Animal Behaviour 105 (2015) 223–230

Contents lists available at ScienceDirect

Animal Behaviour

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

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.

ARTICLE INFO

Article history:
Received 8 December 2014
Initial acceptance 9 January 2015
Final acceptance 12 March 2015
Available online 26 May 2015
MS. number: 14-00991R

Keywords:
behavioural syndromes
metabolism
mixed models
personality
plasticity

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 occurs 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).

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 (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, 1987; 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

* 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).

http://dx.doi.org/10.1016/j.anbehav.2015.04.008
0003-3472/

one of the key criteria for personality is that individual differences in behaviour scores are maintained over time (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, et al., 2009; meta-analysis of behaviour: >750 estimates of *R*) and physiology (Nespolo & Franco, 2007; White, Schimpf, & Cassey, 2013).

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 (Bell, et al., 2009; Lessells & Boag, 1987; 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 repeatability 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.

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. Bell & Peeke, 2012; 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 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).

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 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).