Animal Behaviour 83 (2012) 1325–1334

Contents lists available at SciVerse ScienceDirect

Animal Behaviour

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

# Unpredictable animals: individual differences in intraindividual variability (IIV)

Judy A. Stamps [a,*], Mark Briffa [b,1], Peter A. Biro [c,2]

[a] Department of Evolution & Ecology, University of California Davis, Davis, U.S.A
[b] Marine Biology & Ecology Research Centre, School of Marine Science and Engineering, University of Plymouth, Plymouth, U.K
[c] Centre for Integrative Ecology, School of Life and Environmental Science, Deakin University, Geelong, Australia

ARTICLE INFO

*Article history:*
Received 8 November 2011
Initial acceptance 29 December 2011
Final acceptance 19 January 2012
Available online 4 April 2012
MS. number: A11-00898R

*Keywords:*
behavioural plasticity
censored data
intraindividual variability
personality
random regression
repeatability
sensitivity

When an individual is repeatedly observed or tested in the same context, it does not always express the same behaviour. Intraindividual variability (IIV) refers to the short-term, unpredictable, reversible variation in behaviour that often occurs in this situation. Although individual differences in IIV have been well documented in humans, this topic has been virtually ignored by researchers studying other animals. Here, we review evidence from humans and animals that IIV can vary in important ways across individuals (e.g. as a function of age or prior experience) and that individual differences in IIV may be related to differences in performance. However, most statistical models currently used to study individual differences in behaviour in animals rely on the assumption that IIV does not vary across individuals. Using 'boldness' data for hermit crabs, *Pagurus bernhardus*, and Ward's damselfish, *Pomacentrus wardi*, we show how to measure IIV when behaviour systematically changes over a series of observations (e.g. as a result of habituation), and how to avoid the adverse effects of censored data on estimates of IIV. After controlling for systematic changes in behaviour over time, we observed strong, significant individual differences in IIV in both species. That is, some individuals were much more predictable in the same situation than were others. We conclude by discussing proximate and ultimate factors that might have contributed to interindividual variation in IIV in these species, and the implications of our findings for methods currently used to study individual differences in behaviour in animals.


When individual animals are repeatedly observed or tested over short periods in the same context (i.e. the same external stimulus situation, see Stamps & Groothuis 2010a, b), they rarely behave in exactly the same way on every occasion. Instead, individual behaviour is often variable, even across short intertest intervals of seconds to days, when ontogenetic or seasonal changes in behaviour are not an issue. Psychologists have coined the term intraindividual variability (IIV) to refer to short-term, reversible variability in the behaviour of an individual that is repeatedly measured in the same context (Nesselroade 1991; Siegler 1994; Salthouse 2007; Ram & Gerstorf 2009).

Here, we begin by reviewing the literature on IIV, a topic virtually ignored by students of animal behaviour, but with important implications for the ways we interpret and study behavioural variation. We discuss previous studies of humans and a handful of animals indicating that IIV may vary across individuals as a function of age, experience or other factors, and that individual differences in IIV may be positively or negatively related to indices of performance. In addition, we discuss why individual differences in IIV may bias the results generated by statistical models currently used to study individual differences in behaviour. Then we use data from two species, hermit crabs, *Pagurus bernhardus*, and Ward's damselfish, *Pomacentrus wardi*, to demonstrate that IIV in behaviour differs markedly across individuals within these species, and show how one can test for individual differences in IIV and for relationships between IIV and other variables.

## *Intraindividual Variability in Behaviour: a Primer*

Traditionally, researchers studying individual differences in behaviour have either tended to ignore within-individual behavioural variability (e.g. by testing each subject once), or have acknowledged it by testing each individual twice (rarely, more often), and then using the mean score of each individual in subsequent analyses (Bell et al. 2009). In recent years, however, behavioural biologists have begun to consider short-term within-individual temporal variability in behaviour as an important topic in its own right (Asendorpf 1990; Bell et al. 2009; Dingemanse et al. 2010; Reale & Dingemanse 2010; Pruitt et al. 2011). For instance, 'repeatability' provides a group-level estimate of the fraction of behavioural variation that is due to interindividual differences in

* Correspondence: J. A. Stamps, Department of Evolution & Ecology, University of California Davis, Davis, CA 95616, U.S.A.
*E-mail address:* jastamps@ucdavis.edu (J. A. Stamps).
[1] *E-mail address:* mark.briffa@plymouth.ac.uk (M. Briffa).
[2] *E-mail address:* pete.biro@deakin.edu.au (P. A. Biro).

0003-3472/$38.00 
doi:10.1016/j.anbehav.2012.02.017

behaviour, such that a high repeatability value occurs when within-individual variance is low relative to between-individual variance (Hayes & Jenkins 1997; Bell et al. 2009). Similarly, 'individual stability' (Asendorpf 1992; Sinn et al. 2008) estimates the extent to which the scores of a single individual change from one observation to the next, and thus provides an estimate of within-individual variability for particular individuals within a group. More recently, Pruitt et al. (2011) estimated within-individual temporal consistency in behaviour by dividing the average variance in the responses of all of the subjects by the variance of the responses expressed by a focal individual across a series of tests.

To date, however, these and other indices of within-individual variability have lumped together two different sources of temporal variability: systematic changes in an individual's behaviour over time and unpredictable variability in an individual's behaviour at a given time. When animals are tested over short intertest intervals, systematic changes in an individual's scores on a behavioural assay can occur as a result of familiar processes such as habituation, sensitization, acclimation or motor/sensory fatigue, whereby an individuals' scores increase or decrease as a function of time or observation number. Alternatively, individual scores may change cyclically over short periods (e.g. as a result of circadian rhythms). Various statistical techniques can be used to model systematic changes in behaviour as a function of time, and such models can provide estimates of the expected scores for each individual at any given time (Hoffman 2007; de Kort et al. 2009; Ram & Gerstorf 2009; Martin et al. 2010). However, even after careful accounting for systematic changes in behaviour over time, considerable variability remains in the behaviour of individuals. These seemingly unpredictable fluctuations in behaviour at a given time have been defined as 'intraindividual variability', or IIV (Nesselroade 1991; Siegler 1994; Salthouse 2007; Ram & Gerstorf 2009).

One might suppose that the variability of an individual's behaviour at a given time might be safely ignored, because it simply reflects random 'noise' that is an unavoidable result of measurement error, uncontrolled variation in external stimuli at the time of testing, or other factors that can affect the scores of animals in experimental settings. Indeed, psychologists used to feel the same way. However, for more than 80 years (Fiske & Rice 1955), they have accumulated evidence that nonsystematic variability in behaviour (i.e. IIV) is not simply random 'noise', but that it varies as a result of age and prior experience, may significantly vary across individuals, and may be stable across time within individuals (Nesselroade 1991; Siegler 2007; Ram & Gerstorf 2009; Salthouse & Nesselroade 2010; also see below).

To appreciate why IIV is important in animals as well as humans, we must first briefly consider how IIV is quantified. When behavioural variation is continuous, IIV indicates the extent to which an individual's scores fluctuate over time, after controlling for any systematic changes in behaviour over time (see Figs 1, 2; Fiske & Rice 1955; Nesselroade 1991; Ram & Gerstorf 2009). One useful approach is to use statistical models to estimate each individual's expected score at each time step or observation number, and then use the deviations of each observed value from its expected value (i.e. the residuals) to estimate IIV (see Figs 1, 2; Hoffman 2007; Hultsch et al. 2008; de Kort et al. 2009; Ram & Gerstorf 2009). This procedure generates an estimate of IIV (residual individual standard deviation, or riSD) that is analogous to the familiar standard deviation statistic, except that it estimates the amount of variation around the expected values at every time step instead of estimating the amount of variation around the average of all of the observations. This approach also illustrates why it is helpful to discriminate systematic from nonsystematic variation in behaviour when studying within-individual temporal variability in behaviour. For instance, a high level of within-individual variability might indicate that an individual's behaviour dramatically changed as a function of time (e.g. its slope was far from zero), or it might indicate that an individual's behaviour was highly unpredictable at any given time (its scores fluctuated widely around its predicted trend line; see Fig. 2).

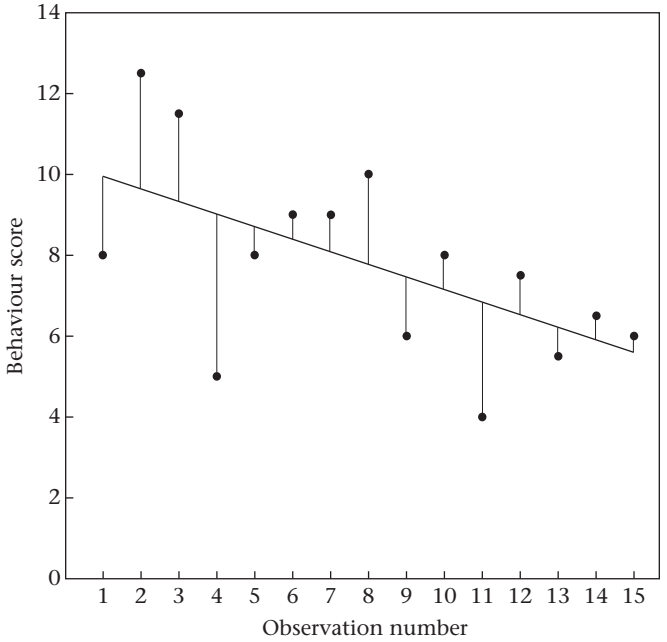


**Figure 1.** Illustration of intraindividual variability (IIV) when behaviour changes systematically over time. Shown is a linear regression fit to the data for a single individual whose behavioural scores decline as a function of observation number. At each observation, the individual's score deviates from its expected score (indicated by the trend line); these deviations are the residuals. These residuals provide an index of IIV, the residual individual standard deviation (riSD).

Intraindividual variability in behaviour can also be quantified when behavioural variation is discrete, using various indices of 'behavioural diversity' (see Fig. 3). The simplest index of behavioural diversity is the number of different behaviour patterns expressed by a given individual (e.g. song repertoire size in birds).

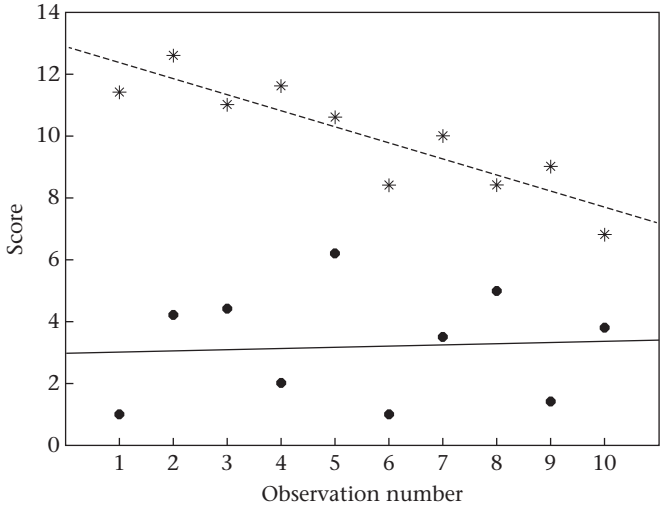


**Figure 2.** Illustration of individual differences in intraindividual variability (IIV) for continuous variation in behaviour. Shown are two individuals with the same within-individual variability (i.e. the same overall standard deviation in their scores). Much of the variability of individual 1 (stars) is due to systematic changes in its behaviour over time, but this is not the case for individual 2 (dots). IIV is substantially higher (predictability is substantially lower) for individual 2 than for individual 1.

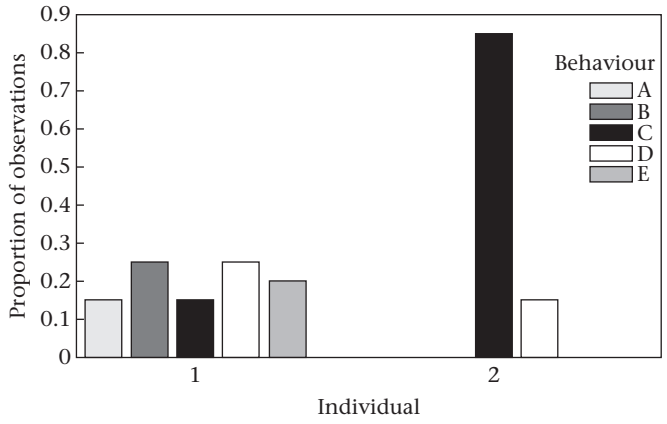


**Figure 3.** Illustration of intraindividual variability (IIV) for discrete behavioural variation. Shown are two animals that differ with respect to the frequency distributions of five different types of behaviour that they might express in a particular context: behavioural diversity is higher (predictably is lower) for individual 1 than for individual 2.

Some authors have advocated using more complicated indices, such as the Shannon–Weaver diversity index ($H$), that consider the relative frequency as well as the number of behaviour patterns expressed by each individual (Schafer 1980; Rosengren & Braswell 2001; Ram & Gerstorf 2009), but thus far, relatively few researchers have adopted this approach (but see Lee et al. 2004).

Finally, to ask whether IIV itself is repeatable over extended periods, or whether it changes as a function of age, life-history stage or season, one can use 'multiple burst' experimental designs. With this design, a set of individuals is repeatedly tested in the same context over a short period (providing an initial set of estimates of IIV), and then this same process is repeated one or more times over longer time intervals (Nesselroade 1991; Salthouse & Nesselroade 2010).

### *Intraindividual Variability: Biological Importance and Statistical Implications*

Using the approaches outlined above, researchers have demonstrated that intraindividual variability is not simply 'random noise', but that it may predictably change as a function of age or experience, differ across individuals, be stable over time, and covary with individual performance. In humans, IIV in many types of behaviour, including personality traits, predictably changes across ontogeny (Rosengren & Braswell 2001; Williams et al. 2005; Noftle & Fleeson 2010). In birds, IIV in song repertoire size (a measure of behavioural diversity) often increases as a function of age, while the IIV of a given song (the inverse of 'song stereotypy') often declines as a function of age (Botero et al. 2009; Byers & Kroodsma 2009; de Kort et al. 2009; Nelson & Poesel 2009; Rivera-Gutierrez et al. 2010). There is also abundant evidence that IIV can be affected by experiences prior to the tests. In humans, the IIV of motor and cognitive skills declines with practise (Cox 1934; Ram et al. 2005), can change in response to reinforcement (Neuringer 2002, 2004), and may increase during the extinction of previously learned responses (Kinloch et al. 2009). Similarly, rodents, pigeons and invertebrates can be trained to either increase or decrease IIV in various types of behaviour (Neuringer 2004; Brembs 2011), and poor nutrition during the nestling phase in birds may reduce song stereotypy (increase IIV) later in life (Searcy et al. 2010). Finally, a few studies of humans indicate that IIV may significantly differ between subjects in the same test (Hoffman 2007) and that these individual differences in IIV may be stable over time (Allaire & Marsiske 2005).

There are also indications that intraindividual variability may be related to individual performance in animals and humans. For instance, larger song repertoire sizes (high IIV) in birds might be advantageous because they help juvenile males better match the songs of established territory owners when they are attempting to establish themselves into a neighbourhood (Nelson & Poesel 2009). Conversely, high song stereotypy (low IIV) may be favoured because female birds might prefer males that sing more consistent songs (Wegrzyn et al. 2010). In humans, individuals with high IIV in response times may experience greater reductions in cognitive function later in life, suggesting that high IIV in response times may reflect underlying deficiencies in neurological function, providing clinicians with a diagnostic tool to predict future cognitive decline (MacDonald et al. 2009; Bielak et al. 2010). In contrast, the diversity of strategies used during problem-solving sessions in children is positively related to future learning (Siegler 2007). These examples also illustrate that the fitness implications of high versus low IIV depend on the behaviour in question (see also Discussion).

Significant individual differences in intraindividual variability are also potentially important for methodological reasons. Currently, most of the statistical models that are used to study individual differences explicitly assume that individual observations are drawn from the same underlying distribution with constant variance (Singer & Willett 2003; Littell et al. 2006; Martin et al. 2010; West et al. 2010; but see Hoffman 2007). For instance, random regression can model changes in behaviour over time within individuals by adding three (co-)variance parameters to the traditional population (mean) level parameters: one that allows the intercept to vary across individuals ($V_I$), a second that allows the slope to vary across individuals ($V_S$), and a third that describes the correlation between predicted intercepts and slopes (the covariance, or cov(IS)). Any nonsystematic within-individual variation that remains (after accounting for covariates) is assumed to arise from a single variance parameter, the residual variance (Vr; Singer & Willett 2003). As a result, the variability around each individual's estimated score for each value of the independent variable (its BLUP, see below) is assumed to be the same for every individual (Littell et al. 2006). Similarly, repeatability in its most familiar form is modelled using ANOVA with individual specified as a fixed factor (Lessells & Boag 1987), and as we all know, ANOVA assumes homogeneity of variance across groups. Although violation of the assumption of equal variance may not have much effect on parameter estimates, unequal variance (heteroscedasticity) is likely to affect the reliability of statistical tests (i.e. $F$ tests, likelihood values, etc.), particularly when means and variances are related in some manner (Singer & Willett 2003; Littell et al. 2006).

Despite these and other reasons why individual differences in IIV are important, this topic has been almost entirely ignored in the field of animal behaviour. Indeed, to our knowledge, no one has tested for individual differences in IIV in any nonhuman animal, although at least one study indicates that such differences are likely (Eriksson et al. 2010). Here, we use data from hermit crabs and Ward's damselfish to test for significant individual differences in IIV in their latency to emerge from a shelter following a disturbance, a common assay of 'boldness'. In the process, we show why the right-censoring of data, which often occurs in studies of latency, can bias estimates of IIV. After demonstrating several-fold, significant individual differences in IIV for both species, we discuss the proximate and ultimate causation of interindividual differences in IIV in these and other animals and suggest potential areas of future research on this topic.

## METHODS

### *General Concerns in Studies of Individual Differences in Intraindividual Variability*

Experimental design and sample size and are the first issues to consider when studying differences between individuals in IIV.

Repeated samples of each individual, conducted in the same context over relatively short periods (e.g. hours to days) are required to obtain reasonable estimates of IIV for each individual, and relatively large numbers of individuals are required to obtain reasonable estimates of the extent to which IIV varies across individuals. Discussion of the sample sizes required for reasonable statistical power for studies of IIV is beyond the scope of the current article (but see Martin et al. 2010). However, here we demonstrate that under strictly controlled laboratory conditions one can detect significant differences between individuals in IIV based on 10 observations per individual, for samples ranging from 22 to 39 individuals.

When behavioural variation is continuous, a second concern in IIV studies is data censoring. Data censoring occurs when an experimental design fails to measure variation in behaviour that lies above or below a given threshold value. Right-censored data is of particular concern when latency scores are used to measure behaviour, because researchers often impose an arbitrary maximum period that they are willing (or able) to observe each subject. In such cases, whenever subjects do not respond prior to the end of that interval they are typically assigned the same right-censored (maximal) score. This practise may lead to biased estimates of IIV at the individual level, since any individual with many right-censored scores will (probably erroneously) appear to have very low IIV. It may also lead to biased estimates of relationships between IIV and mean values at the group level, since individuals with many right-censored scores will also have very high mean values. Finally, it may lead to biased estimates of repeatability, since it erroneously reduces the residual variance.

Although some authors have devised statistical methods to 'correct' censored data after the fact, these methods are based on underlying assumptions about the distributions of the variables of interest (Wang et al. 2009; McBee 2010). Since behavioural variables need not follow particular (e.g. normal) distributions, and because the distributions of behaviour scores can differ across individuals within the same sample (Fleeson 2001), such statistical 'fixes' for censored data are inadvisable. Here, we use data from hermit crabs to show how right-censoring can bias the results of IIV studies, and suggest ways to avoid this problem.

A third important question in studies of intraindividual variability is whether IIV is correlated with mean values across individuals. In humans, relationships between IIV and mean values vary depending on the behaviour, for reasons which are currently obscure. Intraindividual variability in response times are typically positively related to mean response times (Schmiedek et al. 2009), but negative relationships between IIV and mean values have been reported for other types of behaviour. For instance IIV in running speeds is inversely related to mean running speed (i.e. the fastest runners are the least variable) (Hopkins & Hewson 2001), and IIV in a number of cognitive tasks is inversely related to the mean scores on those tasks (Salthouse 2007). Below, we use data from the fish and hermit crabs to show how information about relationships between IIV and mean values can help shed light on proximate factors that might contribute to individual differences in IIV.

Finally, care must be taken in interpreting results based on transformed data when studying intraindividual variability, because transformations have a number of effects on the variability of scores (Asuero & Bueno 2011). In particular, transformations need to be taken into account when investigating relationships between IIV and mean values. For instance, if the standard deviation of a variable $y$ is directly proportional to the mean of $y$, then the standard deviation of ln $y$ will be constant across different values of mean ln $y$ (Asuero & Bueno 2011). This means that if scores are ln-transformed prior to analysis, a correlation of zero between the standard deviation of ln-transformed scores and the mean of ln-transformed scores implies that in the original (untransformed) data set, variability was proportional to the means across the subjects. Since we used ln-transformed latency scores to compute indices of IIV, this issue is directly relevant to the interpretation of our own results.

### *Hermit Crabs*

We measured IIV for latency to emerge from the shell after handling in hermit crabs, using a protocol explicitly designed to avoid censored data. Crabs occupying *Littorina littorea* shells were collected by hand from the intertidal zone at Hannafore Point, Cornwall, U.K. in August 2010, transported back to the laboratory, and held for 48 h in large holding tanks containing filtered and aerated sea water at 15 °C. Crabs were then removed from their gastropod shell, sexed, examined for obvious parasites and loss of appendages and weighed. Only young adult males free from parasites, appendage loss or recent moult were used ($N = 42$; weight range 0.36–1.23 g) and each was supplied with a new shell of the preferred shell weight for the weight of the crab determined using the regression equation from a previous shell selection experiment (Briffa & Elwood 2007). This variation in mass is a small proportion of the total size variation of adult *P. bernhardus*, which ranges from 0.12 g (Briffa et al. 1998) to 65 g (Briffa & Elwood 2005).

Crabs were placed into a 20 cm diameter, white, plastic, flat-bottomed dish containing aerated sea water to a depth of 5 cm, and fed white fish flesh ad libitum throughout the experiment. These dishes were arranged across 1.5 m of bench space over two levels. We measured water temperatures from two dishes at either end, and in the centre, of the bench on both levels. Temperatures ranged from 14.7 to 15.4 °C (mean = 15.01, coefficient of variation, CV = 0.02); ANOVA indicated no significant spatial differences in temperature across the bench (extreme left, middle or extreme right: $F_{2,6} = 2.45$, $P = 0.374$) or across shelves ($F_{1,6} = 1.56$, $P = 0.26$), with no significant interaction effect ($F_{2,6} = 0.67$, $P = 0.55$).

Forty-eight hours later, we obtained a startle response for each individual. Each crab was removed from its dish by hand, held in an inverted position for 5 s and replaced in the dish in the inverted position. This causes crabs to withdraw into their shell. The duration of the startle response was timed to within 0.1 s, from the time when the crab was returned to the dish until the crab emerged from the shell and made contact with the base of the dish with its legs and first attempted to right its position (Briffa et al. 2008). Each individual was observed until it emerged from its shell, so none of the data were right-censored. The minimum score was 0.7 s, so none of the data were left-censored. For each crab, one startle response was evoked every second day between 0900 and 1100 hours, and the order of observations was randomized across individuals. Ten startle responses per individual were observed over 20 days. Water was changed following observations (using pre-aerated sea water at 15 °C), and uneaten food was replaced with fresh food. Three crabs were removed from the experiment because they moulted, reducing our sample to 39 individuals.

To determine how right-censoring might have affected our results, we constructed an artificially censored data set in which we assigned a score of 30 s to each observation in which a crab had not emerged from its shell 0.5 min after being returned to its dish. This data set (censored crabs) was analysed using the same methods described below for the original data set. We have included a copy of the crab latency data as Supplementary Material (SAS program file and crab data), for readers who wish to explore other effects of data censoring on studies of IIV.

No permits or animal care protocols were required for these experiments, but they conformed to the ABS/ASAB ethical guidelines for the treatment of animals in research. None of the crabs

were harmed when they were removed from their original shells, and the crabs were returned to the sea in shells of the optimal size following the experiment.

*Damselfish*

Data for analyses of IIV in latency to emerge from shelter following a disturbance came from a previous study on juvenile Ward's damselfish (*P. wardi*), where this same species was incorrectly first identified as *P. bankanensis* (Biro et al. 2010). In brief, fish were captured in light traps in November 2008 while in the process of metamorphosis from the pelagic larval stage to the benthic juvenile stage. All of the fish were approximately the same age (22–25 days old, based on otolith daily increments; P. A. Biro, unpublished data), and none of them had any experience with a benthic environment prior to transfer to the temperature-controlled laboratory. We selected 30 fish of similar length, placed each into its own home tank with a section of plastic pipe for shelter, fed them ad libitum with live *Artemia* food, and changed 80% of the water every second day with fresh sea water adjusted to the ambient water temperature (29.0 °C). Tank-specific temperature was measured after each set of observations in the morning and afternoon.

Behaviour data were collected twice a day (1000–1200 hours and 1400–1700 hours), over a 1-week period. The data used here represent the first week of data collected by Biro et al. (2010), following a period of acclimation to the experimental protocols (we omitted the first three observations, as was also done in our earlier publication). Latency to emerge from shelter was measured for each fish following a simulated predation event, where the handle of an aquarium dip net was rapidly plunged into the centre of the aquarium. In every case, this resulted in the fish immediately taking refuge inside of the plastic tubing.

Because this study was not originally designed to study IIV, the latency data were right-censored: fish were observed to a maximum of 180 s each, but on a number of occasions, the fish had not yet emerged from their shelter by the end of this period. To reduce the effects of right-censoring on our results, we excluded any individual for which 40% or more of its scores were 180 s, leaving 22 fish for analysis.

These experiments were conducted under permits from the Great Barrier Reef Marine Park Authority (permit number G07/21313.1) and the James Cook University Animal Ethics Committee (permit number A1123), and they conformed to the ABS/ASAB ethical guidelines for the treatment of animals in research.

*Statistical Analyses*

Initial examination of data plots indicated that the latency scores of many individuals in both species systematically increased or decreased as a function of observation number, and that these relationships differed across individuals. Therefore, we used both simple linear regression and random regression methods to model systematic (linear) changes in behaviour over time within individuals. We used separate regressions for each individual because this method makes no assumptions about homogeneity of variance across individuals. We also used random regression because this method is often used to model changes in scores over time, is more parsimonious, and can incorporate covariates that might affect the behaviour in question. In addition, by comparing the estimates of IIV generated by the separate regression and random regression methods (see Results), we were able to demonstrate that some potential errors that can arise when random regression is used to generate individual-specific predictions (Hadfield et al. 2010) were not an issue in the current study.

Those initial data plots also indicated nonlinear relationships between latency scores and time for many individuals, and that the latency data were characterized by many low but relatively few large values, and were thus perhaps lognormally distributed. Thus, for both the separate regression and random regression models (see details below) we compared the fits of models based on $\ln(x+1)$ transformed scores to those based on the raw latency scores. The model fits were comparable for both models for some individuals, but ln-transformation linearized the data and normalized the residuals for many individuals in both the crab and the fish data sets. Thus, we used the ln-transformed latency scores in our analyses.

We used two methods to model systematic changes in behaviour over time for each individual. First, we fitted a separate regression model to each individual (model: $\ln(\text{latency}+1) = a + b(\text{observation number})$), and then used the expected values generated by this model to calculate riSD, as described below. Similarly, we used random regression models to generate expected values for each observation period (using the BLUPs, see below), and then used these expected values to calculate riSDs for each subject.

Random regression (RR) models were implemented using Proc Mixed (SAS Institute, Cary, NC, U.S.A.). We tested for individual variation in the intercepts and/or slopes of the relationship between ln-latency and observation number, by specifying the intercept and observation number as random effects, with any remaining factors treated as fixed effects (Singer & Willett 2003; West et al. 2010). RR generates predictions for individual-specific intercepts and slopes (BLUPs), yielding individual-specific predictions of behaviour at a given time or observation number.

We started model fitting with a saturated model that included three random effects (the two variance parameters $V_I$ and $V_S$, along with a covariance parameter cov(IS) describing the correlation between them) and any fixed effects (the 'covariates') (Singer & Willett 2003; see Supplementary Material for model code). When covariates (e.g. temperature, size) were not significant ($P > 0.2$), they were removed from the model before refitting (Crawley 2005). We used the Kenward–Roger method to calculate degrees of freedom for the fixed effects, using a type III approach (Littell et al. 2006). After removing any covariates from the model, we next assessed fit of the random effects using Akaike's Information Criterion values, corrected for small samples (AICc), comparing model fit with and without a random intercept effect, and a model with random intercept to a model with random slope effect and covariance parameter. In this approach, smaller values indicate better fit, and when two competing models differ by less than a few AICc units (typically $\Delta$ AICc$< 4$), they are considered equally likely, and therefore not 'significantly' different (Burnham & Anderson 1998).

Next, we used the residuals generated from the models above to calculate an index of IIV, the residual individual standard deviation (riSD; Hultsch et al. 2008; Ram & Gerstorf 2009). Calculating riSD is straightforward. For each individual i, and each observation j, one enters the observed score $Y_{ij}$ and the expected score generated by the regression model, $E_{ij}$, into a spreadsheet. Then one computes the variance for each individual ($s^2$) as the sum of the $(Y_{ij} - E_{ij})^2$, divided by $N_i - 1$, where $N_i$ is the number of observations for that individual. Finally, riSD is obtained as the square root of the variance. Note that this procedure is comparable to the procedure used to compute a standard deviation, except that deviations from the expected values for each observation ($E_{ij}$) are used instead of deviations from the expected (average) value across all of the observations.

To test whether IIV differed across individuals, we used two familiar statistical tests to determine whether the variance, $(\text{riSD})^2$, differed across individuals. First, we used the $F$ test for unequal variance, which is calculated as the ratio of the largest to the smallest

variance among groups (here, individuals). Although easily implemented, this test is sensitive to departures from normality. Second, we used the more robust Levene's test of inequality of variances, which produces a statistic, $W$, which is then compared against an $F$ distribution (Crawley 2005). We slightly modified Levene's test to suit our question, because in our data, the expected scores varied as a function of observation number, whereas traditional versions of Levene's test rely upon a single expected value (typically the mean or the median) for each group (here, individuals) in the analysis (see Appendix). For each data set, we calculated Levene's $W$ using a spreadsheet because the statistical applications we used to model linear changes in behaviour over time did not allow for tests of homogeneity of residual variance at the individual level.

For those interested in additional details, we have provided two files as Supplementary Material ('SAS program file and crab data', 'Excel sheet for calculating IIV and modified Levene's test'). The first file is an annotated SAS program file that generates expected values for each observation period for each individual and plots the observed and expected values at each observation period for each individual. The second file illustrates how one can use these residuals to compute the IIV of each subject and test for individual differences in IIV.

Finally, to obtain a measure of within-individual variation that incorporated both systematic and nonsystematic variation in behaviour and that was not affected by transformation, we calculated the standard deviation of the raw latency scores for each individual, hereafter termed the individual standard deviation (iSD).

## RESULTS

### *Comparing a Separate Regression Approach with Random Regression*

For both the fish and crabs, we compared the riSD values (based on ln-transformed latency scores) generated by the random regression model with the riSD values generated by separate regressions. Correlations between the two estimates of riSD were high for the crabs (Pearson correlation: $r_{37} = 0.997$, $P < 0.0001$) and for the fish ($r_{20} = 0.89$, $P < 0.0001$). One fish (fish 28) was poorly fit by the random regression model and was well fit by separate regression; with this individual removed, fit was substantially improved ($r_{19} = 0.96$, $P < 0.0001$). In addition, for all three data sets, when the estimate of riSD generated by random regression was regressed against the estimate of riSD generated by separate regression, the slopes did not significantly differ from 1.0 ($t = -0.18$ to 0.77, $P = 0.45$–$0.85$), indicating that the two methods generated comparable estimates of IIV for each subject. Since the random regression model and the separate regression models produced the same qualitative results, we focus on the more informative and parsimonious random regression models for the remainder of this article.

### *Hermit Crabs*

For the crabs, ln-latency to emerge was independent of mass ($F_{1,37} = 0.3$, $P = 0.58$). On average, ln-latencies increased over time ($F_{1,37} = 4.5$, $P < 0.045$), suggesting sensitization rather than habituation. Results also suggested that changes in behaviour as a function of time differed across individuals, not only with respect to their intercepts ($\Delta$AICc = 237), but also with respect to their slopes ($\Delta$AICc = 17; see Fig. 4). Following Martin et al. (2010), we calculated repeatability of latency for different observation numbers using the variance, co-covariance and residual variance parameters. We found that repeatability was 0.63 for observation 1 and 0.76 for observation 10.

Using the residuals from the random regression analysis, we found a 12.8-fold difference in IIV across individuals as measured by riSD (mean = 0.57, range 0.13–1.67). A simple $F$ test comparing the ratio of the largest to the smallest residual individual variance across individuals indicated that IIV differed significantly across crabs ($F_{9,9} = 164$, $P < 0.0001$). This result was also supported by Levene's test, which indicated significant differences in IIV across individuals ($W_{38,351} = 4.43$, $P < 0.0001$).

There was also a weak, positive relationship between mean ln-latency and riSD across individuals ($F_{1,37} = 5.8$, $r = 0.37$, $P < 0.03$) indicating that crabs with high mean latencies were even more variable than one would expect if variability in latency had increased as a constant proportion of mean latencies in this species (Asuero & Bueno 2011).

To explore the effect of data censuring on these results, we repeated these analyses using the artificially right-censored data set (see Methods). In the censored data set, there was a 10.5-fold difference in riSD across crabs (mean = 0.44 s, range 0.08–0.84 s). In this case, both the $F$ test ($F_{9,9} = 95$, $P < 0.0001$) and the Levene's test ($W_{38,351} = 3.38$, $P < 0.0001$) again indicated significant differences in IIV across individuals. However, with these censored data there was no longer any discernable relationship between mean ln-latency and riSD across individuals ($F_{1,37} = 2.6$, $r = 0.24$, $P = 0.12$). Hence, if we had relied on censored data for the crabs, we would have underestimated the extent of individual differences in IIV in this species, and failed to detect the positive relationship between mean ln-latency and riSD in these animals.

Finally, for each crab we computed its standard deviation (iSD) for latency, a simple index of within-individual variability. There was a 231-fold difference in values of iSD across individuals (range 1.6–369 s). Hence, controlling for systematic temporal changes in behaviour and for the reductions in variability that resulted from ln-transformation dramatically reduced but did not eliminate individual differences in behavioural variability.

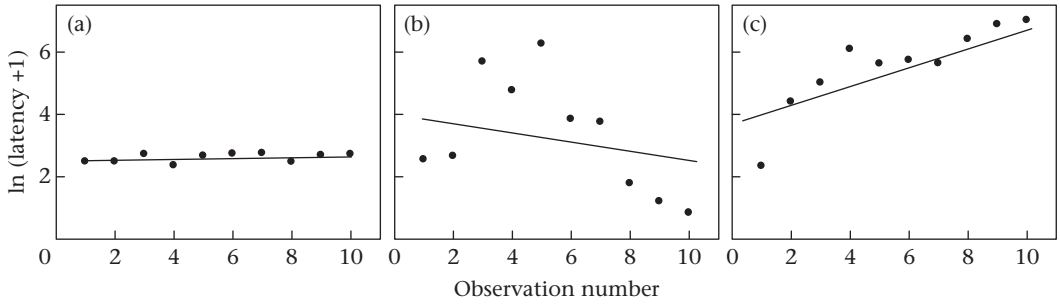


**Figure 4.** Latency to emerge after a disturbance in hermit crabs as a function of observation number. Shown are the linear model fits to the data for individuals that displayed (a) the lowest IIV, (b) the highest IIV and (c) the greatest change in behaviour as a function of time.

*Damselfish*

For the 22 fish whose scores were least affected by right-censored data, ln-latency to emerge from shelter was not influenced by the time of day that the trial was conducted ($P = 0.43$), or by fish size ($P = 0.437$). On average, ln-latencies decreased as a function of observation number ($F_{1,20} = 9.7$, $P < 0.0005$), suggesting habituation; latency may also have been weakly related to water temperature (mean = 29.0 °C, range 25.8–29.5 °C) at the time of the test ($F_{1,194} = 3.6$, $P = 0.06$). After controlling for this possible effect of temperature on latency, results suggested that individuals varied with respect to the intercepts (ΔAICc = 64) and slopes (ΔAICc = 7) of their relationships between latency and observation number (see Fig. 5). Repeatability of latency was estimated to be 0.10 on day 1, and 0.54 on day 10.

There was 3.6-fold variation in riSD across individuals (mean = 0.80 s, range 0.39–1.39 s). As with the crabs, both the *F* test ($F_{8,9} = 12.6$, $P < 0.0005$) and the Levene's test ($W_{21,190} = 2.35$, $P < 0.005$) indicated significant differences in IIV across individuals. In contrast to the crabs, we did not detect a significant positive relationship across individuals between mean ln-latency and riSD ($F_{1,20} = 0.35$, $r = 0.13$, $P = 0.56$). However, these findings should be viewed with caution, since in order to avoid relying on right-censored data we omitted the individuals with the highest mean scores from this analysis.

There was a 49-fold difference across individuals in iSD's (range 1.5–73 s), based on the raw latency scores, a measure of within-individual variability. Thus, as was observed in the crabs, accounting for systematic temporal changes in behaviour and the effects of ln-transformation on variability reduced, but did not eliminate, significant individual differences in behavioural variability.

## DISCUSSION

In aquatic species from two different phyla (hermit crabs and damselfish), intraindividual variability in latency to emerge from a refuge following a disturbance significantly differed across individuals. Significant differences in IIV were detected even after we statistically controlled for systematic changes in behaviour as a function of time, the effects of data transformation on variability, and potentially confounding covariates. Significant differences in IIV across individuals indicate that the latency of an individual cannot be accurately summarized by a single statistic (e.g. its mean value). Instead, individual differences in IIV imply that the expected behaviour of a given individual at a given time is better expressed by a distribution of values that is unique to that individual, and that the scores of some individuals are more widely dispersed around their mean values than are those of others (see also Fleeson 2001; Schmiedek et al. 2009).

Our results also underscore the importance of avoiding using censored data in studies of intraindividual variability. When we artificially censored the hermit crab data set, we subsequently underestimated the extent of individual differences in IIV, and failed to detect the positive correlation between IIV (based on ln-transformed scores) and mean ln-latency in the original data set. Of course, the best way to avoid potential effects of censored data on IIV is to avoid data censoring when designing an experiment. If this is impractical, then an alternative is to omit individuals with a high proportion of censored scores from the analysis, as we did here for the damselfish. However, this approach not only reduces the sample size, but does so nonrandomly, by omitting individuals with the most extreme mean scores (individuals with many maximal scores, in the case of right-censored data, or minimal scores in the case of left-censored data). Currently it is unclear how best to handle censored data in studies of IIV, making this a topic worthy of further study.

Our results suggest that some of the observed differences between individuals in IIV might be a result of constraints in the ability of the subjects to estimate the duration of time intervals. We found that the variability in an individual's raw latency scores increased either as a constant (fish), or as an accelerating (crabs) function of its mean latency. One possible explanation for these positive relationships is that, in both experiments, the subjects were attempting to estimate the duration of a time interval (from a disturbance until it was 'safe' to emerge from shelter). Several studies show that when animals are trained to perceive or remember time intervals, the variability of their estimates increases as a function of the duration being timed (Gibbon 1977; Gallistel & Gibbon 2000; Shettleworth 2010).

However, constraints in the ability of animals to estimate time intervals cannot account for all of the interindividual differences in IIV observed in the current study, because significant individual differences in IIV were observed even after the raw latency scores were ln-transformed, a process that statistically controls for situations in which the standard deviation is directly proportional to the mean. Individual differences in the IIV at a given mean latency might be due to differences between individuals in fluctuations in their internal state or in their sensitivity to subtle fluctuations in external stimuli. Although most studies of IIV to date have focused on the former (Brembs 2011), evidence is mounting that individuals differ in their sensitivity to a variety of stimuli (Aron & Aron 1997; Ellis et al. 2006; Sih & Bell 2008; Biro et al. 2010; Dingemanse et al. 2010; Stamps & Groothuis 2010a). Despite researchers' best efforts to replicate the same context (exactly the same set of external stimuli) in every test, external stimuli may still vary from test to test in subtle, uncontrolled ways that affect the behaviour of a portion of the individuals in the sample. In that situation, individuals with higher contextual plasticity (sensu Stamps & Groothuis 2010a) would show more variable behaviour across a series of tests than individuals with lower contextual plasticity. This hypothesis might

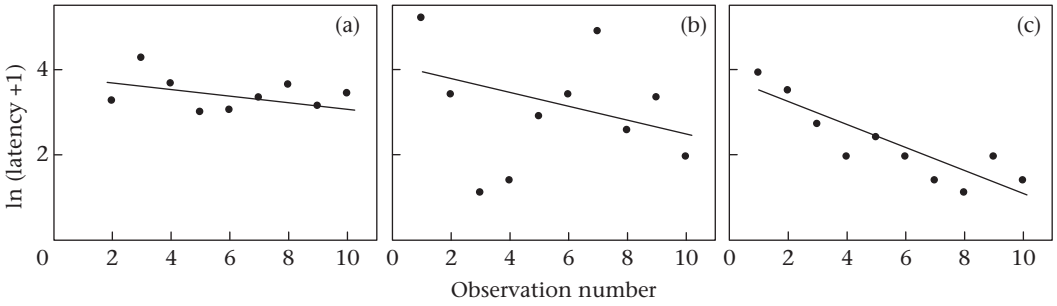


**Figure 5.** Latency to emerge after a disturbance in damselfish as a function of observation number. Shown are the linear model fits to the data for the individuals that displayed (a) the lowest IIV, (b) the highest IIV and (c) the greatest change in behaviour as a function of time.

be tested by measuring the external stimuli during each test more precisely than is usually the case (see Biro et al. 2010), to see whether individuals with higher IIV in an earlier experiment were, in fact, responding to subtle changes in external stimuli that had no discernable effect on the behaviour of individuals with lower IIV.

Our findings of significant individual differences in IIV raise a host of interesting questions about this phenomenon. For instance, it would be interesting to know whether the individual differences in IIV that we detected in hermit crabs and damselfish are consistent across time or contexts. Using multiple burst designs (see Introduction), one could determine whether individual differences in IIV observed at a particular age or time were maintained across longer time intervals (e.g. across different ages, seasons, or life-history stages). Similarly, one could study different behaviours at the same age, and determine whether individual differences in IIV were consistent across contexts (e.g. whether individuals with high IIV in latencies to emerge from shelter also show high IIV in latencies to attack a conspecific opponent). These sorts of analyses might help tell us whether IIV in a given behaviour might be viewed as a trait in its own right, and whether some individuals are generally more unpredictable than others. Spiders, *Latrodectus hesperus*, might prove a useful candidate for this sort of study, given recent indications that within-individual variability in habitat choice in this species is correlated across contexts and affected by food deprivation (Pruitt et al. 2011).

In turn, information about the phenomenology and proximate causation of IIV for a given type of behaviour may help shed light on its functional significance. For instance, if further experiments revealed that near-misses with a natural predator increased IIV in emergence latencies in crabs or fish, we might ask whether and why high IIV might reduce their risk of predation. One possibility is that, in nature, individuals might repeatedly encounter the same individual predators, in which case, variable (unpredictable) emergence timing might help prevent a predator from predicting that prey's behaviour. There are hints that high IIV might be advantageous in other situations in which dyads have a series of win–lose encounters with one another. For instance, coaches of football teams pseudorandomly alternate passing and running plays within games, presumably to prevent their opponent from predicting their next play (McGarrity & Linnen 2010). Alternatively, if IIV in emergence latencies turned out to be correlated across individuals with IIV in other types of behaviour that involve time measurement (e.g. IIV in learned tasks that involve timing; Shettleworth 2010), then we might ask about the fitness implications of variation among individuals in their tendency to accurately measure time intervals.

Individual differences in intraindividual variability also have important implications for the methods currently used to study individual differences in behavioural and labile physiological traits. First, if average IIV is even moderately high in relation to interindividual variability, then many replicate tests of every individual will be required to obtain reasonable estimates of the mean values of the individuals in a sample. Sizeable individual differences in IIV only exacerbate this problem, because adequate sample sizes for measuring mean values will vary across individuals in the same sample. In turn, inaccurate estimates of the mean scores for a given behaviour reduce the chances of detecting significant relationships across individuals between those mean scores and any other variable of interest (e.g. their scores for a different behaviour, a physiological trait, or a component of fitness).

In addition, as noted in the Introduction, virtually all of the statistical models currently used to study individual differences in labile behavioural or physiological traits assume that IIV is the same for all of the individuals in the sample. Our findings that IIV in latency scores varied 13-fold in hermit crabs, even after controlling for systematic changes in behaviour over time and after ln-transformation, suggests that individual differences in IIV are not trivial, and raises the question of whether and how this degree of variation among individuals in IIV would affect the results of these statistical models. Violation of this assumption did not seem to affect the estimates of IIV generated by our random regression models, since the values of riSD generated using that method were virtually identical to those generated by the separate regression method. These results also imply that the random and separate regression methods generated comparable estimates of the slope and intercept of each of our subjects. However, we have no way to tell whether our tests for individual differences in slopes or intercepts, or our inferences about other parameters generated by the random regression models, were affected by violation of this assumption. If future studies detect individual differences in IIV as sizeable as those in the current study, then it might be useful to examine current statistical models to determine whether their results are robust to violations of the assumption of equal IIV.

## Acknowledgments

We are very grateful to David Warton for statistical help and advice, and thank Mark McCormick and the staff at Lizard Island Research Station for their assistance. P.A.B. was supported by an ARC Future Fellowship.

## Supplementary Material

Supplementary material for this article is available, in the online version, at doi:10.1016/j.anbehav.2012.02.017.

## References

**Allaire, J. C. & Marsiske, M.** 2005. Intraindividual variability may not always indicate vulnerability in elders' cognitive performance. *Psychology and Aging*, **20**, 390–401.

**Aron, E. N. & Aron, A.** 1997. Sensory-processing sensitivity and its relation to introversion and emotionality. *Journal of Personality and Social Psychology*, **73**, 345–368.

**Asendorpf, J. B.** 1990. The measurement of individual consistency. *Methodika*, **4**, 1–23.

**Asendorpf, J. B.** 1992. Beyond stability: predicting interindividual differences in intraindividual change. *European Journal of Personality*, **6**, 103–117.

**Asuero, A. G. & Bueno, J. M.** 2011. Fitting straight lines with replicated observations by linear regression. IV. Transforming data. *Critical Reviews in Analytical Chemistry*, **41**, 36–69.

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

**Bielak, A. A. M., Hultsch, D. F., Strauss, E., MacDonald, S. W. S. & Hunter, M. A.** 2010. Intraindividual variability is related to cognitive change in older adults: evidence for within-person coupling. *Psychology and Aging*, **25**, 575–586.

**Biro, P. A., Beckmann, C. & Stamps, J. A.** 2010. Small within-day increases in temperature affects boldness and alters personality in coral reef fish. *Proceedings of the Royal Society B-Biological Sciences*, **277**, 71–77.

**Botero, C. A., Rossman, R. J., Caro, L. M., Stenzler, L. M., Lovette, I. J., de Kort, S. R. & Vehrencamp, S. L.** 2009. Syllable type consistency is related to age, social status and reproductive success in the tropical mockingbird. *Animal Behaviour*, **77**, 701–706.

**Brembs, B.** 2011. Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. *Proceedings of the Royal Society B*, **278**, 930–939.

**Briffa, M. & Elwood, R. W.** 2005. Metabolic consequences of shell choice in *Pagurus bernhardus*: do hermit crabs prefer cryptic or portable shells? *Behavioral Ecology and Sociobiology*, **59**, 143–148.

**Briffa, M. & Elwood, R. W.** 2007. Monoamines and decision making during contests in the hermit crab *Pagurus bernhardus*. *Animal Behaviour*, **73**, 605–612.

**Briffa, M., Elwood, R. W. & Dick, J. T. A.** 1998. Analysis of repeated signals during shell fights in the hermit crab *Pagurus bernhardus*. *Proceedings of the Royal Society B*, **265**, 1467–1474.

**Briffa, M., Rundle, S. D. & Fryer, A.** 2008. Comparing the strength of behavioural plasticity and consistency across situations: animal personalities in the hermit crab *Pagurus bernhardus*. *Proceedings of the Royal Society B*, **275**, 1305–1311.

**Brown, M. B. & Forsythe, A. B.** 1974. Robust tests for equality of variances. *Journal of the American Statistical Association*, **69**, 364–367.

**Burnham, K. P. & Anderson, D. R.** 1998. *Model Selection and Inference: a Practical Information-theoretic Approach*. New York: Springer-Verlag.
**Byers, B. E. & Kroodsma, D. E.** 2009. Female mate choice and songbird song repertoires. *Animal Behaviour*, **77**, 13–22.
**Cox, J. W.** 1934. *Manual Skill: Its Organization and Development*. New York: Cambridge University Press.
**Crawley, M. J.** 2005. *Statistics: an Introduction Using R*. New York: J. Wiley.
**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**, 81–89.
**Ellis, B. J., Jackson, J. J. & Boyce, W. T.** 2006. The stress response systems: universality and adaptive individual differences. *Developmental Review*, **26**, 175–212.
**Eriksson, C. A., Booth, D. J. & Biro, P. A.** 2010. 'Personality' in two species of temperate damselfish. *Marine Ecology Progress Series*, **420**, 273–276.
**Fiske, D. W. & Rice, L.** 1955. Intra-individual response variability. *Psychological Bulletin*, **52**, 217–250.
**Fleeson, W.** 2001. Toward a structure- and process-integrated view of personality: traits as density distributions of states. *Journal of Personality and Social Psychology*, **80**, 1011–1027.
**Gallistel, C. R. & Gibbon, J.** 2000. Time, rate and conditioning. *Psychological Review*, **107**, 289–344.
**Gibbon, J.** 1977. Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, **84**, 279–325.
**Hadfield, J. D., Wilson, A. J., Garant, D., Sheldon, B. C. & Kruuk, L. E. B.** 2010. The misuse of BLUP in ecology and evolution. *American Naturalist*, **175**, 116–125.
**Hayes, J. P. & Jenkins, S. H.** 1997. Individual variation in mammals. *Journal of Mammalogy*, **78**, 274–293.
**Hoffman, L.** 2007. Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research*, **42**, 609–629.
**Hopkins, W. G. & Hewson, D. J.** 2001. Variability of competitive performance of distance runners. *Medicine and Science in Sports and Exercise*, **33**, 1588–1592.
**Hultsch, D. F., Strauss, E., Hunter, M. A. & MacDonald, S. W. S.** 2008. Intraindividual variability, cognition and aging. In: *The Handbook of Aging and Cognition* (Ed. by F. I. M. Craik & T. A. Salthouse), pp. 491–556. New York: Psychology Press.
**Kinloch, J. M., Foster, T. M. & McEwan, J. S. A.** 2009. Extinction-induced variability in human behavior. *Psychological Record*, **59**, 347–369.
**de Kort, S. R., Eldermire, E. R. B., Valderrama, S., Botero, C. A. & Vehrencamp, S. L.** 2009. Trill consistency is an age-related assessment signal in banded wrens. *Proceedings of the Royal Society B*, **276**, 2315–2321.
**Lee, D., Conroy, M., McGreevy, B. & Barraclough, D.** 2004. Reinforcement learning and decision making in monkeys during a competitive game. *Cognitive Brain Research*, **22**, 45–58.
**Lessells, C. M. & Boag, P. T.** 1987. Unrepeatable repeatabilities: a common mistake. *Auk*, **104**, 116–121.
**Littell, R. C., Milliken, G., Stroup, W., Wolfinger, R. & Schabenberger, O.** 2006. *SAS for Mixed Models*. 2nd edn. Cary, North Carolina: SAS Institute.
**McBee, M.** 2010. Modeling outcomes with floor or ceiling effects: an introduction to the Tobit model. *Gifted Child Quarterly*, **54**, 314–320.
**MacDonald, S. W. S., Li, S.-C. & Bäckman, L.** 2009. Neural underpinnings of within-person variability in cognitive functioning. *Psychology and Aging*, **24**, 792–808.
**McGarrity, J. P. & Linnen, B.** 2010. Pass or run: an empirical test of the matching pennies game using data from the National Football League. *Southern Economic Journal*, **76**, 791–810.
**Martin, J. G. A., Nussey, D. H., Wilson, A. J. & Réale, D.** 2010. Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. *Methods in Ecology and Evolution*, **2**, 362–374.
**Nelson, D. A. & Poesel, A.** 2009. Does learning produce song conformity or novelty in white-crowned sparrows, *Zonotrichia leucophrys*? *Animal Behaviour*, **78**, 433–440.
**Nesselroade, J. R.** 1991. The warp and woof of the developmental fabric. In: *Visions of Development, the Environment, and Aesthetics: the Legacy of Joachim F. Wohlwill* (Ed. by R. Downs, L. Liben & D. Palermo), pp. 213–240. Hillsdale, New Jersey: L. Erlbaum.
**Neuringer, A.** 2002. Operant variability: evidence, functions and theory. *Psychonomic Bulletin & Review*, **9**, 672–705.
**Neuringer, A.** 2004. Reinforced variability in animals and people: implications for adaptive action. *American Psychologist*, **59**, 891–906.
**Noftle, E. E. & Fleeson, W.** 2010. Age differences in big five behavior averages and variabilities across the adult life span: moving beyond retrospective, global summary accounts of personality. *Psychology and Aging*, **25**, 95–107.
**Pruitt, J. N., DiRienzo, N., Kralj-Fiser, S., Johnson, J. C. & Sih, A.** 2011. Individual- and condition-dependent effects on habitat choice and choosiness. *Behavioral Ecology and Sociobiology*, **65**, 1987–1995.
**Ram, N. & Gerstorf, D.** 2009. Time-structured and net intraindividual variability: tools for examining the development of dynamic characteristics and processes. *Psychology and Aging*, **24**, 778–791.
**Ram, N., Rabbitt, P., Stollery, B. & Nesselroade, J. R.** 2005. Cognitive performance inconsistency: intraindividual change and variability. *Psychology and Aging*, **20**, 623–633.
**Reale, D. & Dingemanse, N. J.** 2010. Personality and individual specialization. In: *Social Behavior: Genes, Ecology and Evolution* (Ed. by T. Szekely, A. Moore & J. Komdeur), pp. 417–441. Cambridge: Cambridge University Press.
**Rivera-Gutierrez, H. F., Pinxten, R. & Eens, M.** 2010. Multiple signals for multiple messages: great tit, *Parus major*, song signals age and survival. *Animal Behaviour*, **80**, 451–459.
**Rosengren, K. S. & Braswell, G. S.** 2001. Variability in children's reasoning. In: *Advances in Child Development and Behavior* (Ed. by H. W. Reese & R. Kail), pp. 1–40. New York: Academic Press.
**Salthouse, T. A.** 2007. Implications of within-person variability in cognitive and neuropsychological functioning for the interpretation of change. *Neuropsychology*, **21**, 401–411.
**Salthouse, T. A. & Nesselroade, J. R.** 2010. Dealing with short-term fluctuation in longitudinal research. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences*, **65**, 698–705.
**Schafer, W. D.** 1980. Assessment of dispersion in categorical data. *Educational and Psychological Measurement*, **40**, 879–883.
**Schmiedek, F., Lövdén, M. & Lindenberger, U.** 2009. On the relation of mean reaction time and intraindividual reaction time variability. *Psychology and Aging*, **24**, 841–857.
**Searcy, W. A., Peters, S., Kipper, S. & Nowicki, S.** 2010. Female response to song reflects male developmental history in swamp sparrows. *Behavioral Ecology and Sociobiology*, **64**, 1343–1349.
**Shettleworth, S.** 2010. *Cognition, Evolution and Behavior*. 2nd edn. New York: Oxford University Press.
**Siegler, R. S.** 1994. Cognitive variability: a key to understanding cognitive development. *Current Directions in Psychological Science*, **3**, 1–5.
**Siegler, R. S.** 2007. Cognitive variability. *Developmental Science*, **10**, 104–109.
**Sih, A. & Bell, A. M.** 2008. Insights for behavioral ecology from behavioral syndromes. *Advances in the Study of Behavior*, **38**, 227–281.
**Singer, J. D. & Willett, J. B.** 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.
**Sinn, D. L., Gosling, S. D. & Moltschaniwskyj, N. A.** 2008. Development of shy/bold behaviour in squid: context-specific phenotypes associated with developmental plasticity. *Animal Behaviour*, **75**, 433–442.
**Stamps, J. & Groothuis, T. G. G.** 2010a. The development of animal personality: relevance, concepts and perspectives. *Biological Reviews*, **85**, 301–325.
**Stamps, J. & Groothuis, T. G. G.** 2010b. Developmental perspectives on personality: implications for ecological and evolutionary studies of individual differences. *Philosophical Transactions of the Royal Society B*, **365**, 4029–4041.
**Wang, L. J., Zhang, Z. Y., McArdle, J. J. & Salthouse, T. A.** 2009. Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research*, **43**, 476–496.
**Wegrzyn, E., Leniowski, K. & Osiejuk, T. S.** 2010. Whistle duration and consistency reflect philopatry and harem size in great reed warblers. *Animal Behaviour*, **79**, 1363–1372.
**West, S. G., Ryu, E., Kwok, O. M. & Cham, H.** 2010. Multilevel modeling: current and future applications in personality research. *Journal of Personality*, **79**, 2–50.
**Williams, B. R., Hultsch, D. F., Strauss, E. H., Hunter, M. A. & Tannock, R.** 2005. Inconsistency in reaction time across the life span. *Neuropsychology*, **19**, 88–96.

## Appendix

### *Levene's Test for Differences in Intraindividual Variability (IIV) across Individuals*

We used a modified version of Levene's test, because traditional versions of Levene's test rely on a single estimate of the expected score for each group (typically the mean or median value for each group), to compute the test statistic *W* (Brown & Forsythe 1974). However, if an individual's scores systematically change as a function of time or observation number, the expected scores for that individual will vary as a function of time or observation number. To accommodate systematic changes in an individual's behaviour as a function of time, we computed *W* based on the expected scores for each individual i at observation j ($E_{ij}$) generated by our regression models. For each individual and each observation period, we entered the observed score ($Y_{ij}$) and the expected score ($E_{ij}$) into a spreadsheet. Then, $Z_{ij} = |Y_{ij} - E_{ij}|$. The rest of the calculations were the same as those used in standard versions of the Levene's test.

The Levene's test statistic, *W*, is defined as follows:

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2}$$

When $W$ is used to test for variation in IIV across individuals, $k$ is the number of different individuals, $N$ is the total number of observations, $N_i$ is the number of observations for the ith individual, $Y_{ij}$ is the observed value for the jth observation of the ith individual, $Z_{ij} = |Y_{ij} - E_{ij}|$, where $E_{ij}$ is the expected value for the jth observation of the ith individual:

$$Z_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} Z_{ij} \text{ is the mean of all } Z_{ij}.$$

$$Z_{i.} = \frac{1}{N_i}\sum_{j=1}^{N_i} Z_{ij} \text{ is the mean of the } Z_{ij} \text{ for individual i.}$$

The significance of $W$ is tested against $F(\alpha, k-1, N-k)$ where $F$ is a quantile of the $F$ test distribution, with $k-1$ and $N-k$ degrees of freedom, and $\alpha$ is the chosen level of significance.