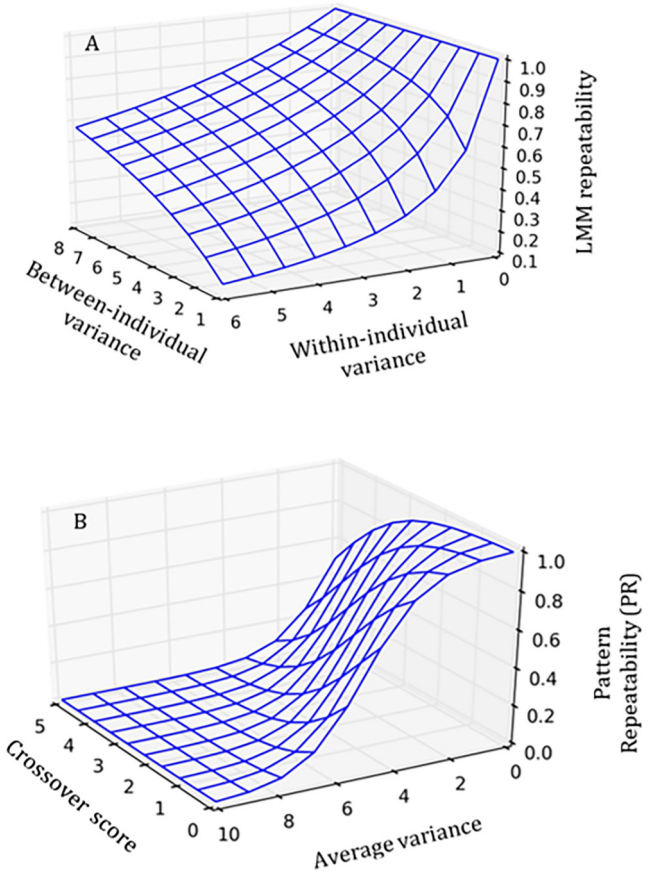

Fig. 5. Relationships between repeatability metrics (A) LMM (linear mixed models) and (B) PR (profile repeatability), for a range of component values. The particular LMM used came from Baugh et al. (2014, p. 158). 'Crossover_score' is a function of the number of times Cort Series trajectories cross one another relative to the number of times they could cross; see text for calculation.

With fewer Cort Series per individual (replicates), there was a dramatic step-wise reduction in the median PR score, with a threshold between 6 and 4 samples. This suggests that PR scores are more robust with increasing number of replicates. In addition, as with reducing the number of time points, reducing the number of Cort Series strongly altered rankings between individuals. The correlation between ranks continually decreased as the number of replicates decreased. The conclusion from this analysis is that more replicates is better. Although the median PR for the group appears to stabilize with more than 6 replicates, the individual ranks did not stabilize. Presumably there will exist a number of replicates where the rank will stabilize, but going from 6 to 8 was insufficient for this data set. In addition, the lowest ranked individual, which was heavily influenced by a single data point, indicates how sensitive the PR metric can be to outliers because of its impact on variance. We note that LMM scores would be similarly influenced by an outlier's impact on variance (see below).

Interestingly, the mean PR values were quite different between the starling data sets. The mean PR score from molting starlings with 4 Cort Series (0.841) was much higher than the mean PR score for the short- and long-day starlings with 4 Cort Series (0.376). Although part of the reason for this difference could be seasonal differences in cort responses (Romero 2002), the mean short- and long-day responses for these birds did not differ when comparing individual time points of the Cort Series (Romero and Remage-Healey, 2000). A more likely possibility is the higher repeatability of molting birds resulted from the strong down-regulation of cort responses during molt (Romero 2004). This leads to much lower variance in cort values across time and across Cort Series, therefore leading to higher PR scores. This is why comparing ranks of cort values rather than absolute cort values is more appropriate when comparing individuals across seasons (Romero and Reed, 2008).

4.3. Comparing PR to LMM

Using linear mixed models (LMM) to determine repeatability works by partitioning within- vs. between-individual variance for a suite of time series replicates, creating an average 'repeatability' value for a suite of individuals. LMM does not calculate repeatability for multiple replicates from a single individual. So the LMM approach, such as that used by Baugh et al. (2014), addresses only part of the question of repeatability. Depending on the choice of fixed effects, the CORT0 and CORT30 LMM analyses provide either an intercept adjustment based on fixed effects or simply the variance across replicates at initial or stress-induced points. The repeatability statistic presented by Baugh et al. is r = variance of intercepts / (variance of residuals + variance of intercepts),

which does not give us a definitive notion of repeatability at the individual level nor does it account for crossovers in temporal profiles among multiple replicates for an individual. In essence, it concludes that suites of time series with relatively low within-individual variance has high repeatability, regardless of replicate trajectories. From our perspective, this approach does not calculate true repeatability of individual time series because it aggregates time series rather than treating each time series independently. As an example, in our synthetic data, the LMM approach reports no difference between Graphs J and K

(Fig. 2), whereas independent biologists did distinguish between them (Table 1).

It is also likely that repeatability is context dependent. For example, in good habitats individuals might have lower variability than in poor habitats, resulting in a higher estimate of repeatability. This would derive from two features: the body's inherent ability to repeat an identical response (i.e., innate ability to respond), and environmental factors that alter the body's desired response (i.e., environmental stochasticity). The ability to determine a reaction norm for a response would depend upon both features. We note that the same relationship holds true for calculating heritability, h^2, which is a similar type of assessment (Falconer and Mackay, 1996). The PR statistic allows us to estimate both components because we can estimate the variance in responses of each individual as well as the average (and variance in) profile repeatability of a group of individuals. LMM, on the other hand, can only estimate the combined effects of both components.

An important point is that repeatability values from PR and LMM approaches are not strictly comparable. In fact, because there is no specific 'real' thing that is repeatability, there can be different metrics of a concept that can be referred to as repeatability. LMM measures repeatability as defined by the component of total variance that is within-individuals. This means that there can be more than one way to get a LMM repeatability value of, for example, 0.6, and it is akin to a percent value. PR measures repeatability as an assessment of variance plus a component of line crossing, and also can result in more than one way to get a repeatability value of 0.6, but it is not a percent. As an analogy, parasite load and tumor burden are both metrics that indicate disease, but their values are not strictly comparable – doubling of tumor size is not equivalent to doubling a parasite load. Similarly, a higher PR and higher LMM both indicate higher repeatability, but their values are not comparable (a PR of 0.8 does not mean that the data are twice as repeatable as a LMM of 0.4). Consequently, comparisons of repeatability of data sets can be made within an index, but not between indices.

We also note that although both metrics are scaled to 0 to 1, the underlying shapes of the relationships are not linear and that they are different from each other (Fig. 5). Ideally, one might wish for a repeatability index that has a simple, linear relationship between low and high values, as is provided by Lessells and Boag's (1987) single-value repeatability metric. However, this is unlikely to be available in the response profiles that are the focus of this paper. We are interested in multi-dimensional responses where there is more than a single combination of values that can give the same repeatability value. The LMM repeatability assessment (e.g. Allegue et al., 2017; Baugh et al., 2014; Hruschka et al., 2005; Westneat et al., 2011, 2014) can increase by either increasing between-subject-variance or by decreasing within-subject-variance. This yields repeatability values that are strongly non-linear across both axes (Fig. 5a). Our PR metric, in contrast can increase by decreasing variance (either maximum variance or average variance) or by decreasing the number of response trajectory crossovers. Because we scale values using a logit function, this results in PR values that are sigmoidal (Fig. 5b). This further illustrates the difficulty of comparing repeatability scores between metric types.

From our perspective, PR has three advantages compared to LMM approaches to determining response profile repeatability. First, PR calculates repeatability for an individual whereas LMM cannot; LMM can only provide average values. PR thus allows analysis at an individual level, which might be very important for understanding genetic/environmental dynamics within individual profile responses. Second, PR incorporates trajectory crosses in a series of replicates whereas LMM does not. PR can thus distinguish degrees of repeatability between individuals where LMM would record identical repeatability scores (see synthetic data graphs). Since the shape of a response (profile) and not just an intermediate or end value is important for understanding repeatability for many phenomena, the PR will provide a more accurate assessment of data repeatability than is provided by LMM for many data sets. Third, as can be seen in Fig. 5, our PR metric has its greatest discrimination ability (i.e., steepest rate of change) in the middle of the distribution, while the LMM metric has its greatest discrimination at high values of repeatability, which are uncommon in natural populations (e.g. Bell et al., 2009; Reale et al., 2007). What we hope this study prompts is the development of a more general statistic (one that is not *ad hoc*) that estimates individual response profile repeatability for all components of repeatability discussed here, including rank repeatability.

Acknowledgements

We are grateful to Durwood Marshall (formerly of Tufts University) for discussions about calculating point and profile repeatability, and we thank two anonymous reviewers for comments that improved this manuscript. We also thank the students and colleagues who ranked the synthetic data for subjective repeatability. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygcen.2018.09.015.

References

Allegue, H., Araya-Ajoy, Y.G., Dingemanse, N.J., Dochtermann, N.A., Garamszegi, L.Z., Nakagawa, S., Réale, D., Schielzeth, H., Westneat, D.F., Hadfield, J., 2017. Statistical Quantification of Individual Differences (SQuID): an educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. Methods Ecol. Evol. 8, 257–267.

Araya-Ajoy, Y.G., Mathot, K.J., Dingemanse, N.J., 2015. An approach to estimate short-term, long-term and reaction norm repeatability. Methods Ecol. Evol. 6, 1462–1473.

Baugh, A.T., van Oers, K., Dingemanse, N.J., Hau, M., 2014. Baseline and stress-induced glucocorticoid concentrations are not repeatable but covary within individual great tits (*Parus major*). Gen. Comp. Endocrinol. 208, 154–163.

Bell, A.M., Hankison, S.J., Laskowski, K.L., 2009. The repeatability of behaviour: a meta-analysis. Anim. Behav. 77, 771–783.

Biro, P.A., Stamps, J.A., 2010. Do consistent individual differences in metabolic rate promote consistent individual differences in behavior? Trends Ecol. Evol. 25, 653–659.

Boake, C.R.B., 1989. Repeatability: its role in evolutionary studies of mating behavior Evol. Ecol. 3, 173–182.

Carere, C., van Oers, K., 2004. Shy and bold great tits (*Parus major*): body temperature and breath rate in response to handling stress. Physiol. Behav. 82, 905–912.

Carter, A.J., Feeney, W.E., Marshall, H.H., Cowlishaw, G., Heinsohn, R., 2013. Animal personality: what are behavioural ecologists measuring? Biol. Rev. Cambridge Phil. Soc. 88, 465–475.

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 Ecol. Evol. 6, 27–37.

Cockrem, J.F., 2013. Individual variation in glucocorticoid stress responses in animals. Gen. Comp. Endocrinol. 181, 45–58.

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. Gen. Comp. Endocrinol. 163, 158–168.

Cockrem, J.F., Silverin, B., 2002. Variation within and between birds in corticosterone responses of great tits (*Parus major*). Gen. Comp. Endocrinol. 125, 197–206.

Cook, K.V., O'Connor, C.M., Gilmour, K.M., Cooke, S.J., 2011. The glucocorticoid stress response is repeatable between years in a wild teleost fish. J. Comp. Physiol. A Neuroethol. Sens. Neural. Behav. Physiol. 197, 1189–1196.

Dingemanse, N.J., Both, C., Drent, P.J., van Oers, K., van Noordwijk, A.J., 2002. Repeatability and heritability of exploratory behaviour in great tits from the wild. Anim. Behav. 64, 929–938.

Dingemanse, N.J., Both, C., van Noordwijk, A.J., Rutten, A.L., Drent, P.J., 2003. Natal dispersal and personalities in great tits (*Parus major*). Proc. Royal Soc. B. Biol. Sci. 270, 741–747.

Dingemanse, N.J., Dochtermann, N.A., 2013. Quantifying individual variation in behaviour: mixed-effect modelling approaches. J. Anim. Ecol. 82, 39–54.

Falconer, D.S., Mackay, T.F.C., 1996. Quantitative Genetics, 4th edn. Pearson Education Limited, Essex, UK.

Fidler, A.E., van Oers, K., Drent, P.J., Kuhn, S., Mueller, J.C., Kempenaers, B., 2007. Drd4 gene polymorphisms are associated with personality variation in a passerine bird. Proc. Royal Soc. B, Biol. Sci. 274, 1685–1691.

Gesquiere, L.R., Learn, N.H., Simao, M.C.M., Onyango, P.O., Alberts, S.C., Altmann, J.,

2011. Life at the top: rank and stress in wild male baboons. Science 333, 357–360.
Hegarty, A., Stanley, G., Kashdan, E., Hodgson, J., Parnell, A.C., 2016. Repeatability analysis of airborne electromagnetic surveys. Mathematics-in-Industry Case Studies 7, 6. https://doi.org/10.1186/s40929-016-0008-1.
Hau, M., Casagrande, S., Ouyang, J.Q., Baugh, A.T., 2016. Glucocorticoid-mediated phenotypes in vertebrates: multilevel variation and evolution. In: Naguib, M., Mitani, J.C., Simmons, L.W., Barrett, L., Healy, S., Zuk, M. (Eds.), Advances in the Study of Behavior, pp. 41–115.
Hau, M., Goymann, W., 2015. Endocrine mechanisms, behavioral phenotypes and plasticity: known relationships and open questions. Frontiers Zool. 12, 15.
Holtmann, B., Lagisz, M., Nakagawa, S., Moore, I., 2017. Metabolic rates, and not hormone levels, are a likely mediator of between-individual differences in behaviour: a meta-analysis. Funct. Ecol. 31, 685–696.
Hruschka, D.J., Kohrt, B.A., Worthman, C.M., 2005. Estimating between- and within-individual variation in cortisol levels using multilevel models. Psychoneuroendocrinol. 30, 698–714.
Jones, B.C., Bebus, S.E., Ferguson, S.M., Bateman, P.W., Schoech, S.J., 2016. The glucocorticoid response in a free-living bird predicts whether long-lasting memories fade or strengthen with time. Anim. Behav. 122, 157–168.
Keyel, A.C., Peck, D.T., Reed, J.M., 2012. No evidence for individual assortment by temperament relative to patch area or patch openness in the bobolink. Condor 114, 212–218.
Koolhaas, J.M., Korte, S.M., De Boer, S.F., Van Der Vegt, B.J., Van Reenen, C.G., Hopster, H., De Jong, I.C., Ruis, M.A.W., Blokhuis, H.J., 1999. Coping styles in animals: current status in behavior and stress-physiology. Neurosci. Biobehav. Rev. 23, 925–935.
Lessells, C.M., Boag, P.T., 1987. Unrepeatable repeatabilities: a common mistake. Auk 104, 116–121.
Mitchell, D.J., Biro, P.A., 2017. Is behavioural plasticity consistent across different environmental gradients and through time? Proc. Royal Soc. B, Biol. Sci. 284, 20170893. https://doi.org/10.1098/rspb.2017.0893.
Mitchell, D.J., Fanson, B.G., Beckmann, C., Biro, P.A., 2016. Towards powerful experimental and statistical approaches to study intraindividual variability in labile traits. Royal Soc. Open Sci. 3https://doi.org/10.1098/rsos.160352. (160352).
Narayan, E.J., Cockrem, J.F., Hero, J.M., 2013. Repeatability of baseline corticosterone and short-term corticosterone stress responses, and their correlation with testosterone and body condition in a terrestrial breeding anuran (*Platymantis vitiana*). Comp. Biochem. Physiol. A Mol. Integr. Physiol. 165, 304–312.
Ouyang, J.Q., Sharp, P., Dawson, A., Hau, M., 2011. Hormone levels predict individual differences in reproductive success in a passerine bird. Integr. Comp. Biol. 51 (E103-E103).
Reale, D., Reader, S.M., Sol, D., McDougall, P.T., Dingemanse, N.J., 2007. Integrating animal temperament within ecology and evolution. Biol. Rev. Cambridge Phil. Soc. 82, 291–318.
Romero, L.M., 2002. Seasonal changes in plasma glucocorticoid concentrations in free-living vertebrates. Gen. Comp. Endocrinol. 128, 1–24.
Romero, L.M., 2004. Physiological stress in ecology: lessons from biomedical research. Trends Ecol. Evol. 19, 249–255.
Romero, L.M., Reed, J.M., 2005. Collecting baseline corticosterone samples in the field: is under 3 min good enough? Comp. Biochem. Physiol. A Mol. Integr. Physiol. 140, 73–79.
Romero, L.M., Reed, J.M., 2008. Repeatability of baseline corticosterone concentrations. Gen. Comp. Endocrinol. 156, 27–33.
Romero, L.M., Remage-Healey, L., 2000. Daily and seasonal variation in response to stress in captive starlings (*Sturnus vulgaris*): Corticosterone. Gen. Comp. Endocrinol. 119, 52–59.
Romero, L.M., Wikelski, M., 2010. Stress physiology as a predictor of survival in Galapagos marine iguanas. Proc. Royal Soc. B. Biol. Sci. 277, 3157–3162.
Romero, L.M., Wingfield, J.C., 2016. Tempests, Poxes, Predators, and Peopld: Stress in Wild Animals and How They Cope. Oxford University Press, Oxford.
Schwagmeyer, P.L., Mock, D.W., 2003. How consistently are good parents good parents? Repeatability of parental care in the house sparrow, *Passer domesticus*. Ethology 109, 303–313.
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*). J. Compar. Physiol. B Biochem. Syst. Environ. Physiol. 185, 119–133.
Smith, B.R., Blumstein, D.T., 2007. Fitness consequences of personality: a meta-analysis. Behav. Ecol. 19, 448–455.
Sparkman, A.M., Bronikowski, A.M., Williams, S., Parsai, S., Manhart, W., Palacios, M.G., 2014. Physiological indices of stress in wild and captive garter snakes: correlations, repeatability, and ecological variation. Comp. Biochem. Physiol. A Mol. Integr. Physiol. 174, 11–17.
van Oers, K., de Jong, G., Drent, P.J., van Noordwijk, A.J., 2004. A genetic analysis of avian personality traits: Correlated, response to artificial selection. Behav. Genetics 34, 611–619.
Wada, H., Breuner, C.W., 2008. Transient elevation of corticosterone alters begging behavior and growth of white-crowned sparrow nestlings. J. Exper. Biol. 211, 1696–1703.
Westneat, D.F., Bokony, V., Burke, T., Chastel, O., Jensen, H., Kvalnes, T., Lendvai, A.Z., Liker, A., Mock, D., Schroeder, J., Schwagmeyer, P.L., Sorci, G., Stewart, I.R.K., 2014. Multiple aspects of plasticity in clutch size vary among populations of a globally distributed songbird. J. Anim. Ecol. 83, 876–887.
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. Amer. Natur. 178, 652–667.
Woods Jr., W.A., Wood, C.A., Ebersole, J., Stevenson, R.D., 2010. Metabolic rate variation over adult lifetime in the butterfly Vanessa cardui (Nymphalidae: Nymphalinae): aging, feeding, and repeatability. Physiol. Biochem. Zool. 83, 858–868.