Understanding different types of repeatability and intra-class correlation for an analysis of biological variation

Shinichi Nakagawa[1,2*], David F. Westneat [1,3], Ayumi Mizuno [1,2], Yimen G. Araya-Ajoy [1,4#], Ned A. Dochtermann [1,5#], Kate L. Laskowski [1,6#], Joel L. Pick [1,7#], Denis Réale [1,8#], Coralie Williams [1,9#], Jonathan Wright [1,4#], and Holger Schielzeth [1,10*]

[1] The Statistical Quantification of Individual Differences (SQuID) Working Group.

[2] Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada

[3] Department of Biology, University of Kentucky, Lexington, KY 40506-0225, USA

[4] Department of Biology, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway

[5] Department of Biological Sciences, North Dakota State University, Fargo ND 58104 USA

[6] Department of Evolution & Ecology, University of California Davis, Davis CA 95616 USA

[7] Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK

[8] Departement des Sciences Biologiques, Université du Québec à Montréal, CP 8888, succursale centre-ville, Montréal, Canada, QC, H3C3P8

[9] School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia

[10] Institute of Ecology and Evolution, Friedrich Schiller University Jena, Jena, Germany

#These authors contributed equally and are listed in alphabetical order

* Equal contributions

*Author for correspondence: S Nakagawa (shinichi.nakagawa@ualberta.ca) and H Schielzeth (holger.schielzeth@uni-jena.de)

Running headline: Flexible estimation of repeatabilities

Data accessibility. No empirical data were used in this study

Authors' contributions. Conceptualization: All, Project administration: SN. Validation: All, Visualization: AM, Writing - Original Draft: SN, HS, DFW, Writing - Review & Editing: All.

Competing interests. We declare we have no competing interests.

Funding. This project was made possible through the grant "Statistical Quantification of Individual Differences (SQuID)" from the Research Council of Norway INTPART project, grant number 309356, to JW. SN was supported by the Australian Research Council Discovery Grant (DP230101248) and the Canada Excellence Research Chair Program (CERC-2022-00074). DFW was supported by a UNSW Visiting Fellowship during the early development of this project. HS was supported by the German Research Foundation (DFG, grant 316099922).

Abstract

Repeatability (more generally known as intraclass correlation) represents an important quantity of interest in many scientific fields. It represents a metric for summarizing variance decomposition to identify sources of variation in an outcome of interest (e.g. organismal traits). The estimation of variance components is often achieved through linear mixed-effect models or their extension, generalized linear mixed-effect models. Here, we review variants of calculating repeatabilities from mixed-effects models for a variety of conditions and applications. We also recommend which variant might be appropriate under what conditions, focusing on behavioural biology/ecology examples. However, the decision is ultimately with the researcher, since it depends upon their research question, and there is no one-size-fits-all solution. We also highlight the importance of the scope of inference, which affects how repeatabilities are used and interpreted. We recommend transparent reporting of statistical results, including all variance components, which are the building blocks of repeatability. This review aims to assist empiricists in choosing an appropriate repeatability variant and interpretation concerning their questions and scope of inference.

Keywords: variance partitioning coefficients, intra-class correlation, mixed-effects modelling, individual differences, repeatability, variance components

1 Introduction

Understanding sources of organismal variation is a central goal in biology. This goal can be achieved by statistical variance decomposition if observations are replicated within a group of interests [1-3]. The choice of the grouping of interests depends on the subdiscipline; for instance, evolutionary biologists are often interested in among-family or among-genotype variance, behavioural ecologists in among-individual variance, community ecologists in among-plot variance, and molecular biologists sometimes in among-strain variance. All disciplines may be interested in measurement errors and variance within and among observers or techniques. The metrics used to partition variances and compare them thus can be critical to advancing a subfield's research agenda.

A popular and efficient tool for estimating variance components is the linear mixed-effect model (LMM) [1-4]. LMMs (and their extension, generalized linear mixed-effects models, GLMMs) allow estimation of variance explained by random effects, where random effects represent groups (or clusters) of observations, and they use the similarity of observations within groups, plus the variation in expected values across groups, to estimate those variance components. However, variances have some properties that create some challenges (cf. [5]). For example, the magnitude of variance components depends on how the outcome (e.g., an organismal trait) is measured, since variances are quantified in squared units of the measurement scale of the target outcome variable. Different scales make it challenging to compare variance components across studies.

A popular approach to this problem is to present variance components relative to the total outcome variance, and such variance-standardized variance components are here called

repeatabilities. Repeatabilities can be interpreted as the expected population-level correlation among outcomes (observations) from the same group and thus represent a case of intra-class correlations, ICCs [2]. We note that repeatabilities also represent a case of variance partition coefficients, VPCs (for more on the difference between ICC and VPC, see [6]). Thus, repeatability is an extremely useful metric in many areas of biology. Because it can be used to address a variety of goals, however, various ways to calculate repeatability have arisen, which can complicate comparisons among studies. Here, we organize types of repeatability and review both their methods of calculation and the questions they can and cannot address. We note that these same concepts also underpin heritability in quantitative genetics [4], where proportionate variance attributable to additive genetic effects parallels the logic of partitioning variance among groups (although we do not discuss heritability in much detail, our arguments on repeatability are directly applicable to heritability) (cf. [7]).

The original application of repeatability was by engineers who were interested in the reproducibility of particular outcomes [8], like the accuracy by which a machine might fill specific amounts of liquid into a bottle or produce a screw that fits a specific nut. More broadly, repeatabilities can be used to separate the among-group variance of interest from the sources of variance within groups. For objects that do not change, such as the width of a screw, the within-object variance represents measurement errors. For objects that do change, such as a young bird undergoing development, the within-object variance represents more than measurement errors—it can be a process that alters values within the group, as if a machine was sensitive to temperature. Biologically, the within-object variance includes plasticity in response to unmeasured environmental variables and other hidden biological processes (e.g., [9-11]). The among-object variance takes on new meaning in this case because it represents consistent

differences in average group (cluster) outcomes relative to the total outcome variance in the population, thus about relative, not absolute, consistency.

In behavioural biology/ecology, repeatability has become the primary metric for quantifying the relative phenotypic consistency of individuals within a population. Repeatability (R) ranges from $R = 0$, meaning there is no consistent difference between groups (i.e., each group exhibits the same range of outcomes as the total population), up to $R = 1$, meaning absolute consistency (all observations within a group are identical but groups differ). Realistically, behavioural repeatabilities often lie between these extremes, typically from near zero up to roughly 0.8, with a meta-analytically estimated average of around 0.37 [12]. A key feature of repeatability is that it is a proportion with a variance component of interest (e.g., among-individual variance) in the numerator and some measure of total variance in the denominator. Because multiple variance components can contribute to the denominator (and sometimes even the numerator), researchers may define "repeatability" differently. For instance, some studies exclude measurement error variance, while others might include additional sources of variation in their denominators. Consequently, comparing published repeatability estimates can be intricate, since not everyone calculates them in the same manner. Moreover, the concept of repeatability can address a range of questions, such as the stability of a trait over time or the influence of measurement conditions, so different "flavours" of repeatability may be relevant to different contexts. Ultimately, the choice of which variance components appear in the numerator versus the denominator depends upon the specific research question. To help clarify these distinctions, we present a taxonomy of repeatability calculations, focusing on the variants that are most commonly encountered. Before delving into these types, we begin by defining the standard or "ordinary" repeatability as it is routinely calculated.

1.1 Ordinary repeatability

The definition of (ordinary) repeatability is:

$$R = \frac{V_G}{V_G+V_W} = \frac{V_G}{V_O}, \quad (1.1)$$

where V_G represents the variance between groups, V_W represents the variance within groups, and V_O represents the total variance in the outcome (often, this is called the phenotypic variance in behavioural and evolutionary biology/ecology; [13]). For this paper, we refer to the V_G as the focal variance component of interest. Furthermore, we call the specific instances of the grouping factor "objects" (these groups are often individuals in behavioural ecology or psychology [14]) and the particular outcome measures "observations". We use "population of interest" in a biological sense, meaning the set of organisms or objects about which we make inferences

Ordinary repeatability can be estimated from a mixed-effects model with a single grouping factor and a residual variance (here, assuming Gaussian distributions, but see for generalizations to other distributions in section 5). The statistical model can be expressed as (Model 1):

$$y_{ij} = \beta_0 + \alpha_i + \varepsilon_{ij}, \quad (1.2)$$

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad (1.3)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad (1.4)$$

where y_{ij} represents the *j*th observation of the *i*th object, β_0 represents the population-level intercept (a fixed effect), α_i represents the object-specific deviation (often called a 'random

intercept') and ε_{ij} is the residual term. In this setup, σ_{α}^2 is the among-object variance (the same as V_G) and σ_{ε}^2 is the residual variance (analogous to V_W although they are not the same beucase σ_{ε}^2 could include variance due to measurement error, for example, but σ_{ε}^2 and V_W often are assumed to be the same). The repeatability from Model 1 is therefore:

$$R = \frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2+\sigma_{\varepsilon}^2}, \quad (1.5)$$

which parallels Equation 1.1. Because this model has a single grouping factor and a single residual term, the resulting *R* is straightforward, hence the term 'ordinary' repeatability. Here, if the outcome is truly constant for each object (i.e., no within-individual plasticity), then 1 – *R* can be interpreted as a measure of measurement error. If the outcome is not strictly fixed (e.g., it changes across time or environments), then 1 – *R* also encompasses within-individual variability (such as phenotypic plasticity).

However, most datasets and questions in biology are more complex, often involving multiple sources of variation that affect the outcome. This leads to multiple ways to calculate repeatability, depending on the conceptual question (which variance components are of interest) and the technical aspects of the statistical model (e.g., multiple random factors, fixed effects, or non-Gaussian data). Various labels have been proposed for these alternative implementations of repeatability. In the sections that follow, we structure these alternatives and illustrate how they can offer different biological insights depending on the outcome variable, data structure, and research goals.

2 More complex random-intercept models

2.1 Multiple random effects

Real datasets often have a more complex hierarchical structure than expressed in Model 1, a structure with multiple grouping levels. This simultaneously increases the options for calculating repeatability and narrows the interpretation of a specific calculation. First, the complexity will have to be reflected in the statistical model so that we have a mixed-effect model with multiple random effects. A mixed-effect model with two random effects can be expressed as (Model 2):

$$y_{hij} = \beta_0 + \gamma_h + \alpha_i + \varepsilon_{hij}, \quad (2.1)$$

$$\gamma_h \sim N(0, \sigma_\gamma^2), \quad (2.2)$$

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad (2.3)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad (2.4)$$

where γ_h represents the deviation for the *h*th level of a second grouping factor (fitted as "random intercepts"), assumed to be normally distributed with mean 0 and population-level variance σ_γ^2. For example, γ_h might capture variation due to different years or different observers while α_i captures variation among individuals. The other terms remain as in Equation (1.2).

Suppose our main interest is still the repeatability at the original group level α. Under Model 2, there are three ways to calculate repeatability, relating to how the random effects (α_i and γ_h) might be nested or crossed and how they combine to form total variance (Figure 1).

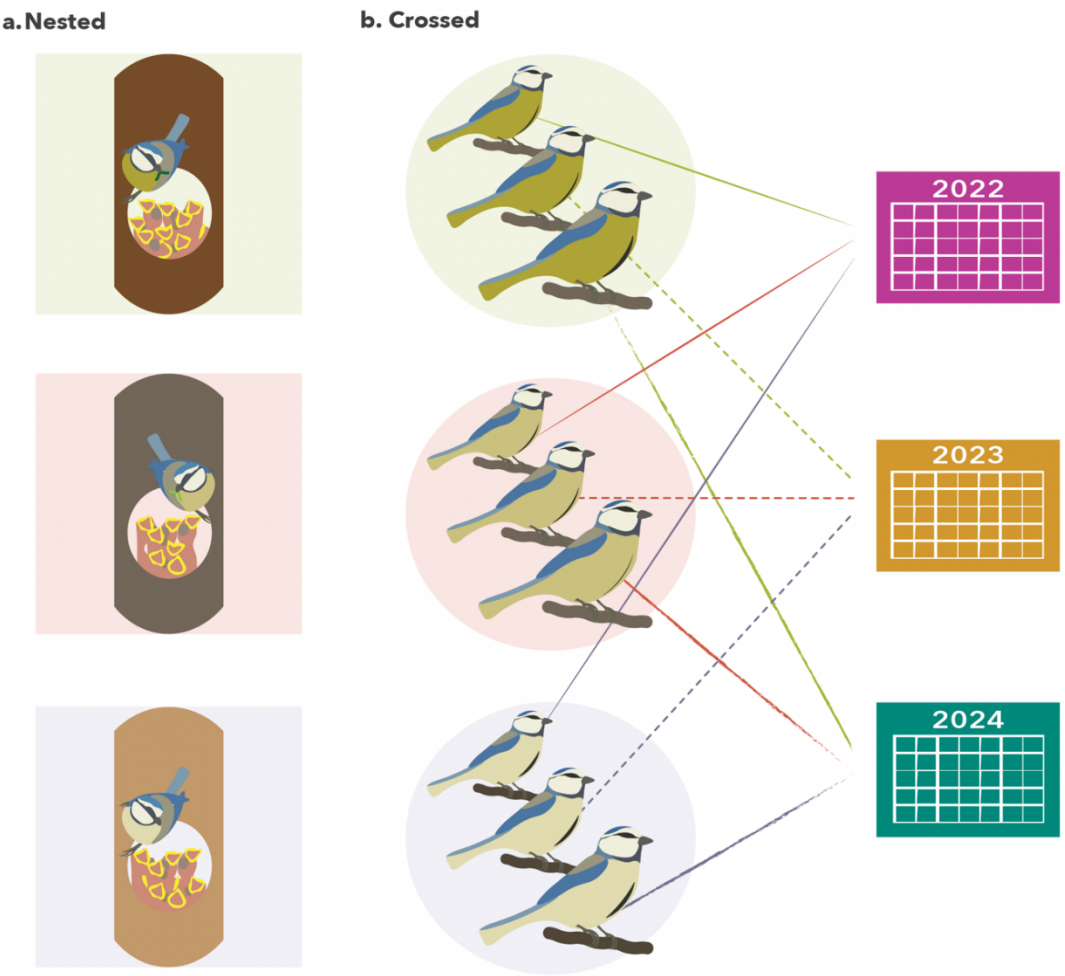


Figure 1. A conceptual example diagram of two types of pairs of random effects: a) Nested random effects, such as nestlings within a nest (and that nest is associated with a particular mother). Here, the mother identity varies at a higher level than nestling identities, creating a strictly nested hierarchical structure. b) Crossed random effects, such as multiple individual birds measured across different years. In a fully crossed design, each bird could be measured in each year (and each year contains multiple birds), although real data might only partially cross these factors if not all birds are observed yearly. The random effects included in a model should reflect the sampling design used to collect data, and it is possible to have multiple nested effects, multiple crossed effects, or a combination of both.

Sometimes, the second random effect of a model captures measurement errors rather than intrinsic biological variation. For example, if multiple observers or instruments were used, then the variability attributable to that factor may be considered 'design-induced'. In such cases, one may exclude the second random effect's variance from the denominator. We call the resulting measure adjusted repeatability, R_{v1}, which is calculated (under Model 2) as [2]:

$$R_{v1} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\varepsilon^2}. \quad (2.5)$$

Note that omitting σ_γ^2 means the denominator of R_{v1}, no longer reflects the total measured variance. However, if γ is indeed an experimental artefact (e.g., observer identity) rather than meaningful biological variation, R_{v1}, can approximate the 'true' proportion of variance that is genuinely biological..

In other instances, the second random effect represents a real biological process, such as multiple years or environmental conditions (Figure 1). If the goal is to express the variance of the focal grouping factor (α) as a fraction of the total outcome variance, the second factor's variance (σ_γ^2) should remain in the denominator. This approach leads to unadjusted repeatability, R_{v2}:

$$R_{v2} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\gamma^2+\sigma_\varepsilon^2}. \quad (2.6)$$

Conceptually, R_{v2} resembles 'ordinary' repeatability (Equation 1.5), except it arises from a more complex dataset and includes σ_γ^2 in the denominator. This version of repeatability is suitable when γ_h is crossed or nested below α_i or simply whenever both factors are legitimate contributors to the total variance, one aims to partition.

Finally, if the two random effects are strictly nested, meaning γ_h lies at a higher level than α_i, one may opt to include both σ_α^2 and σ_γ^2 in both the numerator and the denominator. We refer to this as multi-cluster repeatability, R_{v3}:

$$R_{v3} = \frac{\sigma_\alpha^2+\sigma_\gamma^2}{\sigma_\alpha^2+\sigma_\gamma^2+\sigma_\varepsilon^2}. \quad (2.7)$$

We recommend R_{v3} only when the nested structure is unambiguous and α is truly a sub-level of γ_h. An example is offspring nested within parents (e.g., [15, 16]). If nestling variance (α_i) and parent variance (γ_h) both describe the among-group component of interest, including them together in the numerator and denominator (Equation 2.7) quantifies the overall repeatability of individuals, including "family" effects. Indeed, if one ignores parents entirely and fits a simpler Model 1, the results will approximate R_{v3}. However, if the researcher's focal question is the repeatability across parents, treating nestling variance as part of the denominator, then R_{v2} (Equation 2.6) is more appropriate.

Crossed designs also arise when, for example, individuals are measured repeatedly across years (Figure 1b). In that scenario, the variance for year does not feed into the among-individual variance; rather, the two factors are independent. Hence, multi-cluster repeatability, R_{v3}, is inappropriate. The choice between adjusted or unadjusted repeatability (R_{v1} vs. R_{v2}) depends upon whether 'year' is a nuisance factor or a meaningful contributor to the total variance. In many studies, it may also be valuable to calculate both variants to show how results change when controlling for year versus treating year as part of the phenotypic background.

2.2 Single fixed effects

Most biological studies have broader goals than simply partitioning variance among random effects. Researchers often incorporate treatments or covariates as fixed effects, reflecting hypotheses about how these factors influence the outcome. Accordingly, the resulting models are termed 'mixed-effect' models because they include both random and fixed components. For instance, consider a relatively simple model with one fixed effect (Model 3):

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \alpha_i + \varepsilon_{ij}, \quad (2.8)$$

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad (2.9)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad (2.10)$$

where β_1 is the population-level regression coefficient for a covariate x_{1ij} and other terms follow Model 1. Because model output typically does not provide a direct variance estimate associated with the fixed effect term ($\beta_1 x_{1ij}$), one must custom calculate it (e.g., following [17]):

$$\sigma_F^2 = Var(\beta_1 x_{1ij}) \quad (2.11)$$

or

$$\sigma_F^2 = \beta_1^2 \cdot Var(x_{1ij}), \quad (2.12)$$

where $Var(x_{1ij})$ represents the variance in the covariate or predictor. With this additional variance component σ_F^2, as with the random effects above (Section 2.1), we can calculate three repeatabilities, which differ in whether or not the fixed effect variance σ_F^2 is included in the denominator and, if so, whether it is also included in the numerator.

Suppose a dataset comes from two labs analysing the same subject strains. Although 'lab' might conceivably be modelled as a random effect, having only two levels often leads us to treat it as a fixed effect. The variance associated with labs, however, is likely an artefact of the conditions and measures of those labs. Accordingly, we can calculate adjusted repeatability by omitting lab-induced variance from the denominator, just as in Equation 2.5. In this particular case, we can use our earlier, simple calculation of repeatability:

$$R_{v1} = \frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2+\sigma_{\varepsilon}^2}, \qquad (2.13)$$

note that now the denominator no longer represents the total variance. Many published repeatabilities in mixed-effect models with fixed effects are of this adjusted kind [2, 12]. However, this is frequently an unintended outcome because standard software packages typically only provide random-effect variances by default, thereby excluding fixed-effect variances from the denominator, often with no explicit decision by the researcher.

When one's goal is simply to remove design artefacts (e.g., lab differences) from the total variance, R_{v1} can be appropriate. But if a fixed effect represents a genuinely biological process (e.g., a temperature gradient), subtracting its variance from the denominator does yield a biased estimate of population-level repeatability. Instead, it only gives the repeatability at a single reference value (e.g., the mean temperature or one specific lab). Hence, removing biologically relevant sources of variation understates the total phenotypic variance.

Most fixed effects in statistical models capture processes that affect trait variation. If a factor or covariate varies within the focal grouping, it typically pulls variance from the residual term, in

which case repeatability should not be adjusted; the fixed-effect variance then belongs to the denominator. An unadjusted repeatability that includes the fixed-effect variance, R_{v2}, might be calculated as:

$$R_{v2} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_F^2+\sigma_\varepsilon^2}. \quad (2.14)$$

As in Equation 2.6, $\sigma_\alpha^2 + \sigma_F^2 + \sigma_\varepsilon^2$ represents an estimate of the total outcome variance. This is useful if we are interested in putting variance components in perspective of the total phenotypic variance [7].

In some situations, a fixed effect is itself a group-level attribute (e.g., the altitude of each plot). In that case, adjusting for altitude would partially reduce the among-plot variance (σ_α^2), thus understating the natural variation among plots. One may instead calculate a repeatability that restores this fixed-effect variance to the numerator and the denominator, for instance:

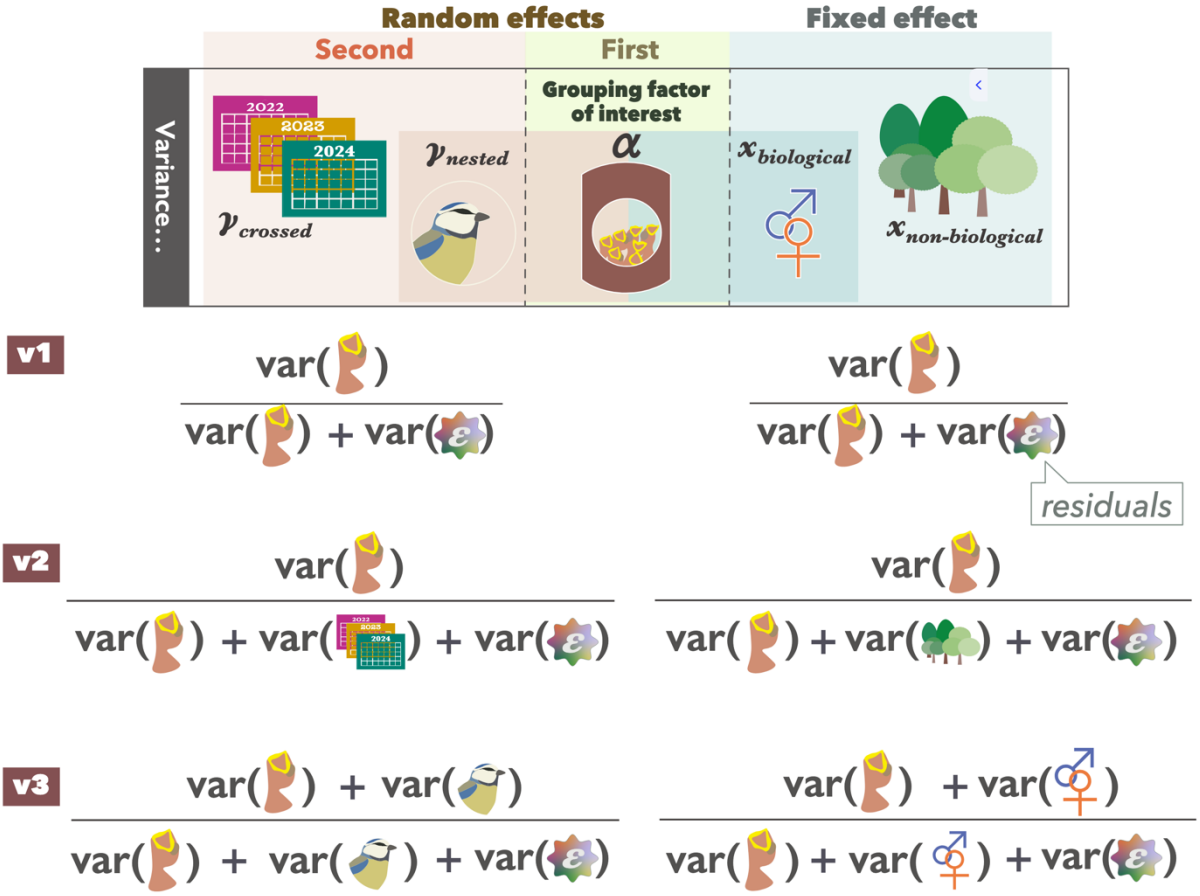
$$R_{v3} = \frac{\sigma_\alpha^2+\sigma_F^2}{\sigma_\alpha^2+\sigma_F^2+\sigma_\varepsilon^2}, \quad (2.15)$$

which has previously been called “enhanced” repeatability [3]. We note that this is equivalent to ordinary repeatability, if we had omitted the altitude of the plot in the model.

As seen in Sections 2.1 and 2.2, the term ‘adjusted’ repeatability (Equations 2.5 and 2.13) does not indicate which sources of variance were excluded. Complex fixed-effect structures—especially those that vary at multiple levels—can yield blends of adjusted and enhanced repeatabilities. It is therefore crucial for authors to explicitly state how they compute repeatability if the goal is to allow cross-study comparisons. In behavioural and evolutionary

research, non-biological factors (e.g., instruments, lab IDs) can often be omitted, whereas biological factors (e.g., temperature, altitude, body size) should generally remain in the denominator (and sometimes the numerator) unless one is specifically interested in the value of repeatability at a reference setting.

For example, if one measures individual birds in two forests [18], using R_{v1} might make sense if the two-forest difference is truly an experimental artefact (e.g., forced sampling from both sites). Conversely, if individuals might select forests according to habitat preference, or if the forests reflect genuine environmental gradients, then treating 'forest' as part of phenotypic variance (an unadjusted or "enhanced" version; R_{v2} or R_{v3}) is likely more biologically meaningful. Ultimately, the data alone cannot dictate which repeatability variant is 'correct'; it depends on the biological context and the research question (see Figure 2).



296

297 **Figure 2. A conceptual diagram of three types of repeatability with an extra random effect**
298 **or a fixed effect**: 1) R_{v1} represents adjusted repeatability controlling for an extra random effect
299 or fixed effect, 2) R_{v2} represents another type of adjusted repeatability where the extra variance
300 components are not part of the focal variance (e.g., individual ID), 3) R_{v3} represents either multi-
301 cluster repeatability (with nested random-effect structure) and enhanced repeatability (with a
302 fixed effect that is biological or a part of focal variance). Note, however, that it is sometimes
303 difficult to know what is biological and what is not (see the main text).

304

2.3 Multiple fixed effects

Real datasets in many areas of biology often include multiple fixed effect predictors. The presence of two or more fixed effects adds both conceptual and technical issues to repeatability estimates. A model fitting two fixed effect predictors can be expressed as (Model 4):

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \alpha_i + \varepsilon_{ij} \quad (2.16)$$

$$\alpha_i \sim N(0, \sigma_\alpha^2) \quad (2.17)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \quad (2.18)$$

where β_2 represents the population-level regression coefficient for the second covariate or predictor, x_{2ij} represents the covariate value for the *i*th objects measured at the *j*th observation of the second predictor, and other terms are the same as in Model 3.

The variance explained by a set of fixed effects, regardless of correlations among the fixed effects, can be lumped into a single variance component by calculating the variance in the linear predictor:

$$\sigma_F^2 = var(\boldsymbol{\beta}\mathbf{X}) \quad (2.19)$$

where $\boldsymbol{\beta}$ represents the vector of regression coefficients, and $\mathbf{X}$ represents the fixed effect design matrix. The variance σ_F^2 can then be treated as in section 2.2.

However, the situation is more complicated when fixed effects covary and the two covarying fixed effects are to be treated differently with respect to repeatability variants R_{v1}, R_{v2} and R_{v3}. If predictors are partially correlated, there is no unequivocal splitting of the outcome variance into a part of the variance explained by x_1 and a part explained by x_2. Parts of the variance will be explained by both x_1 and x_2. Note that while in section 2.2, $\sigma_F^2 = Var(\beta_1 x_1)$ and $\sigma_F^2 = \beta_1^2 Var(x_1)$ were conceptually equivalent (even if numerically often slightly different) $Var(\mathbf{\beta X}) \neq \sum \beta_k^2 \, Var(x_k)$ with correlated predictors unless the variance is explained by both predictors x_1 and x_2 is subtracted on the right-hand side of the equation. We do not have the same issue when there are multiple random effects, which are modelled as independent of each other (cf. [19]).

There are options for partitioning the variance using partial variances explained [20]), but the solutions will be highly case-specific. Some outcome variance is likely explained by both predictors, and the question is where the shared variance components are assigned. In the general case with multiple predictors and interactions, where interaction terms are correlated to their main effects, there are multiple ways to attribute the variance to each predictor (i.e., input variable; *sensu* [1]) and their interaction terms (see [20]).

3 Random-slope models

Many biological subfields are interested in how an outcome variable changes with a predictor and whether that change varies among objects in a grouping factor. For instance, individual or genetic variation in phenotypic plasticity is a major topic in evolutionary biology. Models that capture this variation are often called ‘random-regression’ or ‘random-slope’ models because

342 they fit an interaction between a fixed effect and a random effect, allowing each group (e.g.,
343 individual, genotype) to have its own slope.

344 A simple random-slope model with one covariate and one grouping factor can be expressed as
345 (Model 5):

346 $$y_{hij} = \beta_0 + (\beta_1 + \alpha_{1i})x_{1ij} + \alpha_{0i} + \varepsilon_{ij} \quad (3.1)$$

347 $$[\alpha_{0i}, \alpha_{1i}] \sim N([0,0], \mathbf{\Sigma}) \quad (3.2)$$

348 $$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2_{\alpha 0} & \rho\sigma_{\alpha 0}\sigma_{\alpha 1} \\ \rho\sigma_{\alpha 0}\sigma_{\alpha 1} & \sigma^2_{\alpha 1} \end{bmatrix} \quad (3.3)$$

349 $$\varepsilon_{ij} \sim N(0, \sigma^2_{\varepsilon}) \quad (3.4)$$

350 where β_1 represents the population-level slope for an observation-level predictor x_1 and α_{1i}
351 represents the object-specific deviation from the population slope. The random components α_{0i}
352 and α_{1i} are assumed to be multivariate normally distributed with means of 0 and variance and
353 covariances summarized by the matrix $\mathbf{\Sigma}$, where $\sigma^2_{\alpha 0}$ and $\sigma^2_{\alpha 1}$ are the variances and $\rho\sigma_{\alpha 0}\sigma_{\alpha 1}$ the
354 covariance (ρ represents the correlation between random intercept and slopes). The other terms
355 are as in Model 3.

356 With such variation in the slope depending upon the level of the grouping factor, the
357 repeatability calculation becomes even more involved. There is no universal repeatability
358 anymore, since the variance explained by the grouping factor of interest changes as the covariate
359 changes [21]. In these cases, repeatability can be estimated at particular points along the

360 covariate, termed a ‘conditional repeatability’ (as it is conditioned on the value of the covariate),
361 or an overall repeatability can be estimated across the entire range of the covariate, typically
362 called the ‘marginalized repeatability’. In the following, we first treat this variability in the form
363 of conditional repeatabilities and then suggest the marginalized repeatability as a valuable
364 benchmark for overall repeatability in the outcome with respect to the grouping factor of interest.

365 ## 3.1 Conditional repeatabilities

366 Although we have a random-intercept component and residual variance as in Model 3, a
367 repeatability calculation following Equation 2.13 will result in a repeatability estimate that
368 applies only to the point where the covariate (as fitted in the model) is zero. This point is often
369 arbitrary and does not represent the overall relative magnitude of group-level variances [22],
370 since in a random-slope model, the group-level variance itself changes with the covariate’s
371 value.

372 With random-slope models, any specific repeatability estimate is conditional on the precise value
373 of the covariate, and is therefore referred to as the ‘conditional’ repeatability [22] (see also [2]).
374 Notably, conditional repeatability is not an ‘adjusted’ repeatability (as in Equations 2.5 and
375 2.16), because it is conditioned on a specific point with a specific group-level variance, rather
376 than being adjusted for the entire range (where variance may differ substantially at different x-
377 values). Indeed, conditional repeatability is effectively a function of x (the covariate) rather than
378 a single constant [22].

379 The random-slope variance component $\sigma^2_{\alpha 1}$ determines the magnitude of change across values of
380 the covariate (at the extreme end if $\sigma^2_{\alpha 1} = 0$, the model reduced to a random-intercept model),

while the covariance $\rho\sigma_{\alpha 0}\sigma_{\alpha 1}$ determines where (sign) and how far (magnitude) from $x_1 = 0$ where the minimum value of the conditional repeatability is located (Figure 3a). Negative covariances occur if the minimum of the among-group variance component is located at positive covariate values, while a positive covariance occurs if the minimum is located at negative covariate values [22]). In other words (in strictly linear models), the correlation between random intercepts and slopes can approach +1 or −1 as the absolute vale of *x* grows large in either direction, which may be biologically unrealistic unless the range of *x* is bounded [23].

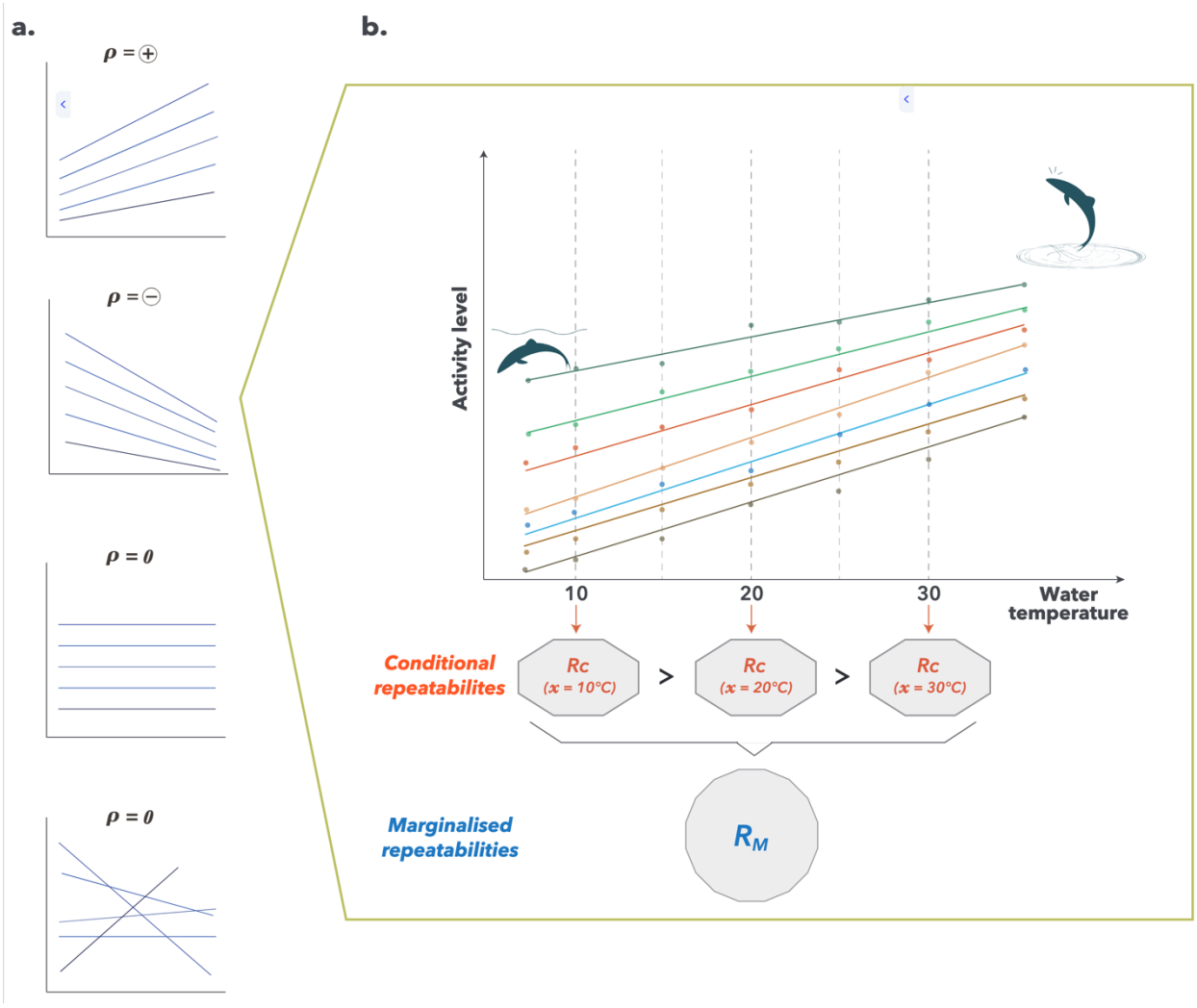
The conditional repeatability (R_C) for any point of the covariate can be calculated as:

$$R_C = \frac{\sigma^2_{\alpha 0} + \sigma^2_{\alpha 1} x_1^* + 2\rho\sigma_{\alpha 0}\sigma_{\alpha 1} x_1^*}{\sigma^2_{\alpha 0} + \sigma^2_{\alpha 1} x_1^* + 2\rho\sigma_{\alpha 0}\sigma_{\alpha 1} x_1^* + \sigma^2_{\varepsilon}}, \qquad (3.5)$$

where x_1^* represents a specific value of x_1 (like 25°C° for temperature, 245 days for age, or 45% for humidity).

Conditional repeatabilities have rarely been explored (but see [24-26]), perhaps because it is not generally understood that random-slope models imply variable group-level variances. Repeatabilities can dramatically change along the range of the covariate (see Figure 3b). We note that x_1^* can be a binary variable (including dummy coded variables). For example, $x_1^* = 0$ it might represent one environment (forest A) and $x_1^* = 1$, another environment (forest B). With Equations 3.5, we can obtain environment-specific repeatability estimates in this case. Note, however, that this assumes that the residual variance σ^2_{ε} is the same in both environments (forests).

We can combine adjusted repeatability and conditional repeatability. If there are more predictors in the model, we can condition on some and adjust (or not) for others (choice of R_{v1}, R_{v2}, and R_{v3}). Indeed, all three types of repeatabilities (R_{v1}, R_{v2}, and R_{v3}) can be used with conditional repeatability and marginalized repeatability (which we will introduce in the next section see Figure 3), given we have several fixed effects or covariates.



405

406 **Figure 3. A conceptual diagram of conditional and marginalized repeatability. using**
407 **'Activity Level' in an aquatic species as the outcome and 'Water Temperature' as the**
408 **covariate.** a) The four panels show different possibilities for the intercept–slope correlation ρ. b)
409 The panel depicts hypothetical individual responses (lines) to changing water temperature. Each
410 individual has its intercept and slope, leading to varying differences in activity at each
411 temperature point (10°C, 20°C, 30°C). The conditional repeatability R_C is estimated at a specific
412 temperature x, while the marginalized repeatability R_M averages across the entire temperature
413 range.

3.2 Marginalized repeatabilities

In random-slope models, no single value of R_C will be representative of the relative magnitude of group-level variances (the larger the random-slope variance the less representative a specific value of R_C will be) across the entire range of the covariate (except maybe for binary covariates when we are interested in environment-specific repeatabilities as described above). However, assuming that the range of the covariate as it occurs in the data is representative of the range of conditions in the population of interest, we can calculate what we call 'marginalized repeatability' that effectively averages group-level variances across the range of the covariate (cf. [17, 27]). The marginalized among-group variance can be calculated as:

$$\tilde{\sigma}_{\alpha}^2 = \sigma_{\alpha 0}^2 + \sigma_{\alpha 1}^2 Var(x_1) + \bar{x}_1^2 \sigma_{\alpha 1}^2 + 2\bar{x}_1 \rho \sigma_{\alpha 0} \sigma_{\alpha 1}, \qquad (3.6)$$

where $\bar{x}_1$ is the population-level mean value of the fixed effect x_1. We can obtain marginalized repeatability as (note that we assume here that there is only one fixed effect in the model):

$$R_M = \frac{\tilde{\sigma}_{\alpha}^2}{\tilde{\sigma}_{\alpha}^2 + \sigma_{\varepsilon}^2}, \qquad (3.7)$$

where σ_F^2 is the outcome variance induced by the population-level slope as in Equations 2.11 and 2.12. Note that calculations are considerably more tedious for multiple correlated predictors, although manageable (see [22]).

4 Post-stratification

For outcome variance explained by fixed effects (Equations 2.11, 2.12, and 2.19) and for marginalized repeatabilities (Equation 3.6), the resulting repeatability values depend on the

433 distribution of the covariate (specifically, the variance in the covariate and, in the case of
434 marginalized repeatabilities, also on the mean). In those calculations, we need not use the exact
435 covariate values observed in our data. Other values may be more representative of the population
436 of interest. Specifically, the variance of the covariate can be replaced by theoretically justified
437 values, a procedure known as 'post-stratification' (*sensu* [28]). This process modifies the scope
438 of inference (see Section 6 for further discussion).

439 An intuitive example of the usefulness of post-stratification comes from an animal study with
440 two sexes. Suppose that, due to a particular experimental design, we sampled disproportionately
441 more females than males, whereas the natural population has an even sex ratio. If we compute
442 the variance explained by sex (Equations 2.11, 2.12, and 2.19) using our sampled data, we
443 misrepresent the actual variance in the population. Likewise, when calculating a marginalized
444 repeatability (Equation 3.6), both the mean and variance of sex (coded as a covariate) would
445 reflect our sample rather than the population. By applying post-stratification, we can weigh our
446 calculations to match the true sex ratio, ensuring the fixed-effect variance for 'sex' aligns with
447 the population distribution.

448 Extrapolation to more extreme ranges of the covariate or to unobserved variances and means can
449 be risky if done without strong justification. However, interpolation may be useful. In many
450 experiments, environmental conditions are deliberately pushed toward extremes (e.g., setting
451 temperature near a species' upper tolerance) to magnify potential effects. If we assume the
452 linearity of the effect, we can interpolate to a narrower or less extreme range of the covariate,
453 thereby avoiding an inflated total phenotypic variance. For instance, [7] illustrates how one
454 might shift from experimental extremes to more moderate conditions by adjusting the fixed-

effect variance. Such interpolation aligns the functional insights from experiments with the natural variation present in wild populations.

5 Non-Gaussian models

All the examples above have used Gaussian data-generating processes with (implicit) identity links. Models 1-5 are thus often called linear mixed-effect models (LMM). However, all the calculations can be generalized to non-Gaussian outcomes and links other than identity links. Generalized linear mixed-effect models (GLMM) can be expressed as (Model 6):

$$\ell_{ij} = \beta_0 + \alpha_i, \quad (5.1)$$

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad (5.2)$$

$$\eta_{ij} = g^{-1}(\ell_{ij}), \quad (5.3)$$

$$y_{ij} = D(\eta_{ij}, \theta), \quad (5.4)$$

where ℓ_{ij} represents the expected value for the *i*th object at the *j*th observation on the link scale, β_0 represents the intercept on the link scale, α_i represents the object-specific random deviations on the link scale, $g^{-1}(\ell_{ij})$ represents the inverse of the link function, $D(\eta_{ij}, \theta)$ represents the process-generating distribution, with a linear predictor η_{ij} and potentially other distribution-specific parameters θ.

The parameterization of the distribution will differ between models. For example, a Poisson model holds only a single rate parameter, such that $D(\eta_{ij}, \theta)$ in the general example above,

reduces to $Pois(\eta_{ij})$ where η_{ij} is the rate of occurrence of an event of interest. In the case of a binomial model, $D(\eta_{ij}, \theta)$ transfers to $B(n, \eta_{ij})$ where n is the number of trials and η_{ij} is the probability of success. As with many other distributions, the Poisson and the Binomial distribution do not include a parameter for the variance, since the variance depends directly on the expectation: η_{ij} for the rate in Poisson and η_{ij} for the probability of success in the Binomial distribution. In such cases, it is usually necessary to include another residual term on the link scale in the form of an observation-level random effect (OLRE) with as many levels as there are observations ([29-31]; alternatively, negative binomial distributions could be used). Such a GLMM with an OLRE can be expressed as (Model 7):

$$\ell_{ij} = \beta_0 + \alpha_i + \omega_{ij}, \quad (5.5)$$

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad (5.6)$$

$$\omega_{ij} \sim N(0, \sigma_\omega^2), \quad (5.7)$$

$$\eta_{ij} = g^{-1}(\ell_{ij}), \quad (5.8)$$

$$y_{ij} = D(\eta_{ij}, \theta), \quad (5.9)$$

where ω_{ij} represents the residual (observation-level) deviation on the link scale, which is assumed to be normally distributed with a mean of 0 and a population-level variance of σ_ω^2 on the link scale (i.e., the variance of the observation-level random effect), and other terms as in Model 5. Also note that ω_{ij} is known as the additive dispersion term, which contrasts with the multiplicative dispersion term, implemented, for example, in generalized linear models fitted by

492 the *glm* function in the software R (for more details on these two types of dispersion terms, see
493 [2]).

494 Non-Gaussian models can be thought of as having three scales ([32]). (1) The link scale of ℓ_{ij} on
495 which terms are linear, and the link-scale model behaves like a Gaussian model. (2) The
496 expected data scale of η_{ij} that represents a non-linear transformation of the link scale by the
497 inverse of the link function. (3) The observed data scale of y_{ij} on which outcomes are observed.
498 The observed data scale is the scale of the outcome, and we usually want the repeatabilities on
499 that scale.

500 Among multiple alternatives (see [2]), the most straightforward way to calculate the repeatability
501 for non-Gaussian models is to infer the distribution-specific variance σ_D^2 that depends on the
502 process-generating distribution and the link function. This distribution-specific variance can then
503 be added to the denominator if appropriate:

504 $$R = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\omega^2+\sigma_D^2}. \qquad (5.10)$$

505 The distribution-specific variance σ_D^2 is known for many popular distributions and link functions
506 [2, 27] and can be inferred for other distributions using the Delta method [27]. Consideration
507 should be given to whether the distribution-specific variance should be added to the denominator
508 [33]. This variance is related to the data generating process, and may represent sampling
509 variance, for example in the case of count data. Often count data represents a sample (e.g.,
510 number of individuals in a sampling area, number of occurrences of a behaviour in a sampling
511 period), and the Poisson variance represents the uncertainty generated by the sampling process.

This sampling variance is a function of the mean and so relies upon the sampling effort. The proportion of the total variance due to the Poisson sampling variance is reduced as the mean increases, and so the maximum attainable repeatability will increase with sampling effort [33]. In the case where the counts are based on a sample, it is more appropriate, therefore, to exclude this distribution-specific variance in the denominator.

In case fixed effects are fitted on the link scale, there will be an additional source of variance; the outcome variance due to fixed effects and considerations as in Sections 2.2, 2.3 and 3.2 do apply. However, in generalized linear mixed-effects models (GLMMs), we have the non-linear link function. A heuristic approach is to calculate the link-scale fixed effect variance following Equation 2.11. However, since the link functions are inherently non-linear, it is better to use integration over the range of the covariates to derive the variance on the expected data and, finally, the observed data scale [32]. Nonetheless, repeatabilities on the link (latent) scale described above is easier to obtain and can be useful [27].

6 A matter of scope

Above, we have reviewed some of the different options for calculating repeatabilities. The choice depends on the research question. In the context of behavioural biology/ecology and the study of individual differences, we are usually interested in having all components that represent or are attributes of individuals in the numerator and all biologically relevant components that contributed to outcome (phenotypic) variance in the denominator.

6.1 Pseudo-repeatabilities

There are some pitfalls, however, and they can give rise to "pseudo-repeatabilities" that do not represent the relative magnitude of individual (group) differences as desired [34, 35]. Often, the issue of pseudo-repeatabilities comes down to a mismatch between the scope of the data (used to derive estimates) and the intended scope of inference (the estimand of interest, i.e. the target parameter). If the scope of inference is larger than the scope of the data, we have to generalize beyond the data, which can be problematic and misleading.

All estimates of variances are also specific to the environment in which data were collected and the sampling regime that was followed. Even the magnitude of among-group variances might differ across environments, or between subpopulations. For example, repeatability of activity level could differ substantially between warm and cold days [36], or between the sexes. Similarly, juvenile repeatabilities might not represent those of adults. In genetics, a parallel issue exists with heritability when data are collected from a restricted environment, but inferences are drawn for a broader context (see e.g., [37]).

6.2 Understanding scope and timing

Similar considerations apply to the time scale of data collection. Repeatabilities over short time scales are often comparatively high because individuals (objects) experience similar external influences (over short frames), and these influences may differ more among individuals than within them. If the scope of inference is about long-term repeatabilities—such as lifetimconsistency in behaviour—short-term estimates are likely misleading, since they tend to be upward biased by semi-stable external influences. That is, the estimate is not aligned with the target of estimation (i.e., the estimand). To estimate long-term repeatabilities, we need data at an

553 appropriate scope—e.g., repeated measurements per individual over their entire lifespan (cf.
554 [38]).

555 The choice between alternative ways of quantifying repeatabilities (described in Sections 2 and
556 3) ultimately depends on the scope of inferences. For instance, deciding whether to adjust for sex
557 (R_{v1}) or not (R_{v2}) depends on whether one aims to compare within-sex differences or across the
558 entire population. Since sex—in most gonochoric animals—is an inherent attribute of
559 individuals, some researchers treat variance explained by sex as part of the among-individual
560 variance, whereas others treat it as an additional explanatory factor that is removed from the
561 denominator. Neither option is inherently wrong, but each leads to a different implication. This
562 point again highlights that which variance components appear in the repeatability calculation is
563 dictated by the research question—albeit with the caveat that any component in the numerator
564 should also appear in the denominator (cf. [37]).

565 Another general timing problem arises if observations occur at a fixed time (e.g., midday), but
566 inference is sought about a broader range of times. Repeatabilities based on a single time point
567 can appear overestimated in that case. Similarly, if objects were measured across different times
568 but the time-level variance is excluded (i.e., using an adjusted repeatability), inferences may
569 again be incorrect. For instance, if the goal is to describe group-level variance over the full day,
570 an adjusted repeatability that removes all 'time' variance would be 'over-adjusted', so we could
571 call it "over-adjusted repeatability". In other words, one might intend to estimate R_{v2} but instead
572 (by ignoring time) end up with R_{v1}.

573 ## 6.3 Missing components and over-adjustment

Sometimes, not all relevant components will be known, or no data have been sampled to model certain fixed and/or random effects. Such sampling regimes can lead to biased repeatability estimates. In particular, omitting a higher-level random effect that is not a legitimate attribute of the focal groups can also yield pseudo-repeatabilities. The lower-level random effects will then absorb variance that belongs at a higher level [19, 39]. That is, although we might be calculating R_{v1} following Equation 2.5, we implicitly calculate a version of R_{v3} when, in fact, we want R_{v2}.

Notably, many published studies over-adjust repeatability. For example, Bell and colleagues [12] found that most estimates were adjusted repeatabilities excluding any fixed-effect variance from the denominator—even when those fixed effects captured biologically meaningful variation. For instance, we might not want to remove age effects on a trait if it is truly relevant (age can contribute both among- and within-individual variance; e.g., [24, 40]). At the same time, there are cases where we do wish to isolate a trait's repeatability at a particular age, so it can be valid to adjust for age—provided the goal is explicitly stated and we also compare it to the unadjusted counterpart.

Ultimately, researchers should be clear whether they are over-adjusting (i.e., omitting meaningful biological variance from the denominator). Ensuring that all relevant higher-level effects (e.g., multi-year or multi-site data) are included can help avoid pseudo-repeatabilities. Conversely, if certain effects are purely methodological (such as instrumentation) and not of biological interest, excluding them may yield a more appropriate estimate. The key is that the chosen scope of inference aligns with how the variance components are assembled into the repeatability measure.

7 Concluding remarks

596 Repeatability, which we reviewed here (Table 1), is an important metric in many areas of
597 biology. It has expanded well beyond its original purpose of assessing the quality of
598 measurement techniques [41]. If used appropriately, repeatability can allow for new insights into
599 sources of phenotypic variance and how such variance evolves. Many of the issues we have
600 presented here also apply to measures of heritability (a standardized measure of additive genetic
601 variance for a referent population). We also note that there are alternatives for standardizing
602 variances, in particular mean standardization, in which variances are divided by the square of the
603 mean outcome value rather than the sum of variance components (the square root of this quantity
604 is known as CV or coefficient of variation; see [42, 43]). Our equations and considerations apply
605 equally when the population-level mean (or expectation) of the outcome average y^2 or better
606 $E(y)^2$ replaces denominators in each equation. Such mean-standardized variance measurements
607 remove the need to consider what is included in the denominator but still require careful
608 consideration of what is included in the numerator. Therefore, what we have described for
609 repeatabilities (ICC) helps obtain a mean-standardized variance value (e.g., CV) of interest.

610 The challenges researchers face is that any real dataset has unique structures that make simple
611 calculations of repeatability (or heritability) problematic. Our survey of these complications and
612 the types of repeatabilities that can be calculated tackles the most common complexities that may
613 exist. Navigating these in specific cases will take careful thought. We emphasize that having a
614 clear goal of presenting repeatability and a thorough understanding of how the data may or may
615 not be suitable for that goal is of primary importance. Usually, the goal of using any standardized
616 variance estimate is to make an inference about a natural population. If so, then the central task is
617 to understand how a dataset is or is not representative of that population. This is not easy, but it
618 is necessary and means that some thought is required and decisions made about what to use in

the calculation. Thus, we advise that any published repeatability estimate comes with explicit descriptions of how repeatability was calculated and why; more importantly, all variance components should be reported regardless. Because we often do not know how well a dataset matches the referent population, we might need several repeatabilities calculated under different hypotheses for the potential mismatches. Such details will allow repeatability to be appropriately used for comparative analyses. The potential for insight into the structure and evolution of phenotypic variance can be achieved only with such details.

Table 1. Types of repeatability, the symbol used for them in the text, the data structure of relevance, and the inference goal for repeatability.

Name	Symbol	Data structure	Inference
Ordinary	R	One grouping variable with sampling within and among	How well the observation is predicted by the group it is in
Adjusted	R_{v1}	As above plus one or more random or fixed effects that explain artefactual variance (such as lab or assay plate)	Adjustment to accurately assess how group identity predicts values when aspects of data collection produce artefacts or biases
Unadjusted	R_{v2}	Grouping variable plus additional variables (either fixed or random) that are not artefacts	Same goal as for ordinary repeatability but with more complex data
Multi-cluster	R_{v3}	Hierarchical nested random effects	To assess how well lower random effect predicts values
Enhanced	R_{v3}	When a fixed effect varies at the grouping level	To include explained variance at the group level, as it would contribute to group identity predicting values
Conditional	R_C	Data having 1+ random effect and 1+ fixed effect with the magnitude of the fixed effect depending on group identity (random slope)	To assess how well group identity predicts values in a specified condition
Marginalized	R_{v3}	Same as for Conditional	Given random slopes, to assess how well group identity predicts values

			across the range of the covariate
Extrapolated	-	Data structure is non-representative of the population at large	Same goal as for Ordinary but using known bias to extrapolate to the referent population
Interpolated	-	Data structure is artificially exaggerated (e.g., experimental conditions exceed natural ones)	Same goal as for Ordinary but using known range of data to adjust for the referent population

628

8 Literature

[1] Gelman, A. & Hill, J. 2007 *Data analysis using regression and multilevel/hierarchical models*. Cambridge ; New York, Cambridge University Press; xxii, 625 p. p.

[2] Nakagawa, S. & Schielzeth, H. 2010 Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol Rev* **85**, 935-956. (doi:10.1111/j.1469-185X.2010.00141.x).

[3] Stoffel, M.A., Nakagawa, S. & Schielzeth, H. 2017 rptR: repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods Ecol Evol* **8**, 1639-1644. (doi:10.1111/2041-210x.12797).

[4] Henderson, C.R. 1950 Estimation of Genetic Parameters. *Ann Math Stat* **21**, 309-310.

[5] Pick, J.L., Kasper, C., Allegue, H., Dingemanse, N.J., Dochtermann, N.A., Laskowski, K.L., Lima, M.R., Schielzeth, H., Westneat, D.F., Wright, J., et al. 2023 Describing posterior distributions of variance components: Problems and the use of null distributions to aid interpretation. *Methods Ecol Evol* **14**, 2557-2574. (doi:10.1111/2041-210x.14200).

[6] Goldstein, H. 2011 *Multilevel statistical models*. 4th ed. Chichester, West Sussex, Wiley; xxi, 358 p. p.

[7] de Villemereuil, P., Morrissey, M.B., Nakagawa, S. & Schielzeth, H. 2018 Fixed-effect variance and the estimation of repeatabilities and heritabilities: issues and solutions. *J Evolution Biol* **31**, 621-632. (doi:10.1111/jeb.13232).

[8] Shewhart, W.A. 1931 *Economic control of quality of manufactured product*. New York,, D. Van Nostrand Company, Inc.; xiv, 501 p. p.

[9] Westneat, D.F., Wright, J. & Dingemanse, N.J. 2015 The biology hidden inside residual within-individual phenotypic variation. *Biol Rev* **90**, 729-743. (doi:10.1111/brv.12131).

[10] Berdal, M.A. & Dochtermann, N.A. 2019 Adaptive Alignment of Plasticity With Genetic Variation and Selection. *J Hered* **110**, 514-521. (doi:10.1093/jhered/esz022).

[11] Dochtermann, N.A. 2023 The role of plasticity, trade-offs, and feedbacks in shaping behavioral correlations. *Behav Ecol* **34**, 913-918. (doi:10.1093/beheco/arad056).

[12] Bell, A.M., Hankison, S.J. & Laskowski, K.L. 2009 The repeatability of behaviour: a meta-analysis. *Anim Behav* **77**, 771-783. (doi:10.1016/j.anbehav.2008.12.022).

[13] Allegue, H., Araya-Ajoy, Y.G., Dingemanse, N.J., Dochtermann, N.A., Garamszegi, L.Z., Nakagawa, S., Réale, D., Schielzeth, H. & Westneat, D.F. 2017 Statistical Quantification of Individual Differences (SQuID): an educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. *Methods Ecol Evol* **8**, 257-267. (doi:10.1111/2041-210x.12659).

[14] Shrout, P.E. & Fleiss, J.L. 1979 Intraclass Correlations - Uses in Assessing Rater Reliability. *Psychol Bull* **86**, 420-428. (doi:Doi 10.1037/0033-2909.86.2.420).

[15] Arct, A., Drobniak, S.M., Mellinger, S., Gustafsson, L. & Cichon, M. 2019 Parental genetic similarity and offspring performance in blue tits in relation to brood size manipulation. *Ecol Evol* **9**, 10085-10091. (doi:10.1002/ece3.5367).

[16] Magierecka, A., Cooper, B., Sloman, K.A. & Metcalfe, N.B. 2023 Unpredictability of maternal environment shapes offspring behaviour without affecting stress-induced cortisol in an annual vertebrate. *Horm Behav* **154**. (doi:ARTN 10539610.1016/j.yhbeh.2023.105396).

[17] Nakagawa, S. & Schielzeth, H. 2013 A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol Evol* **4**, 133-142. (doi:10.1111/j.2041-210x.2012.00261.x).

[18] Araya-Ajoy, Y.G. & Dingemanse, N.J. 2017 Repeatability, heritability, and age-dependence of seasonal plasticity in aggressiveness in a wild passerine bird. *J Anim Ecol* **86**, 227-238. (doi:10.1111/1365-2656.12621).

[19] Schielzeth, H. & Nakagawa, S. 2013 Nested by design: model fitting and interpretation in a mixed model era. *Methods Ecol Evol* **4**, 14-24. (doi:10.1111/j.2041-210x.2012.00251.x).

[20] Stoffel, M.A., Nakagawa, S. & Schielzeth, H. 2021 partR2: partitioning R in generalized linear mixed models. *Peerj* **9**. (doi:ARTN e1141410.7717/peerj.11414).

[21] Martin, J.G.A., Nussey, D.H., Wilson, A.J. & Réale, D. 2011 Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. *Methods Ecol Evol* **2**, 362-374. (doi:10.1111/j.2041-210X.2010.00084.x).

[22] Schielzeth, H. & Nakagawa, S. 2022 Conditional repeatability and the variance explained by reaction norm variation in random slope models. *Methods Ecol Evol* **13**, 1214-1223. (doi:10.1111/2041-210x.13856).

[23] Mitchell, D.J. & Houslay, T.M. 2021 Context-dependent trait covariances: how plasticity shapes behavioral syndromes. *Behav Ecol* **32**, 25-29. (doi:10.1093/beheco/araa115).

[24] Holtmann, B., Santos, E.S.A., Lara, C.E. & Nakagawa, S. 2017 Personality-matching habitat choice, rather than behavioural plasticity, is a likely driver of a phenotype-environment covariance. *P Roy Soc B-Biol Sci* **284**. (doi:ARTN 20170943

10.1098/rspb.2017.0943).

[25] Briffa, M., Sneddon, L.U. & Wilson, A.J. 2015 Animal personality as a cause and consequence of contest behaviour. *Biol Letters* **11**. (doi:ARTN 2014100710.1098/rsbl.2014.1007).

[26] Westneat, D.F. 2024 Biological Links between Personality and Plasticity: Testing Some Alternative Hypotheses. *Am Nat* **203**, 174-188. (doi:10.1086/727700).

[27] Nakagawa, S., Johnson, P.C.D. & Schielzeth, H. 2017 The coefficient of determination and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J R Soc Interface* **14**. (doi:ARTN 20170213

10.1098/rsif.2017.0213).

[28] Gelman, A., Hill, J. & Vehtari, A. 2020 *Regression and other stories*, Cambridge University Press.

[29] Hinde, J. 1982 Compound Poisson regression models. In *Glim 82: Proceedings of the international conference on generalised linear models* (pp. 109-121, Springer.

[30] Harrison, X.A. 2014 Using observation-level random effects to model overdispersion in count data in ecology and evolution. *Peerj* **2**. (doi:ARTN e61610.7717/peerj.616).

[31] Harrison, X.A. 2015 A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution. *Peerj* **3**. (doi:ARTN e1114

10.7717/peerj.1114).

[32] de Villemereuil, P., Schielzeth, H., Nakagawa, S. & Morrissey, M. 2016 General Methods for Evolutionary Quantitative Genetic Inference from Generalized Mixed Models. *Genetics* **204**, 1281-+. (doi:10.1534/genetics.115.186536).

[33] Pick, J.L., Khwaja, N., Spence, M.A., Ihle, M. & Nakagawa, S. 2023 Counter culture: causes, extent and solutions of systematic bias in the analysis of behavioural counts. *Peerj* **11**. (doi:ARTN e1505910.7717/peerj.15059).

[34] Dingemanse, N.J. & Dochtermann, N.A. 2013 Quantifying individual variation in behaviour: mixed-effect modelling approaches. *J Anim Ecol* **82**, 39-54. (doi:10.1111/1365-2656.12013).

[35] Westneat, D.F., Hatch, M.I., Wetzel, D.P. & Ensminger, A.L. 2011 Individual Variation in Parental Care Reaction Norms: Integration of Personality and Plasticity. *Am Nat* **178**, 652-667. (doi:10.1086/662173).

[36] Biro, P.A., Beckmann, C. & Stamps, J.A. 2010 Small within-day increases in temperature affects boldness and alters personality in coral reef fish. *P Roy Soc B-Biol Sci* **277**, 71-77. (doi:10.1098/rspb.2009.1346).

[37] Wilson, A. 2008 Why h 2 does not always equal VA/VP? *J Evolution Biol* **21**, 647-650.

[38] Araya-Ajoy, Y.G., Mathot, K.J. & Dingemanse, N.J. 2015 An approach to estimate short-term, long-term and reaction norm repeatability. *Methods Ecol Evol* **6**, 1462-1473. (doi:10.1111/2041-210x.12430).

[39] Schielzeth, H., Dingemanse, N.J., Nakagawa, S., Westneat, D.F., Allegue, H., Teplitsky, C., Reale, D., Dochtermann, N.A., Garamszegi, L.Z. & Araya-Ajoy, Y.G. 2020 Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol Evol* **11**, 1141-1152. (doi:10.1111/2041-210x.13434).

[40] Bouwhuis, S., Sheldon, B.C., Verhulst, S. & Charmantier, A. 2009 Great tits growing old: selective disappearance and the partitioning of senescence to stages within the breeding cycle. *P Roy Soc B-Biol Sci* **276**, 2769-2777. (doi:10.1098/rspb.2009.0457).

[41] Lessells, C.M. & Boag, P.T. 1987 Unrepeatable Repeatabilities - a Common Mistake. *Auk* **104**, 116-121. (doi:Doi 10.2307/4087240).

[42] Dochtermann, N.A. & Royauté, R. 2019 The mean matters: going beyond repeatability to interpret behavioural variation. *Anim Behav* **153**, 147-150. (doi:10.1016/j.anbehav.2019.05.012).

[43] Stirling, D.G., Réale, D. & Roff, D.A. 2002 Selection, structure and the heritability of behaviour. *J Evolution Biol* **15**, 277-289. (doi:DOI 10.1046/j.1420-9101.2002.00389.x).