Research regarding among-individual variation, typically under the label of ‘animal personality’, often makes use of the estimation of repeatability for inference. This usage makes sense: variance standardized ratios like repeatability allow for straightforward interpretations. For example, repeatabilities greater than 0.5 demonstrate that most of the observed variation in behaviour is due to differences among individuals. However, behavioural researchers are interested in a variety of questions vis-à-vis ‘personality’ and, for some of these questions, the inferences that can be drawn from repeatability will be limited (Wilson, 2018). Here we discuss these limitations and additional metrics that may be useful to researchers.

Repeatability, defined as the among-individual variance ($V_I$) divided by the total variance ($V_P$; Boake, 1989; Dingemansen & Dochtermann, 2013; Hayes & Jenkins, 1997; Nakagawa & Schielzeth, 2010), is what is known as a variance standardized measure (Houle, 1992). Variance standardized measures have limitations of which potential users should be aware. First, comparing these measures between sexes, among populations, taxa, trait types, etc., is problematic. This is because these ratios can differ between groups because of differences in either the numerator or denominator. Whether repeatability is found to be similar or different among groups is therefore ultimately uninformative without considering the specific changes in variance components.

Second, repeatability is often discussed as a metric of ‘individual consistency’ (de Villemereuil, Morrissey, Nakagawa, & Schielzeth, 2018). Unfortunately consistency on its own is difficult to interpret and it is instead consistency relative to the mean that will often be more biologically important. Put another way, how variable individuals are only makes sense on the scale of the behaviour of interest. In this regard, the inferential limitations for the use of repeatability are similar to those associated with heritability. For example, as detailed by Houle (1992) and Hansen, Pélabon, and Houle (2011), standardizing additive genetic variances ($V_A$) by total phenotypic variances ($V_P$)—that is, calculating narrow-sense heritability—is not appropriate for most of the questions researchers ask. This assertion is based on the observation that most of the variation expressed in a trait can be genetically based (i.e. high heritability) even when there is little total variation in that trait. In such a case the observed genetic differences result in only small differences in trait values. Similarly, for repeatability, individuals can exhibit high ‘consistency’, and thus high repeatability, but actually vary little on the scale of the traits of interest. Therefore, repeatability says nothing about whether individuals exhibit large differences in phenotype: individuals can exhibit substantial repeatability but ultimately differ little from one another (e.g. Fig. 1c versus Fig. 1d). Instead, whether repeatability—which combines additive genetic variation, nonadditive genetic variation and permanent environmental effects (Boake, 1989; Dochtermann,
represents biologically relevant variation depends on the mean of a trait and how individuals vary around that mean. One alternative is the use of mean standardization rather than variance standardization. Mean standardization places the observed magnitude of variation directly in the context of the mean. One mean-standardization approach is to divide the among-individual variation ($V_I$) in a trait by the square of its mean (for applications of this approach to understanding heritability see Hansen et al., 2011; Houle, 1992; also see Fig. 2c):

$$I = 100 \times \frac{V_I}{\bar{x}^2}$$

where $\bar{x}$ is the mean of the population.

This mean standardization can also be applied to any other estimate of variance, including residual variances (e.g. $V_I$ for $V_R$; following Houle, 1992; Hansen et al., 2011) and variances due to fixed effects ($I_p$ for $V_p$), with $V_I$ calculated following de Villemereuil et al., 2018; Nakagawa & Schielzeth, 2013). Mean standardization allows direct comparison of magnitudes of variation across traits measured on different scales between groups (Fig. 2). In contrast, repeatability only indirectly provides an indication of relative magnitude. There are other dimensionless mean standardizations that similarly allow the direct comparison of magnitudes, including the coefficient of variation (CV, $\frac{V_I}{\bar{x}}$), which is likely more familiar than $I$ (see also Holtmann, Lagisz, & Nakagawa, 2017). We prefer the use of $I$ over CV and other measures because the former is more directly linked to understanding evolutionary responses to selection (Hansen et al., 2011; Houle, 1992). For example, when calculated for additive genetic variance, $I$ estimates the percentage of change in a trait’s mean under selection (Hansen et al., 2011). By using $I$, the study of individual variation in behaviour can therefore be more directly linked to the broader evolutionary literature. Regardless of whether $I$ or CV is preferred, we encourage authors to report all the unstandardized variance components that their statistical models estimate, along with behavioural means, so as to facilitate future meta-analyses.

Mean standardization provides several benefits that complement those of repeatability. Most notably, mean standardization allows variation to be compared across scales of measurement and traits (e.g. time allocated to parental care versus propensity to engage in agonistic interactions), without the assumptions necessary for doing so with repeatability. Mean standardization allows for different biological inferences to be made than can be made based on variance standardization alone. For example, in regards again to heritability, it is a general prediction that traits that are closely linked to fitness will exhibit lower variation and this is generally borne out when examining heritabilities (Mousseau & Roff, 1987; Stirling, Réale, & Roff, 2002). However, when this question is asked using mean standardization it becomes clear that there is considerable additive genetic variation in fitness affecting traits and that the observation of low heritabilities is due to there also being considerable environmental variation in these traits.

Figure 1. Populations can have the same repeatability yet have drastically different magnitudes of among-individual variance relative to their means. For example, two populations of unicorns simulated by Wilson (2008) had the same average aggression and the same repeatability despite different amounts of variability (a, b). This can be more clearly observed when aggression of 100 randomly selected individuals from each population is measured three times (c, d). The comparison of individuals from Population 1 (c) and Population 2 (d) demonstrates that, despite the equal repeatability, unicorns in Population 2 exhibited much larger differences from each other than did unicorns in Population 1.
Similarly, behavioural traits linked to mating, habitat selection and aggression typically show high repeatability, whereas activity, mate preference and migration have lower repeatability (Bell, Hankson, & Laskowski, 2009). However, because mean standardized ratios have not typically been reported in behavioural studies, it is not possible to determine how these results translate to magnitudes of behavioural differences among individuals. Whether traits differ in the magnitude of individual differences consequently remains an important open topic for behavioural ecologists to tackle. More generally, the use of mean standardization allows researchers to ask new and exciting questions about how different types of behaviours differ and the magnitude of effects on behavioural variation.

As an example of how mean standardization complements repeatability, we can revisit data collected (simulated) by Wilson (2008) for a variety of morphological and behavioural traits of unicorns (see also de Villemereuil et al., 2018). Two populations of unicorns exhibited the same average aggression and the same repeatability of aggression ($r = 0.80$; Fig. 1). However, unicorns in Population 1 exhibited less variability both among individuals and across repeated measurements of the same individuals than observed for Population 2 (Fig. 1c and d). This difference in variability is apparent when directly comparing variance components (Fig. 2a) but obscured under variance standardization (Fig. 2b). The use of $I$ values recaptures this difference between the two populations at both the among-individual and within-individual levels (Fig. 2c). The use of $I$ therefore highlights that while the unicorns of both populations were equally aggressive and exhibited the same repeatability, individual unicorns in Population 2 differed from each other to a greater degree than did those in Population 1 (Fig. 2c). This is an inference that could not have been drawn from the comparison of repeatabilities and suggests interesting future questions regarding the contribution of genetic and environmental factors to aggression in unicorns.

There are, of course, caveats to the use of mean standardization. First, it only applies to traits on ratio or log scales (although the evolutionary implications discussed above and by Hansen et al. (2011) are not applicable on log scales). Ratio scale does not refer to values expressed as ratios but rather to whether how something has an objective zero value and if the comparison of two measurements of the same kind estimates the ratio of the difference in their magnitude (Hansen et al., 2011; Houle, 1992). This issue of a trait having a real zero is particularly important as researchers frequently standardize measures based on their standard deviations and centred to their mean prior to analysis (e.g., Hadfield, 2010). This approach can facilitate statistical model fitting but prohibits mean standardization and thus limits the inferences that can be made. Finally, while mean standardization can be conducted on the data scale for non-normally distributed traits following de Villemereuil, Schielzeth, Nakagawa, and Morrissey (2016), its interpretation in such cases is not clear.
The use of variance standardized measures like repeatabilities versus mean standardization has been discussed a great deal in the evolutionary literature, as have the points that we make above (e.g. Hansen et al., 2011; Hansen & Houle, 2008; Holtmann et al., 2017; Houle, 1992). However, given the manner in which repeatability is frequently discussed in animal behaviour studies, these points require continued emphasis (see also Wilson, 2018). Importantly, we are not suggesting that behavioural ecologists stop using repeatability, but rather want to emphasize that variance ratios like repeatabilities and mean standardized estimates allow for different inferences. The study of ‘animal personality’ and behavioural variation continues to ask increasingly sophisticated questions regarding when and why behavioural variation is manifested. This expanding range of questions necessitates an expanded quantitative toolbox as not all questions will be properly addressed with repeatability. Mean standardization is one such tool. Which tool should be used fundamentally comes down to the questions being asked.

Thus, given their differences, what questions are appropriate for available approaches? Variance ratios like repeatability allow inferences regarding sources of variation while mean standardized estimates allow inferences regarding the magnitude of variation (Table 1; see also de Villemereuil et al., 2018; Nakagawa & Schielzeth, 2010; Wilson, 2018). For example, if a researcher is interested in whether most variation among individuals is attributable to developmental differences, comparing variance ratios and direct comparison of variances is most appropriate (e.g. Royauté & Dochtermann, 2017; Royauté, Garrison, Dalos, Berdal, & Dochtermann, 2019; Rudin, Simmons, & Tomkins, 2018). If, instead, a researcher is interested in whether individuals exhibit large differences in their behaviours, mean standardization will be most appropriate. Researchers should carefully consider their questions when drawing inferences from these estimates. We encourage researchers to present the most complete information possible. This complete information includes not only derived values like repeatability and mean standardized estimates but their constituent components as well.

Acknowledgments

We thank A. Wilson and the authors of de Villemereuil et al. (2018) for helpful discussions. Two anonymous referees provided valuable feedback that improved the readability and clarity of the paper. We also thank A. Wilson for providing the unicorn data from Wilson (2008). N.A.D. and R.R. were funded by the National Science Foundation (NSF IOS 1557951).

Table 1

<table>
<thead>
<tr>
<th>Metric</th>
<th>Interpretation</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variance standardized measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted repeatability</td>
<td>$\frac{V_I}{V_I + V_F}$</td>
<td>Proportion of variation due to intrinsic differences among individuals (i.e. due to genetic and permanent environmental effects)</td>
</tr>
<tr>
<td>Unadjusted repeatability</td>
<td>$\frac{V_I}{V_I + V_F}$</td>
<td>Proportion of total variation that is attributable to differences among individuals conditioned on measurable, fixed effects ($V_F$). Fixed effects might include either procedural/experimental controls or biological</td>
</tr>
<tr>
<td><strong>Mean standardized measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean-scaled individual variation ($I_I$)</td>
<td>$100 \times \frac{V_I}{\sigma_X^2}$</td>
<td>Percentage of variation relative to the mean explained by differences in trait expression among individuals</td>
</tr>
<tr>
<td>Mean-scaled fixed effect variation ($I_F$)</td>
<td>$100 \times \frac{V_F}{\sigma_X^2}$</td>
<td>Percentage of variation relative to the mean explained by differences in trait expression due to fixed effects</td>
</tr>
<tr>
<td>Mean-scaled residual variation ($I_R$)</td>
<td>$100 \times \frac{V_R}{\sigma_X^2}$</td>
<td>Percentage of variation relative to the mean explained by differences in trait expression due to unmeasured sources of variation</td>
</tr>
</tbody>
</table>

**References**