Title
Using repeatability to study physiological and behavioural traits: ignore time-related change at your peril

Authors
Biro, Peter A
Stamps, Judy A

Publication Date
2015-07-01

DOI
10.1016/j.anbehav.2015.04.008

Peer reviewed

Commentary

Using repeatability to study physiological and behavioural traits: ignore time-related change at your peril

Peter A. Biro[a,*] Judy A. Stamps[b]

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

[b] Department of Evolution & Ecology, University of California Davis, Davis, CA, U.S.A.

Received 8 December 2014

Initial acceptance 9 January 2015

Final acceptance 12 March 2015

MS. number: 14-00991R

*Correspondence: P. A. Biro, Centre for Integrative Ecology, School of Life and Environmental Science, Deakin University, Geelong 3216, Australia.

E-mail address: pete.biro@deakin.edu.au (P. A. Biro).

Broad sense repeatability, which refers to the extent to which individual differences in trait scores are maintained over time, is of increasing interest to researchers studying behavioural or physiological traits. Broad sense repeatability is most often inferred from the statistic *R* (the intraclass correlation, or narrow sense repeatability). However, *R* ignores change over time, despite the inherent longitudinal nature of the data (repeated measures over time). Here, we begin by showing that most studies ignore time-related change when estimating broad sense repeatability, and estimate *R* with low statistical power. Given this problem, we (1) outline how and why ignoring time-related change in scores (that occur for whatever reason) can seriously affect estimates of the broad sense repeatability of behavioural or physiological traits, (2) discuss conditions in which various indices of *R* can or cannot provide reliable estimates of broad sense repeatability, and (3) provide suggestions for experimental designs for future studies. Finally, given that we already have abundant evidence that many labile traits are 'repeatable' in that broad sense (i.e. $R > 0$), we suggest a shift in focus towards obtaining robust estimates of the repeatability of behavioural and physiological traits. Given how labile these traits are, this will require greater experimental (and/or statistical) control and larger sample sizes in order to detect and quantify change over time (if present).

Keywords:

behavioural syndromes, metabolism, mixed models, personality, plasticity,

A major challenge in studying and describing behavioural and physiological traits is their lability. In contrast to morphological traits, physiology and behaviour are labile traits that can change over short periods (e.g. seconds to days) in response to changes in internal and external stimuli (Wolak, Fairbairn, & Paulsen, 2012). High lability implies that

individual differences in behavioural or physiological traits observed at one point in time might not be observed if the same set of individuals were observed again on one or more occasions, even under highly controlled conditions.

Various terms, including repeatability, differential consistency and differential stability have been used by biologists and psychologists to refer to the extent to which individual differences in behavioural or physiological scores are maintained over time (Alison M. Bell, Hankison, & Laskowski, 2009; Caspi & Roberts, 2001; Hayes & Jenkins, 1997; Roberts, Caspi, & Moffitt, 2001; Stamps & Groothuis, 2010). However, the term 'repeatability' also refers to a statistic, *R*, which has traditionally been used in quantitative genetics to estimate the proportion of trait variation that is attributed to individual differences (see equation 1; Hayes & Jenkins, 1997; Lessells & Boag, 1987b; McGraw & Wong, 1996; Nakagawa & Schielzeth, 2010; Wolak, et al., 2012). Because of the potential confusion over the two meanings of the term repeatability, here we use 'broad sense repeatability' to refer to the extent to which individual differences in scores are maintained over time (in a given context) and 'narrow sense repeatability' to refer to *R*. Importantly, although *R* can sometimes provide reasonable estimates of broad sense repeatability, this is not always the case. As we discuss below, *R* makes no implicit inferences about time-related change (there is no term for time in its formulation). Thus, if our longitudinal data contain individual or mean level changes over time not accounted for in the underlying statistical model, then inferences about broad sense repeatability will not be correct because model assumptions are violated.

Broad sense repeatability is of interest in many areas of research because it indicates that a given type of behaviour or physiology can be considered to be a characteristic of an individual (i.e. a trait), and may reflect heritability (e.g. Falconer, 1981)but see (Dohm, 2002). Recently, broad sense repeatability has attracted considerable interest from researchers interested in animal personality, because one of the key criteria for personality is

that individual differences in behaviour scores are maintained over time (Alison M. Bell, et al., 2009; Stamps & Groothuis, 2010). Similarly, in recent years physiologists have increasingly focused on individual differences that are consistent over time (Careau, Gifford, & Biro, 2014; Nespolo & Franco, 2007; Williams, 2008). Assessing broad sense repeatability is often a key part of studies of individual differences in labile traits (Nakagawa & Schielzeth, 2010; Wolak, et al., 2012), and the statistic *R* has been calculated hundreds of time to infer broad sense repeatability of behaviour (e.g. Bell 2009; meta-analysis of behaviour: >750 estimates of *R*) and physiology (Nespolo & Franco, 2007; White, Schimpf, & Cassey, 2013).

<H1>Issues surrounding the use of *R*

Here, we raise some important issues relating to the use and interpretation of *R* when it is used to estimate broad sense repeatability. Longitudinal data (repeated measures over time) are necessarily at the core of any study of individual differences in labile traits, but most empirical studies ignore time-related change within and across individuals (see below, and Appendix Table A1). One of the indices that has been widely used to estimate the broad sense repeatability of labile traits is the intraclass correlation, or the ICC (Alison M. Bell, et al., 2009; Lessells & Boag, 1987a; Nakagawa & Schielzeth, 2010; Nespolo & Franco, 2007; Wolak, et al., 2012). Unfortunately, as was stressed long ago, the ICC ignores trait changes over time, which will lead to invalid and biased estimates of broad sense repeatability if such changes are present (Hayes & Jenkins, 1997; McGraw & Wong, 1996). Because the ICC is one of several different types of intraclass correlations (McGraw & Wong, 1996), to avoid confusion we follow earlier suggestions and refer to this index of *R* as ‘agreement R’, R_A (McGraw & Wong, 1996; Nakagawa & Schielzeth, 2010). Note that R_A can be calculated using a variety of different models, including single-factor ANOVA (e.g. see Lessells & Boag 1987) or mixed-effects models (e.g. see Nakagawa & Schielzeth 2010).

Unfortunately, if temporal patterns exist in the data, then R_A is not necessarily a good measure of broad sense repeatability, and we provide examples to illustrate why this is so. Critically, R_A assumes there is no temporal change in behaviour (i.e. there is no term for time in the underlying statistical model, see below). If such changes exist, R_A will provide an inaccurate estimate of broad sense repeatability, because key assumptions of that model have been violated (Hayes & Jenkins, 1997; McGraw & Wong, 1996). The remedy for the problem, discussed further below, is to include a term for time elapsed between repeated measures (when unequally spaced in time) or observation number in the model. In addition to satisfying model assumptions, incorporating change over time (a 'time effect') in the model serves the purpose of accounting for any changes in internal state, external stimuli and interactions between them that may have generated systematic temporal changes in behaviour at the mean or individual levels. A 'time effect' should not replace, but rather be used in addition to any obvious factors such as size, hunger, sex or temperature that could affect variation in the data across individuals and/or across successive measurements.

More generally, *R* will yield inaccurate estimates of broad sense repeatabiity if investigators ignore any factors, whether they be due to change over time or variation in some identifiable variable (variation in contexts), that might affect *R*. For instance, some investigators have estimated 'conservative' values of *R*, by deliberately excluding factors that might affect variation in the data (Laskowski & Bell, 2013; Nakagawa & Schielzeth, 2010). While this approach may be sufficient to test whether values of *R* are significantly greater than zero, it necessarily underestimates *R*, and may also violate assumptions of the statistical model used to estimate it (see below). Therefore, we advocate that researchers include predictors for both time-related change and change due to temporal variation in external stimuli (e.g. temperature) and factors such as sex and maturity when estimating *R*. We elaborate on this in later sections.

<H1>Effects of time are usually ignored

Despite cautions raised long ago (Hayes & Jenkins, 1997; McGraw & Wong, 1996), and despite a growing number of recent publications focusing on how to quantify individual differences in labile traits (e.g. Dingemanse, Kazem, Réale, & Wright, 2010; Martin, Nussey, Wilson, & Réale, 2011; Nakagawa & Schielzeth, 2010; Wolak, et al., 2012) and recent papers that explicitly consider temporal change (e.g. A. M. Bell & Peeke, 2012; Peter A. Biro, 2012; Dingemanse et al., 2012), the importance of including time when computing and intepreting *R* none the less continues to be ignored by most empiricists studying labile traits in nonhuman animals. For instance, we reviewed empirical studies published in three prominent behavioural journals (*Animal Behaviour*, *Behavioral Ecology*, *Behavioral Ecology and Sociobiology*) in 2011–2014, using the search keyword ‘repeatability’ in Web of Science. Of 41 relevant studies that reported repeatability to make inferences about consistency over time, only 39% tested for mean level (shared) effects of time on behaviour, and only 15% tested for individual differences in responses over time on behaviour (see Appendix Table A1). Thus, our aim is to educate those that are not aware of these issues, using simple examples that show how temporal change can seriously affect our estimates of broad sense repeatability.

Indeed, many authors either implicitly assume that behavioural or physiological traits are highly consistent over time, and then sample each individual only once (reviewed in Beckmann & Biro, 2013; Garamszegi, Markó, & Herczeg, 2012), or test for broad sense repeatability, but do so by only testing each subject twice (reviewed by Alison M. Bell, et al., 2009; Nespolo & Franco, 2007; Wolak, et al., 2012). This low level of replicates per individual implies that few investigators have explicitly considered just how labile physiological and behavioural traits can be, nor have they considered changes in behaviour over time, since multiple observations per individual are required to provide reasonable estimates of R_A, even in the absence of any time-related change (Wolak, et al., 2012). By

contrast, psychologists have a long tradition of explicitly modelling temporal variation in behaviour (Singer & Willett, 2003).

<H1>How temporally consistent are labile traits?

Currently, estimates of *R* reported in the empirical literature for nonhuman animals are rather low (mean = 0.4 or less) for both behavioural and physiological traits (reviewed by Alison M. Bell, et al., 2009; Nespolo & Franco, 2007; White, et al., 2013; Wolak, et al., 2012). Although many studies refer to *R* = 0.4 as 'substantial', the reality is that it can be very difficult to distinguish between individuals and ascertain consistency over time for samples with this value of *R* (e.g. see Fig. 1c). Low values of *R* might occur because (1) most of the variation resides within rather than across individuals, (2) broad sense repeatability is low (i.e. individual differences in scores are not maintained over the observation period) or (3) an investigator has failed to account (or control) for factors, including time, that affect trait variation (Hayes & Jenkins, 1997; McGraw & Wong, 1996; Nakagawa & Schielzeth, 2010).

<H1>What is narrow sense repeatability, *R*?

R is the proportion of the total variance in scores in a single context that is due to variance across individuals in their expected (mean) scores:

$$R = \frac{VAR\,across}{VAR\,across + VARresid} \qquad (1)$$

VAR_{across} indicates the variance across individuals in their expected values and VAR_{resid} is any unexplained residual (within-individual) variance in the data. Several assumptions must be satisfied for *R* to provide a valid estimate of the proportion of the total variance that is due to individual differences in expected values. Arguably, the most important of these is

that there is a common population (residual) variance for all measurement conditions (McGraw & Wong, 1996). Following from this are the related assumptions that residuals are random, independent and normally distributed (for Gaussian data). In practice, this means that for longitudinal data the VAR_{resid} should not change over time, that every individual in the sample should have the same residual variance around its expected value, and that the residuals around each individual's expected value should follow a normal distribution. For instance, if the assumption of a common population variance is not met due to the omission of a key factor(s) in the underlying model such as time, then it 'would be meaningless' to calculate any index of *R* (see also Hayes & Jenkins, 1997; p. 37, McGraw & Wong, 1996).

Importantly, even though *R* is often interpreted as an estimate of the extent to which individual differences in scores are maintained over time (Alison M. Bell, et al., 2009; Nespolo & Franco, 2007; Wolak, et al., 2012), one can plainly see that there is no term for time in equation 1. Therefore, if behaviour does systematically change over time, either in the same way in all of the subjects, or in different ways in different subjects, but these temporal changes are not accounted for in the model that is used to estimate *R*, then *R* should not be used to infer broad sense repeatability. Below we show why ignoring temporal changes in behaviour, if present, can lead to problems when *R* is used to estimate broad sense repeatability.

<H1>Different indices of *R:* which to use and when

The variances used to calculate *R* can be generated by a statistical model that contains different terms to address the effects of time, which change the relative size of each variance component in equation 1, and therefore any inferences about broad sense repeatability that follow from them. We outline the three major indices of *R* below, their assumptions about

change over time, and what they may or may not tell us about broad sense repeatability; in Table 1 we describe the underlying statistical model for each.

<H2>'Agreement' repeatability (R_A)

The most widely used version of R is R_A, an index that provides a measure of the agreement (or reproducibility) of the scores of different individuals (Hayes & Jenkins, 1997; McGraw & Wong, 1996; Nakagawa & Schielzeth, 2010). Traditionally, R_A has been measured using a single-factor ANOVA, in which there is no term for time and individual identity is the only predictor variable (Hayes & Jenkins, 1997; Lessells & Boag, 1987b; McGraw & Wong, 1996). More recently, mixed-effects models have been used to estimate R_A, where individual identity is specified as a random intercept effect. Here we focus on the latter models, because they provide direct estimates of variance (for any index of R), and handle unbalanced and missing data.

Because there is no term for time in the underlying model (see Table 1), R_A implicitly assumes that every individual's trend line over time is horizontal (see Fig. 1). If this is true, and if other assumptions are satisfied (mentioned above), then R_A can provide a useful estimate of broad sense repeatability (McGraw & Wong, 1996). Data that do satisfy the assumptions for R_A are simulated in Fig. 1. Here, the expected score of each individual does not change over time, and (within each sample) the residual variance around the expected values (VAR_{resid}) is the same for every individual. However, because VAR_{resid} differs between the samples, R_A also differs between Fig. 1a, b and c. Thus, even though the VAR_{across} is the same for all three samples (i.e. the individual intercepts are the same in Fig. 1a, b, c), individual differences in scores are more strongly maintained over time when $R_A = 0.9$ than when $R_A = 0.4$. As a result, broad sense repeatability is higher in Fig. 1a than in Fig. 1c.

Alternatively, of course, R_A would also vary across samples if the VAR_{resid} were the same for every sample, but VAR_{across} was higher in some samples than in others.

When mixed-effects models are used to generate estimates of R_A, these models specify an intercept for the fixed-effects portion (representing the population mean) and a variance parameter to describe VAR_{across}, which is given as the variance in individual intercepts VAR_{int}, termed a 'random intercept effect'; see R_A in Table 1).

<H2>'Consistency' repeatability (R_C)

If scores change systematically over time, then R_A provides biased estimates of broad sense repeatability. When shared changes over time exist (i.e. individual expected values over time are parallel, but not horizontal), then R_A cannot provide a good estimate of broad sense repeatability unless one accounts for these mean level changes in scores over time in the statistical model, yielding an index of R that has been called 'consistency' R, or R_C (McGraw & Wong, 1996; Nakagawa & Schielzeth, 2010). R_C is an index of R that accounts for any factor with equal effects on all of the individuals in the sample. An example of such a model is presented in Table 1. Failure to account for mean level change over time will lead to the residual variance changing over time, violating the constant variance assumption. This occurs because we are implicitly fitting horizontal trend lines for each individual, when all of the trend lines should instead be increasing or decreasing, with the same slopes. In turn, this leads to underestimates of broad sense repeatability, where the extent of the discrepancy depends on the extent to which mean level scores change over time (see Fig. 2). Here, R_C is the same for all three samples because neither VAR_{across} nor VAR_{resid} varies across samples (Fig. 1a, b, c): thus R_C correctly indicates that broad sense repeatability is the same for all three samples. However, if we ignore these mean level changes in scores over time, and use R_A instead, we would erroneously conclude that broad sense repeatability was substantially

higher in Fig. 2a than in Fig. 2b or c; this occurs because any shared within-individual change over time incorrectly becomes part of VAR_{resid}.

<u>*<H2>'Conditional' repeatability (R|condition)*</u>

When scores change over time, but the extent of change differs between individuals (Fig. 3a, b), then neither R_A nor R_C should be used to infer broad sense repeatability. In this situation, using R_A as an index of *R* is invalid because the key assumption of equal residual variance across individuals is violated: individuals whose behaviour changes markedly over time (a substantial time trend) have higher residual variance than individuals who maintain the same expected values over time (no time trend). Similarly, R_C cannot provide a valid index of *R* because it assumes that individuals all have the same time trends. If individuals differ in their time trends (Fig. 3b), then VAR_{across} necessarily also changes over time, and so must *R* as well. In other words, *R* varies as a function of time. In this case, the appropriate index of *R* has been termed 'conditional *R*' (Nakagawa & Schielzeth, 2010), where *R* is specific (conditional) to a particular value of time (here, *R*|time).

Unfortunately, *R*|time cannot be used to estimate broad sense repeatability, because a value of *R* that is specific to only one point in time cannot tell us about the extent to which individual differences in scores are maintained across the observation period. Rather, *R*|time tells us the extent to which individuals differ at a given point in time, under the assumption that within-individual (residual) variance is constant across individuals and over time (Fig 3b). If values of *R*|time change dramatically across observations, this implies that broad sense repeatability is low, but there is no simple mapping between *R*|time and broad sense repeatability.

A statistical model that can be used to determine whether individuals have different trend lines over time is outlined in Table 1. Detailed descriptions of this type of mixed

model, called ‘random regression’, can be found in several good texts (e.g. Singer & Willett, 2003; Verbeke & Molenberghs, 2009; Zuur, Ieno, Walker, Saveliev, & Smith, 2009). Briefly, if there is significant variance across individuals in their estimated slopes (VAR_{slopes}), then individuals differ in trends over time, and *R* must therefore vary as a function of time (Table 1).

<H1>Summary of what *R* tells us about broad sense repeatability

Currently, most empirical studies use R_A to estimate broad sense repeatability, and then use these estimates of R_A to infer the extent to which individual differences in scores are maintained over time (Alison M. Bell, et al., 2009; Nakagawa & Schielzeth, 2010; Wolak, et al., 2012), Appendix Table A1). However, R_A only provides a valid estimate of broad sense repeatability if behaviour does not change over time. If there are shared trends over time (i.e. a significant fixed effect of time), then R_C should be used instead of R_A. The indices R_A and R_C can provide reasonable estimates of broad sense repeatability only if individual trends over time all have zero slopes, or if individual trends are nonzero but parallel, respectively (see above). Finally, if the functional relationships between behaviour and time differ significantly between individuals (i.e. VAR_{slopes} is significant), then *R*|time can be used to estimate the extent to which individuals differ at a given point in time. However, in this situation none of the indices of *R* discussed above can provide valid estimates of broad sense repeatability (see above; Table 1).

<H1>Assumptions when choosing an index of *R*

Before using any index of *R* to estimate the level of broad sense repeatability in a sample, we must verify that we have not violated assumptions of the model used to generate that index. Testing assumptions should begin by first asking whether individuals have different slopes (see *R*|time, Table 1) or different residual variation. If there is no indication

that slopes differ between individuals, or that individuals differ in residual variation, then one can test for a shared effect of time (see R_C, Table 1). If there is a shared (fixed) effect of time, then R_C can be used to assess broad sense repeatability. If not (time effect $P > 0.1$), then one may simplify the underlying model further by removing the fixed effect of time and then use R_A to estimate broad sense repeatability (Table 1). An essential part of this process is to plot model predicted values against the raw data for each individual in the sample, to ensure that model predictions are meaningful, and to verify assumptions about residuals (see above). In addition, if the focus of a given study is on individual differences, then one should report individual level data and model predictions in relation to the repeated measures.

One practical difficulty with testing assumptions is that detecting individual differences in slopes with reasonable power requires very large samples. Depending upon assumptions about the size of VAR_{resid}, this can require total sample sizes (individuals and repeated measures per individual) of nearly 1000 (Martin, et al., 2011; van de Pol, 2012). To date, most studies of labile traits reporting *R* measure about 30 individuals twice each (Alison M. Bell, et al., 2009; Nespolo & Franco, 2007; Wolak, et al., 2012), which is clearly insufficient to detect individual differences in slopes with power or precision (Martin, et al., 2011; van de Pol, 2012). With such small samples, one could not conclude much if a statistical test for shared or nonshared time trends yielded a statistically nonsignificant result.

In a situation in which significant differences in individual slopes (trends over time) are detected, how can one obtain a reasonable index of broad sense repeatability, given that none of the indices of *R* are valid? At present we do not have a solution to this problem. This is because broad sense repeatability refers to the extent to which individual differences in scores are maintained over time; it does not refer to the extent to which individual differences in expected values are maintained over time. If one is interested in the temporal consistency of expected values (as opposed to the raw scores), then this might be explored using an effect

size estimator of the variation in individual slopes over time (Singer & Willett, 2003). Alternatively, the range of *R*|time values across the observation period might provide an index of the extent to which individual differences in expected values were maintained across the observation period.

<H1>Sample sizes and confounding factors

Throughout this discussion we have assumed that behaviour or physiology is measured under carefully controlled conditions, that repeated measures for all of the subjects were all taken in a single context (same set of external stimuli) using protocols that controlled for variation in many of the other factors that contribute to behavioural variability (e.g. time of day, feeding history, sex, age, etc.). Failure to control experimentally for these sources of variability could inflate estimates of *R* (in the case of sex differences) or underestimate *R* (in the case of time of day variation). For instance, the repeatability of metabolism declines with time between successive measures (White, et al., 2013), suggesting that ontogenetic or ageing effects may confound our estimates of broad sense repeatability if we do not account for time effects. In some cases, with sufficient samples, it may be possible to measure and then control statistically for sources of variability (other than time) using additional fixed and/or random effects (Peter A Biro, Adriaenssens, & Sampson, 2014). However, the greater the number of such effects, the greater the chance that individual differences will be confounded with these effects, reducing the power to detect and estimate broad sense repeatability (see also discussion by Martin & Reale, 2008).

A related issue is whether sample sizes are sufficient to provide reasonably precise estimates of the value of *R* (and by extension, reasonable estimates of individual means), as opposed to simply testing whether $R > 0$. For instance, even in the absence of any time-related change in scores at the mean or individual levels, or any other confounding fixed

effects, one would need to sample about 100 individuals, five times each (or 250 individuals twice each), in order to estimate an R_A value of 0.4 with reasonable precision (see Figure 3 in Wolak et al. 2012). Using data from Bell et al. (2009), of some 759 estimates of behavioural repeatability we estimated that the average study (with ca. 40 individuals and two repeated measures) have only 20% of the required sample size mentioned above. Thus, both past studies (Bell et al. 2009) and recent ones (Wolak et al. 2012) typically have sample sizes too low for rigorous estimates of R_A. At the same time, larger sample sizes provide more robust estimates of individual predicted mean values which, in addition to estimates of R, aid in exploring links between traits at the across-individual level (see Adolph & Hardin, 2007).

Of course, it is obvious that large sample sizes and careful controls over environmental conditions are much easier to achieve in the laboratory than in the field. Even so, researchers studying free-living animals have been able to gather substantial numbers of repeated measures (Carter, Heinsohn, Goldizen, & Biro, 2012), and have been able to detect not only changes in mean level behaviour as a function of time (e.g. Martin & Reale, 2008), but also significant individual differences in the rates of change in behaviour as a function of time (e.g. Dingemanse, et al., 2012). Hence it is clearly feasible for investigators studying free-living animals to determine an appropriate index of *R* (or none!) to estimate broad sense repeatability of those animals. Thus, it should be possible to increase the number of samples per individual beyond the $N = 2$ that is still common in many field studies reporting *R*.

<H1>Concluding Remarks

We hope to have convinced the reader that using *R* to infer broad sense repeatability is not as simple as commonly supposed, and requires much larger sample sizes than is usually the case. There are different indices of *R*, and whether any of them can provide a useful index of the temporal consistency of individual scores requires us to explicitly consider the

possibility that trait values might systematically change over time. If they do, then using indices of *R* that ignore changes in scores over time can result in invalid (due to violations of assumptions) or seriously biased estimates of broad sense repeatability. More generally, now that there is abundant empirical evidence that many labile traits are 'repeatable' we suggest that researchers, especially those studying animals in the laboratory, pay less attention to whether or not *R* is significantly greater than zero, and more attention to obtaining robust estimates of the repeatability of behavioural and physiological traits.

References

Adolph, S., & Hardin, J. (2007). Estimating phenotypic correlations: correcting for bias due to intraindividual variability. *Functional Ecology*, *21*, 178-184.

Beckmann, C., & Biro, P. A. (2013). On the validity of a single (boldness) assay in personality research. *Ethology*, *119*(11), 937-947.

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

Bell, A. M., & Peeke, H. V. S. (2012). Individual variation in habituation: behaviour over time toward different stimuli in threespine sticklebacks (Gasterosteus aculeatus). *BEHAVIOUR*, *149*(13-14), 1339-1365.

Biro, P. A. (2012). Do rapid assays predict repeatability in labile (behavioural) traits? [doi: 10.1016/j.anbehav.2012.01.036]. *Animal Behaviour*, *83*(5), 1295-1300.

Biro, P. A., Adriaenssens, B., & Sampson, P. (2014). Individual and sex-specific differences in intrinsic growth rate covary with consistent individual differences in behaviour. *Journal of Animal Ecology*, *83*, 1186-1195.

Careau, V., Gifford, M. E., & Biro, P. A. (2014). Individual (co-)variation in thermal reaction norms of standard and maximal metabolic rates in wild-caught slimy salamanders. *Functional Ecology*, *In press*.

Carter, A. J., Heinsohn, R., Goldizen, A. W., & Biro, P. A. (2012). Boldness, trappability and sampling bias in wild lizards. [doi: 10.1016/j.anbehav.2012.01.033]. *Animal Behaviour*, *83*(4), 1051-1058.

Caspi, A., & Roberts, B. W. (2001). Personality development across the life course: The argument for change and continuity. *Psychological Inquiry*, *12*(2), 49-66.

Dingemanse, N. J., Bouwman, K. M., van de Pol, M., van Overveld, T., Patrick, S. C., Matthysen, E., et al. (2012). Variation in personality and behavioural plasticity across four populations of the great tit Parus major. *Journal of Animal Ecology*, *81*(1), 116-126.

Dingemanse, N. J., Kazem, A. J. N., Réale, D., & Wright, J. (2010). Behavioural reaction norms: animal personality meets individual plasticity. *Trends in Ecology & Evolution*, *25*(2), 81-89.

Dohm, M. R. (2002). Repeatability estimates do not always set an upper limit to heritability. *Functional Ecology*, *16*(2), 273-280.

Falconer, D. S. (1981). *Introduction to quantitative genetics*, *2nd ed*. London: Longman.

Garamszegi, L., Markó, G., & Herczeg, G. (2012). A meta-analysis of correlated behaviours with implications for behavioural syndromes: mean effect size, publication bias, phylogenetic effects and the role of mediator variables. *Evolutionary Ecology*, *26*(5), 1213-1235.

Hayes, J. P., & Jenkins, S. H. (1997). Individual variation in mammals. *Journal of Mammalogy*, 274-293.

Laskowski, K. L., & Bell, A. M. (2013). Competition avoidance drives individual differences in response to a changing food resource in sticklebacks. *Ecology Letters*, *16*(6), 746-753.

Lessells, C. M., & Boag, P. T. (1987a). Unrepeatable repeatabilities - a common mistake. *Auk*, *104*(1), 116-121.

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

Martin, J. G. A., Nussey, D. H., Wilson, A. J., & Réale, D. (2011). Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. *Methods in Ecology and Evolution*, *4*, 362-374.

Martin, J. G. A., & Reale, D. (2008). Temperament, risk assessment and habituation to novelty in eastern chipmunks, *Tamias striatus*. *Animal Behaviour*, *75*(1), 309-318.

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

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, *85*(4), 935-956.

Nespolo, R. F., & Franco, M. (2007). Whole-animal metabolic rate is a repeatable trait: a meta-analysis. *Journal of Experimental Biology*, *210*(11), 2000-2005.

Roberts, B. W., Caspi, A., & Moffitt, T. E. (2001). The kids are alright: Growth and stability in personality development from adolescence to adulthood. *Journal of Personality and Social Psychology*, *81*(4), 670-683.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence*. New York: Oxford University Press.

Stamps, J. A., & Groothuis, T. G. G. (2010). The development of animal personality: relevance, concepts and perspectives. *Biological Reviews*, *85*, 301-325.

van de Pol, M. (2012). Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models. *Methods in Ecology and Evolution*, *3*(2), 268-280.

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*: Springer.

White, C. R., Schimpf, N. G., & Cassey, P. (2013). The repeatability of metabolic rate declines with time. *Journal of Experimental Biology*, *216*(10), 1763-1765.

Williams, T. D. (2008). Individual variation in endocrine systems: moving beyond the 'tyranny of the Golden Mean'. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *363*(1497), 1687-1698.

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution*, *3*(1), 129-137.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*: Springer.

Figure 1. Hypothetical (simulated) samples of six individuals sampled repeatedly over time. Within each sample (a–c), the residual variance (around the expected values) is the same for every individual, and neither the expected behaviour nor the residual values of each individual change as a function of time. Although VAR_{across} is the same in a, b and c (individual expected values, i.e. the intercepts, are the same), the residual variance (VAR_{resid}) differs, generating R_A values of (a) 0.9, (b) 0.6 and (c) 0.4. At present, many behavioural and physiological studies report R_A values of less than 0.4 (Alison M. Bell, et al., 2009; Nespolo & Franco, 2007).

Figure 2. Simulated data showing the effect of shared (mean level) change over time on estimates of R_A and R_C, when VAR_{across} (variance in individual intercepts) and VAR_{resid} (within-individual variation) are both held constant. (a) $R_A = 0.9$, $R_C = 0.9$, slope = 0. (b) $R_A = 0.77$, $R_C = 0.9$, slope = 1. (c) $R_A = 0.45$, $R_C = 0.9$, slope = 2. Individual intercepts are also identical in a, b and c. In this example VAR_{resid} is assumed to be very low in order to more clearly distinguish individual trends from one another.

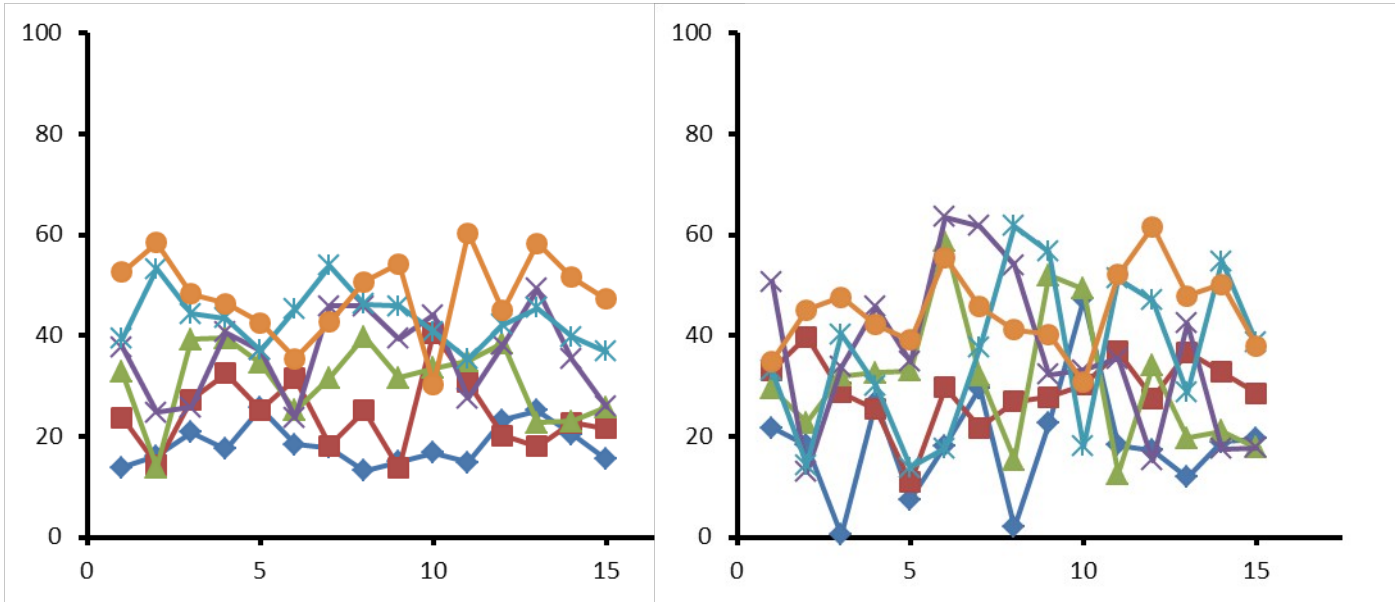
Figure 3. Simulated data showing how the extent to which individuals differ in their trends (expected values) over time affects the various indices of R (i.e. VAR_{slopes} differs between a, b). (a) $R_A = 0.75$, $R_C = 0.9$, $R|$'time = 1' = 0.9, $R|$'time = 15' = 0.9. (b) $R_A = 0.5$, $R_C = 0.75$, $R|$'time = 1' = 0.7, $R|$'time = 15' = 0.96. For simplicity, individual intercepts are held constant, but individual slopes differ, in a and b. Residual variance is identical and low in a and b to aid the reader in distinguishing the individual trend lines.

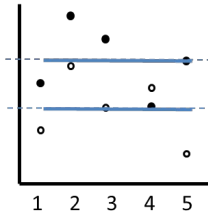
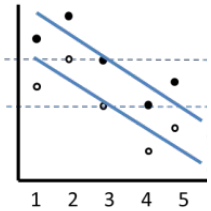
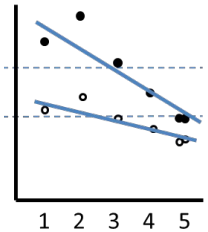
461Appendix

462

Mean-level time considered?	Individual-level time effect considered?	Year	Journal	Authors	Title of article
Yes	No	2014	Anim Behav	Watts et al.	Diel patterns of foraging aggression and antipredator behaviour
Yes	Yes	2014	Anim Behav	Davy et al.	When righting is wrong: performance measures require rank rep
No	No	2014	Anim Behav	Trnka and Grim	Testing for correlations between behaviours in a cuckoo host: wl
No	No	2014	Anim Behav	Jacobs et al.	Personality-dependent response to field playback in great tits: sl
No	No	2014	Anim Behav	Taylor et al.	Colour use by tiny predators: jumping spiders show colour biase
No	No	2014	Anim Behav	Laskowski and Bell	Strong personalities, Not social niches, drive individual differenc
No	No	2014	Anim Behav	Sussman et al.	Tenure in current captive setting and age predict personality cha
No	No	2013	Anim Behav	Petelle et al.	Development of boldness and docility in yellow-bellied marmots
No	No	2013	Anim Behav	Nandi and Balakrishnan	Call intensity is a repeatable and dominant acoustic feature dete
No	No	2013	Anim Behav	Jennings et al.	Personality and predictability in fallow deer fighting behaviour: t
Yes	No	2013	Anim Behav	Fowler-Finn and Rodriguez	Repeatability of mate preference functions in EncheNopa treeho
No	No	2012	Anim Behav	Dammhahn and Almeling	Is risk taking during foraging a personality trait? A field test for c
No	No	2012	Anim Behav	Seltmann et al.	Stress responsiveness, age and body condition interactively affe
No	No	2012	Anim Behav	Deb et al.	Females of a tree cricket prefer larger males but Not the lower f
No	No	2012	Anim Behav	Kluen et al.	A simple cage test captures intrinsic differences in aspects of per
Yes	Yes	2012	Anim Behav	Stamps et al.	Unpredictable animals: individual differences in intraindividual v
Yes	Yes	2012	Anim Behav	Biro	Do rapid assays predict repeatability in labile (behavioural) traits
Yes	Yes	2012	Anim Behav	Carter et al.	Boldness, trappability and sampling bias in wild lizards
Yes	No	2012	Anim Behav	Betini et al.	The relationship between personality and plasticity in tree swall
No	No	2011	Anim Behav	David et al.	Personality affects zebra finch feeding success in a producer-scro
No	No	2011	Anim Behav	Jenkins	Sex differences in repeatability of food-hoarding behaviour of ka
No	No	2011	Anim Behav	David et al.	Personality predicts social dominance in female zebra finches, Ta

No	No	2014	Behav Ecol Socio	Kortet et al.	Behavioral variation shows heritability in juvenile brown trout Sa
Yes	Yes	2014	Behav Ecol Socio	Grim et al.	The repeatability of avian egg ejection behaviors across differen
Yes	No	2014	Behav Ecol Socio	Boulton et al.	How stable are personalities? A multivariate view of behaviourai
No	No	2014	Behav Ecol Socio	ToscaNo et al.	Effect of predation threat on repeatability of individual crab beh
No	No	2014	Behav Ecol Socio	Kekalainen et al.	Do brain parasites alter host personality? - Experimental study ir
No	No	2014	Behav Ecol Socio	Kluen et al.	Testing for between individual correlations of personality and ph
No	No	2013	Behav Ecol Socio	Fitzsimmons et al.	Signaling effort does Not predict aggressiveness in male spring fi
Yes	No	2013	Behav Ecol Socio	Cordes et al.	Risk-taking behavior in the lesser wax moth: disentangling withir
Yes	Yes	2012	Behav Ecol Socio	Lupold et al.	Seasonal variation in ejaculate traits of male red-winged blackbi
No	No	2012	Behav Ecol Socio	Hedrick and Kortet	Sex differences in the repeatability of boldness over metamorph
Yes	No	2011	Behav Ecol Socio	Koski	Social personality traits in chimpanzees: temporal stability and s
No	No	2011	Behav Ecol Socio	Gladbach et al.	Can faecal glucocorticoid metabolites be used to monitor body c
No	No	2014	Behav Ecol	Wignall et al.	Extreme short-term repeatability of male courtship performance
Yes	No	2014	Behav Ecol	Grunst et al.	Age-dependent relationships between multiple sexual pigments
Yes	No	2014	Behav Ecol	Perez et al.	When males are more inclined to stay at home: insights into the
Yes	No	2013	Behav Ecol	Carvalho et al.	Personality traits are related to ecology across a biological invasi
No	No	2013	Behav Ecol	Kluen and Brommer	Context-specific repeatability of personality traits in a wild bird:
No	No	2012	Behav Ecol	Edelaar et al.	Tonic immobility is a measure of boldness toward predators: an
Yes	No	2012	Behav Ecol	Low et al.	Food availability and offspring demand influence sex-specific pat



Index of R	Assumptions	Graphical depiction of the assumptions (each line represents one individual)	Assumptions in terms of statistical effects	Statistical model effects: fixed effects (mean-level trend)	Statistical model effects: random effects (across-individual variance)
R_A	Individual expected values do not change over time		(a) individual differences in intercepts	Y = intercept	VAR_{int}
R_C	Individual expected values change identically over time		(a) individual differences in intercepts (b) mean-level effect of TIME	Y = intercept + TIME	VAR_{int}
$R\|time$	Individual expected values change over time differently		(a) individual differences in intercepts (b) mean-level effect of TIME (c) individual differences	Y = intercept + TIME	VAR_{int} VAR_{slope} $COV_{i,s}$

in change over time (slopes)

(d) covariance between individual intercepts and slopes ($COV_{i,s}$)

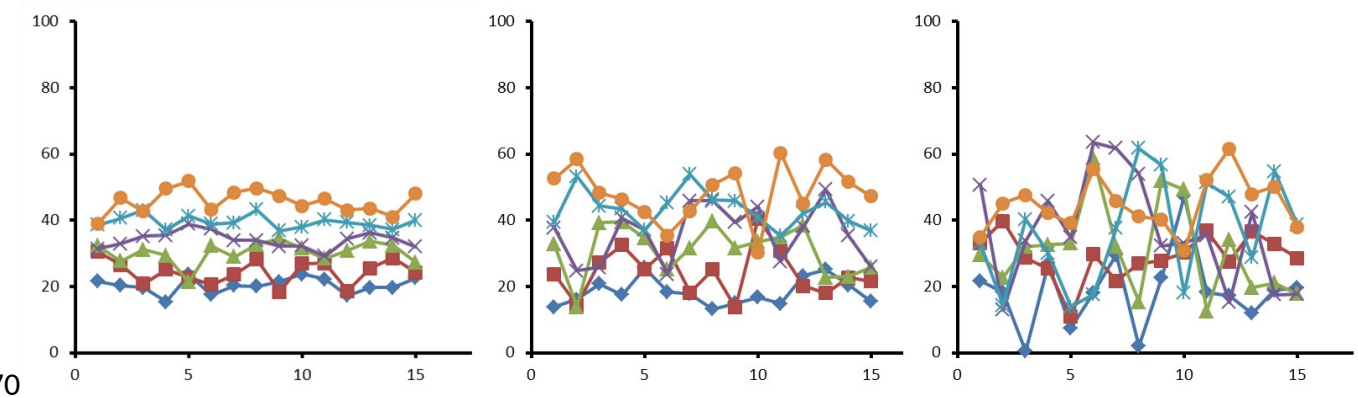


Figure 1

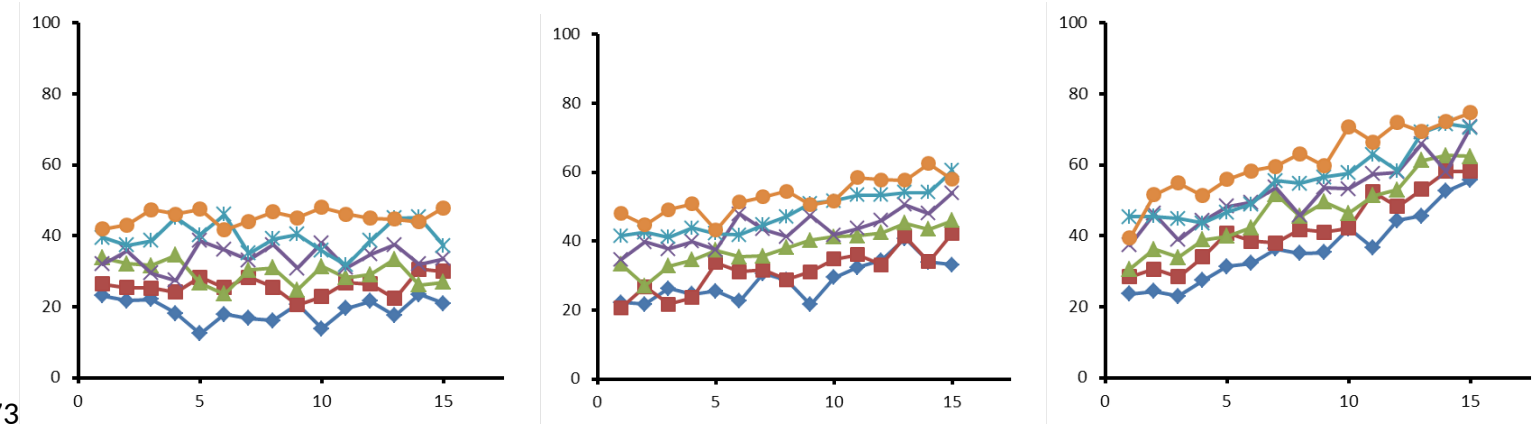

100
80
60
40
20
0
0
5
10
15
100
80
60
40
20
0
0
5
10
15
100
80
60
40
20
0
0
5
10
15

Figure 2