UC Davis

UC Davis Previously Published Works

Title

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

Permalink

<https://escholarship.org/uc/item/5qq6246x>

Authors

Biro, PA
Stamps, JA

Publication Date

2015-07-01

DOI

10.1016/j.anbehav.2015.04.008

Peer reviewed

1Commentary**2Using repeatability to study physiological and behavioural traits: ignore time-related
3change at your peril**

4Peter A. Biro^{a,*} Judy A. Stamps^b

5^a Centre for Integrative Ecology, School of Life and Environmental Science, Deakin

6University, Geelong, Australia

7^b Department of Evolution & Ecology, University of California Davis, Davis, CA, U.S.A.

8Received 8 December 2014

9Initial acceptance 9 January 2015

10Final acceptance 12 March 2015

11MS. number: 14-00991R

12*Correspondence: P. A. Biro, Centre for Integrative Ecology, School of Life and

13Environmental Science, Deakin University, Geelong 3216, Australia.

14E-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,

35

36 A major challenge in studying and describing behavioural and physiological traits is
37 their lability. In contrast to morphological traits, physiology and behaviour are labile traits
38 that can change over short periods (e.g. seconds to days) in response to changes in internal
39 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

65that individual differences in behaviour scores are maintained over time (Alison M. Bell, et
66al., 2009; Stamps & Groothuis, 2010). Similarly, in recent years physiologists have
67increasingly focused on individual differences that are consistent over time (Careau, Gifford,
68& Biro, 2014; Nespolo & Franco, 2007; Williams, 2008). Assessing broad sense repeatability
69is often a key part of studies of individual differences in labile traits (Nakagawa & Schielzeth,
702010; Wolak, et al., 2012), and the statistic R has been calculated hundreds of time to infer
71broad sense repeatability of behaviour (e.g. Bell 2009; meta-analysis of behaviour: >750
72estimates of R) and physiology (Nespolo & Franco, 2007; White, Schimpf, & Cassey, 2013).

73<H1>Issues surrounding the use of R

74 Here, we raise some important issues relating to the use and interpretation of R when
75it is used to estimate broad sense repeatability. Longitudinal data (repeated measures over
76time) are necessarily at the core of any study of individual differences in labile traits, but
77most empirical studies ignore time-related change within and across individuals (see below,
78and Appendix Table A1). One of the indices that has been widely used to estimate the broad
79sense repeatability of labile traits is the intraclass correlation, or the ICC (Alison M. Bell, et
80al., 2009; Lessells & Boag, 1987a; Nakagawa & Schielzeth, 2010; Nespolo & Franco, 2007;
81Wolak, et al., 2012). Unfortunately, as was stressed long ago, the ICC ignores trait changes
82over time, which will lead to invalid and biased estimates of broad sense repeatability if such
83changes are present (Hayes & Jenkins, 1997; McGraw & Wong, 1996). Because the ICC is
84one of several different types of intraclass correlations (McGraw & Wong, 1996), to avoid
85confusion we follow earlier suggestions and refer to this index of R as ‘agreement R ’, R_A
86(McGraw & Wong, 1996; Nakagawa & Schielzeth, 2010). Note that R_A can be calculated
87using a variety of different models, including single-factor ANOVA (e.g. see Lessells & Boag
881987) or mixed-effects models (e.g. see [Nakagawa & Schielzeth 2010](#)).

10

89 Unfortunately, if temporal patterns exist in the data, then R_A is not necessarily a good
90measure of broad sense repeatability, and we provide examples to illustrate why this is so.
91Critically, R_A assumes there is no temporal change in behaviour (i.e. there is no term for time
92in the underlying statistical model, see below). If such changes exist, R_A will provide an
93inaccurate estimate of broad sense repeatability, because key assumptions of that model have
94been violated (Hayes & Jenkins, 1997; McGraw & Wong, 1996). The remedy for the
95problem, discussed further below, is to include a term for time elapsed between repeated
96measures (when unequally spaced in time) or observation number in the model. In addition
97to satisfying model assumptions, incorporating change over time (a ‘time effect’) in the
98model serves the purpose of accounting for any changes in internal state, external stimuli and
99interactions between them that may have generated systematic temporal changes in behaviour
100at the mean or individual levels. A ‘time effect’ should not replace, but rather be used in
101addition to any obvious factors such as size, hunger, sex or temperature that could affect
102variation in the data across individuals and/or across successive measurements.

103 More generally, R will yield inaccurate estimates of broad sense repeatability if
104investigators ignore any factors, whether they be due to change over time or variation in some
105identifiable variable (variation in contexts), that might affect R . For instance, some
106investigators have estimated ‘conservative’ values of R , by deliberately excluding factors that
107might affect variation in the data (Laskowski & Bell, 2013; Nakagawa & Schielzeth, 2010).
108While this approach may be sufficient to test whether values of R are significantly greater
109than zero, it necessarily underestimates R , and may also violate assumptions of the statistical
110model used to estimate it (see below). Therefore, we advocate that researchers include
111predictors for both time-related change and change due to temporal variation in external
112stimuli (e.g. temperature) and factors such as sex and maturity when estimating R . We
113elaborate on this in later sections.

114<H1>Effects of time are usually ignored

115 Despite cautions raised long ago (Hayes & Jenkins, 1997; McGraw & Wong, 1996),
116and despite a growing number of recent publications focusing on how to quantify individual
117differences in labile traits (e.g. Dingemanse, Kazem, Réale, & Wright, 2010; Martin, Nussey,
118Wilson, & Réale, 2011; Nakagawa & Schielzeth, 2010; Wolak, et al., 2012) and recent papers
119that explicitly consider temporal change (e.g. A. M. Bell & Peeke, 2012; Peter A. Biro, 2012;
120Dingemanse et al., 2012), the importance of including time when computing and interpreting
121R none the less continues to be ignored by most empiricists studying labile traits in
122nonhuman animals. For instance, we reviewed empirical studies published in three prominent
123behavioural journals (*Animal Behaviour*, *Behavioral Ecology*, *Behavioral Ecology and*
124*Sociobiology*) in 2011–2014, using the search keyword ‘repeatability’ in Web of Science. Of
12541 relevant studies that reported repeatability to make inferences about consistency over time,
126only 39% tested for mean level (shared) effects of time on behaviour, and only 15% tested for
127individual differences in responses over time on behaviour (see Appendix Table A1). Thus,
128our aim is to educate those that are not aware of these issues, using simple examples that
129show how temporal change can seriously affect our estimates of broad sense repeatability.

130 Indeed, many authors either implicitly assume that behavioural or physiological traits
131are highly consistent over time, and then sample each individual only once (reviewed in
132Beckmann & Biro, 2013; Garamszegi, Markó, & Herczeg, 2012), or test for broad sense
133repeatability, but do so by only testing each subject twice (reviewed by Alison M. Bell, et al.,
1342009; Nespolo & Franco, 2007; Wolak, et al., 2012). This low level of replicates per
135individual implies that few investigators have explicitly considered just how labile
136physiological and behavioural traits can be, nor have they considered changes in behaviour
137over time, since multiple observations per individual are required to provide reasonable
138estimates of R_A , even in the absence of any time-related change (Wolak, et al., 2012). By

139contrast, psychologists have a long tradition of explicitly modelling temporal variation in
140behaviour (Singer & Willett, 2003).

141<H1>**How temporally consistent are labile traits?**

142 Currently, estimates of R reported in the empirical literature for nonhuman animals are
143rather low (mean = 0.4 or less) for both behavioural and physiological traits (reviewed by
144Alison M. Bell, et al., 2009; Nespolo & Franco, 2007; White, et al., 2013; Wolak, et al.,
1452012). Although many studies refer to $R = 0.4$ as ‘substantial’, the reality is that it can be
146very difficult to distinguish between individuals and ascertain consistency over time for
147samples with this value of R (e.g. see Fig. 1c). Low values of R might occur because (1)
148most of the variation resides within rather than across individuals, (2) broad sense
149repeatability is low (i.e. individual differences in scores are not maintained over the
150observation period) or (3) an investigator has failed to account (or control) for factors,
151including time, that affect trait variation (Hayes & Jenkins, 1997; McGraw & Wong, 1996;
152Nakagawa & Schielzeth, 2010).

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

154 R is the proportion of the total variance in scores in a single context that is due to
155variance across individuals in their expected (mean) scores:

$$156R = \frac{VAR_{across}}{VAR_{across} + VAR_{resid}} \quad (1)$$

157 VAR_{across} indicates the variance across individuals in their expected values and VAR_{resid}
158is any unexplained residual (within-individual) variance in the data. Several assumptions
159must be satisfied for R to provide a valid estimate of the proportion of the total variance that
160is due to individual differences in expected values. Arguably, the most important of these is

161that there is a common population (residual) variance for all measurement conditions
162(McGraw & Wong, 1996). Following from this are the related assumptions that residuals are
163random, independent and normally distributed (for Gaussian data). In practice, this means
164that for longitudinal data the VAR_{resid} should not change over time, that every individual in the
165sample should have the same residual variance around its expected value, and that the
166residuals around each individual's expected value should follow a normal distribution. For
167instance, if the assumption of a common population variance is not met due to the omission
168of a key factor(s) in the underlying model such as time, then it 'would be meaningless' to
169calculate any index of R (see also Hayes & Jenkins, 1997; p. 37, McGraw & Wong, 1996).

170 Importantly, even though R is often interpreted as an estimate of the extent to which
171individual differences in scores are maintained over time (Alison M. Bell, et al., 2009;
172Nespolo & Franco, 2007; Wolak, et al., 2012), one can plainly see that there is no term for
173time in equation 1. Therefore, if behaviour does systematically change over time, either in
174the same way in all of the subjects, or in different ways in different subjects, but these
175temporal changes are not accounted for in the model that is used to estimate R , then R should
176not be used to infer broad sense repeatability. Below we show why ignoring temporal
177changes in behaviour, if present, can lead to problems when R is used to estimate broad sense
178repeatability.

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

180 The variances used to calculate R can be generated by a statistical model that contains
181different terms to address the effects of time, which change the relative size of each variance
182component in equation 1, and therefore any inferences about broad sense repeatability that
183follow 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 ($\text{VAR}_{\text{resid}}$) is the same for every individual. However, because $\text{VAR}_{\text{resid}}$ differs between the samples, R_A also differs between Fig. 1a, b and c. Thus, even though the $\text{VAR}_{\text{across}}$ 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.

207Alternatively, of course, R_A would also vary across samples if the VAR_{resid} were the same for
208every sample, but VAR_{across} was higher in some samples than in others.

209 When mixed-effects models are used to generate estimates of R_A , these models
210specify an intercept for the fixed-effects portion (representing the population mean) and a
211variance parameter to describe VAR_{across} , which is given as the variance in individual
212intercepts VAR_{int} , termed a ‘random intercept effect’; see R_A in Table 1).

213<H2>‘Consistency’ repeatability (R_C)

214 If scores change systematically over time, then R_A provides biased estimates of broad
215sense repeatability. When shared changes over time exist (i.e. individual expected values
216over time are parallel, but not horizontal), then R_A cannot provide a good estimate of broad
217sense repeatability unless one accounts for these mean level changes in scores over time in
218the statistical model, yielding an index of R that has been called ‘consistency’ R , or R_C
219(McGraw & Wong, 1996; Nakagawa & Schielzeth, 2010). R_C is an index of R that accounts
220for any factor with equal effects on all of the individuals in the sample. An example of such a
221model is presented in Table 1. Failure to account for mean level change over time will lead to
222the residual variance changing over time, violating the constant variance assumption. This
223occurs because we are implicitly fitting horizontal trend lines for each individual, when all of
224the trend lines should instead be increasing or decreasing, with the same slopes. In turn, this
225leads to underestimates of broad sense repeatability, where the extent of the discrepancy
226depends on the extent to which mean level scores change over time (see Fig. 2). Here, R_C is
227the same for all three samples because neither VAR_{across} nor VAR_{resid} varies across samples
228(Fig. 1a, b, c): thus R_C correctly indicates that broad sense repeatability is the same for all
229three samples. However, if we ignore these mean level changes in scores over time, and use
230 R_A instead, we would erroneously conclude that broad sense repeatability was substantially

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

233 <H2> ‘Conditional’ repeatability ($R|condition$)

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

245 Unfortunately, $R|time$ cannot be used to estimate broad sense repeatability, because a
246 value of R that is specific to only one point in time cannot tell us about the extent to which
247 individual differences in scores are maintained across the observation period. Rather, $R|time$
248 tells us the extent to which individuals differ at a given point in time, under the assumption
249 that within-individual (residual) variance is constant across individuals and over time (Fig
250 3b). If values of $R|time$ change dramatically across observations, this implies that broad
251 sense repeatability is low, but there is no simple mapping between $R|time$ and broad sense
252 repeatability.

253 A statistical model that can be used to determine whether individuals have different
254 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

304size estimator of the variation in individual slopes over time (Singer & Willett, 2003).

305Alternatively, the range of R |time values across the observation period might provide an
306index of the extent to which individual differences in expected values were maintained across
307the observation period.

308<H1>Sample sizes and confounding factors

309 Throughout this discussion we have assumed that behaviour or physiology is
310measured under carefully controlled conditions, that repeated measures for all of the subjects
311were all taken in a single context (same set of external stimuli) using protocols that controlled
312for variation in many of the other factors that contribute to behavioural variability (e.g. time
313of day, feeding history, sex, age, etc.). Failure to control experimentally for these sources of
314variability could inflate estimates of R (in the case of sex differences) or underestimate R (in
315the case of time of day variation). For instance, the repeatability of metabolism declines with
316time between successive measures (White, et al., 2013), suggesting that ontogenetic or ageing
317effects may confound our estimates of broad sense repeatability if we do not account for time
318effects. In some cases, with sufficient samples, it may be possible to measure and then
319control statistically for sources of variability (other than time) using additional fixed and/or
320random effects (Peter A Biro, Adriaenssens, & Sampson, 2014). However, the greater the
321number of such effects, the greater the chance that individual differences will be confounded
322with these effects, reducing the power to detect and estimate broad sense repeatability (see
323also discussion by Martin & Reale, 2008).

324 A related issue is whether sample sizes are sufficient to provide reasonably precise
325estimates of the value of R (and by extension, reasonable estimates of individual means), as
326opposed to simply testing whether $R > 0$. For instance, even in the absence of any time-
327related 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.

359

360References

- 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.

- 387Dingemanse, N. J., Kazem, A. J. N., Réale, D., & Wright, J. (2010). Behavioural reaction norms:
388 animal personality meets individual plasticity. *Trends in Ecology & Evolution*, 25(2),
389 81-89.
- 390Dohm, M. R. (2002). Repeatability estimates do not always set an upper limit to heritability.
391 *Functional Ecology*, 16(2), 273-280.
- 392Falconer, D. S. (1981). *Introduction to quantitative genetics*, 2nd ed. London: Longman.
- 393Garamszegi, L., Markó, G., & Herczeg, G. (2012). A meta-analysis of correlated behaviours
394 with implications for behavioural syndromes: mean effect size, publication bias,
395 phylogenetic effects and the role of mediator variables. *Evolutionary Ecology*, 26(5),
396 1213-1235.
- 397Hayes, J. P., & Jenkins, S. H. (1997). Individual variation in mammals. *Journal of Mammalogy*,
398 274-293.
- 399Laskowski, K. L., & Bell, A. M. (2013). Competition avoidance drives individual differences in
400 response to a changing food resource in sticklebacks. *Ecology Letters*, 16(6), 746-753.
- 401Lessells, C. M., & Boag, P. T. (1987a). Unrepeatable repeatabilities - a common mistake. *Auk*,
402 104(1), 116-121.
- 403Lessells, C. M., & Boag, P. T. (1987b). Unrepeatable repeatabilities: a common mistake. *The*
404 *Auk*, 104, 116-121.
- 405Martin, J. G. A., Nussey, D. H., Wilson, A. J., & Réale, D. (2011). Measuring individual
406 differences in reaction norms in field and experimental studies: a power analysis of
407 random regression models. *Methods in Ecology and Evolution*, 4, 362-374.
- 408Martin, J. G. A., & Reale, D. (2008). Temperament, risk assessment and habituation to
409 novelty in eastern chipmunks, *Tamias striatus*. *Animal Behaviour*, 75(1), 309-318.
- 410McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation
411 coefficients. *Psychological methods*, 1(1), 30.
- 412Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a
413 practical guide for biologists. *Biological Reviews*, 85(4), 935-956.
- 414Nespolo, R. F., & Franco, M. (2007). Whole-animal metabolic rate is a repeatable trait: a
415 meta-analysis. *Journal of Experimental Biology*, 210(11), 2000-2005.
- 416Roberts, B. W., Caspi, A., & Moffitt, T. E. (2001). The kids are alright: Growth and stability in
417 personality development from adolescence to adulthood. *Journal of Personality and*
418 *Social Psychology*, 81(4), 670-683.
- 419Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: modeling change and*
420 *event occurrence*. New York: Oxford University Press.
- 421Stamps, J. A., & Groothuis, T. G. G. (2010). The development of animal personality:
422 relevance, concepts and perspectives. *Biological Reviews*, 85, 301-325.
- 423van de Pol, M. (2012). Quantifying individual variation in reaction norms: how study design
424 affects the accuracy, precision and power of random regression models. *Methods in*
425 *Ecology and Evolution*, 3(2), 268-280.
- 426Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*: Springer.
- 427White, C. R., Schimpf, N. G., & Cassey, P. (2013). The repeatability of metabolic rate declines
428 with time. *Journal of Experimental Biology*, 216(10), 1763-1765.
- 429Williams, T. D. (2008). Individual variation in endocrine systems: moving beyond the 'tyranny
430 of the Golden Mean'. *Philosophical Transactions of the Royal Society B-Biological*
431 *Sciences*, 363(1497), 1687-1698.
- 432Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability.
433 *Methods in Ecology and Evolution*, 3(1), 129-137.

434Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models*
435 *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).

446

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.

453

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.

460

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: wh
No	No	2014	Anim Behav	Jacobs et al.	Personality-dependent response to field playback in great tits: sh
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 affec
No	No	2012	Anim Behav	Deb et al.	Females of a tree cricket prefer larger males but Not the lower fr
No	No	2012	Anim Behav	Klueen 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

41
42

21

No	No	2014	Behav Ecol Socio	Kortet et al.
Yes	Yes	2014	Behav Ecol Socio	Grim et al.
Yes	No	2014	Behav Ecol Socio	Boulton et al.
No	No	2014	Behav Ecol Socio	Toscano et al.
No	No	2014	Behav Ecol Socio	Kekalainen et al.
No	No	2014	Behav Ecol Socio	Kluen et al.
No	No	2013	Behav Ecol Socio	Fitzsimmons et al.
Yes	No	2013	Behav Ecol Socio	Cordes et al.
Yes	Yes	2012	Behav Ecol Socio	Lupold et al.
No	No	2012	Behav Ecol Socio	Hedrick and Kortet
Yes	No	2011	Behav Ecol Socio	Koski
No	No	2011	Behav Ecol Socio	Gladbach et al.
No	No	2014	Behav Ecol	Wignall et al.
Yes	No	2014	Behav Ecol	Grunst et al.
Yes	No	2014	Behav Ecol	Perez et al.
Yes	No	2013	Behav Ecol	Carvalho et al.
No	No	2013	Behav Ecol	Kluen and Brommer
No	No	2012	Behav Ecol	Edelaar et al.
Yes	No	2012	Behav Ecol	Low et al.

Behavioral variation shows heritability in juvenile brown trout *Salmo trutta* L.
The repeatability of avian egg ejection behaviors across different trials
How stable are personalities? A multivariate view of behavioural plasticity
Effect of predation threat on repeatability of individual crab behaviour
Do brain parasites alter host personality? - Experimental study in *Parasitica*
Testing for between individual correlations of personality and phenotype
Signaling effort does Not predict aggressiveness in male spring fish
Risk-taking behavior in the lesser wax moth: disentangling within and between individual effects
Seasonal variation in ejaculate traits of male red-winged blackbirds
Sex differences in the repeatability of boldness over metamorphosis
Social personality traits in chimpanzees: temporal stability and state-dependence
Can faecal glucocorticoid metabolites be used to monitor body condition in wild primates?
Extreme short-term repeatability of male courtship performance in a songbird
Age-dependent relationships between multiple sexual pigments in a fish
When males are more inclined to stay at home: insights into the evolution of parental care
Personality traits are related to ecology across a biological invasion
Context-specific repeatability of personality traits in a wild bird
Tonic immobility is a measure of boldness toward predators: an experimental test
Food availability and offspring demand influence sex-specific parental investment

463
464
465
466
467

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

45
46

23

in change over time (slopes)

(d) covariance between individual

intercepts and slopes ($COV_{i,s}$)

468

469

47
48

470

471

472Figure 1

49
50

473
474Figure 2