A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations

Shinichi Nakagawa*[1], Daniel W. A. Noble[2], Malgorzata Lagisz[1], Rebecca Spake[3], Wolfgang Viechtbauer[4] & Alistair M. Senior*[5]

Address:

[1] Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia

[2] Division of Ecology and Evolution, Research School of Biology, The Australian National University, Canberra, ACT, Australia

[3] School of Biological Sciences, University of Reading, Whiteknights Campus, Reading RG6 6AS, UK

[4] Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Faculty of Health, Medicine, and Life Sciences, Maastricht University, 6200 MD Maastricht, The Netherlands

[5] Charles Perkins Centre and School of Life and Environmental Sciences, University of Sydney, Camperdown, NSW 2006, Australia.

Running head: Improved meta-analyses of the response ratio

* Correspondence: Shinichi Nakagawa (s.nakagawa@unsw.edu.au); Alistair M. Senior (alistair.senior@Sydney.edu.au)

AUTHORSHIP STATEMENT

SN and WV came up with the initial idea and statistical methods, which were discussed and expanded by the other co-authors. AMS led the simulation study, and DWAN put together Supplementary Information with the others' inputs. SN, AMS & DWAN wrote the first draft and all the authors edited and commented on earlier versions of the manuscript.

CONFLICT OF INTEREST

We declare no conflict of interest.

Type of article: Method

Word count: 150 (abstract) and 4996 (main text)

Number of references: 35

Number of figures and tables: 4 figures, 2 tables and 0 text boxes (4 supplementary figures)

DATA AVAILABILITY

All relevant code and data can be found at the GitHub repository (https://github.com/AlistairMcNairSenior/Miss_SD_Sim) and will be archived in Zenodo upon acceptance.

KEYWORDS

Missing data, multiple imputation, meta-regression, robust variance estimation, research synthesis

Abstract

The log response ratio, lnRR, is the most frequently used effect size statistic for meta-analysis in ecology. However, often missing standard deviations (SDs) prevent estimation of the sampling variance of lnRR. We propose new methods to deal with missing SDs *via* a weighted average coefficient of variation (CV) estimated from studies in the dataset that do report SDs. Across a suite of simulated conditions, we find that using the average CV to estimate sampling variances for all observations, regardless of missingness, performs with minimal bias. Surprisingly, even with missing SDs, this simple method outperforms the conventional approach (basing each effect size on its individual study-specific CV) with complete data. This is because the conventional method ultimately yields less precise estimates of the sampling variances than using the pooled CV from multiple studies. Our approach is broadly applicable and can be implemented in all meta-analyses of lnRR, regardless of 'missingness'.

INTRODUCTION

Meta-analyses are frequently used to quantitatively synthesize the outcomes of ecological studies and explain inconsistencies among findings (Gurevitch *et al.* 2018). Meta-analyses often compare the means of two groups, and the most widely used effect sizes for this are the standardized mean difference, SMD (i.e., Cohen's *d* and Hedges' *g*), and the natural logarithm of the response ratio, lnRR (Hedges *et al.* 1999; Nakagawa & Santos 2012; Koricheva & Gurevitch 2014). Both the SMD and lnRR require the standard deviations (SDs) of the two groups to estimate the effect size's precision (i.e., sampling variance). However, many empirical papers do not report SDs or derived statistics from which SDs can be calculated (e.g., standard errors, SEs, and confidence intervals, CIs). A recent review found incomplete reporting of SDs is pervasive and threatens the validity of meta-analytic evidence. Of 505 ecological meta-analytic studies, nearly 70% of the datasets included studies with missing SDs (Kambach *et al.* 2020). The same review also showed that many meta-analysts exclude studies with missing SDs, also known as a 'complete-case' analysis. Unfortunately, simply excluding studies with missing SDs reduces the overall sample size (i.e., number of included studies) can result in biased results (Kambach *et al.* 2020).

An alternative to excluding studies with incomplete data is to impute the missing SDs *via* multiple imputation (MI; Ellington *et al.* 2015; Kambach *et al.* 2020). As a tool to handle missing data, MI was introduced to ecologists more than a decade ago (Nakagawa & Freckleton 2008). However, MI is not widely used in the context of meta-analysis likely for two major reasons. First, for many ecologists the implementation of MI is tedious because it involves three steps: 1) creating *m* (e.g., *m* = 100) replicate versions of the dataset, each containing its own set of imputed values for the missing SDs, 2) analyzing each of these *m* datasets separately, and 3) aggregating the *m* parameter estimates (e.g., regression coefficients) *via* Rubin's rules (Rubin 1987) (for details, see Nakagawa 2015; van Buuren 2018). The second reason MI is not widely used in meta-analysis is uncertainty around its implementation. For example, it is unclear if Rubin's rules are always appropriate for

aggregating estimates of variance/heterogeneity (e.g., τ^2, I^2 and R^2) or information criteria (e.g., AIC, BIC; cf. Nakagawa & Freckleton 2011). Furthermore, MI cannot easily be implemented for multilevel (mixed-effects / hierarchical) meta-analyses, and those implementations that do exist are limited to relatively simple models (van Buuren 2018). For example, as far as we are aware, there is no off-the-shelf implementation of MI for phylogenetic multilevel meta-analytic models despite these models being recommended for meta-analyses that include multiple species – a nearly universal feature of ecological meta-analyses (Cinar *et al.* 2022).

Another common alternative to excluding studies with missing SDs (i.e., complete-case analysis) is to perform an 'unweighted' meta-analysis with lnRR (Koricheva & Gurevitch 2014; O'Dea *et al.* 2021). This approach does not include the sampling variances of effect sizes and thus does not require SDs. However, unweighted analyses are generally inferior to 'formal' meta-analyses for two reasons (cf. Buck *et al.* 2022). First, formal meta-analyses appropriately give more weight to the more precisely estimated effect sizes in the dataset (e.g., those studies with larger sample sizes and hence smaller sampling variances). This weighting improves precision of model parameter estimates, and imparts resilience to publication bias (Hedges & Olkin 1985; Gurevitch *et al.* 2018), because especially smaller studies, which are down-weighted in a weighted analysis, tend to be affected most by this phenomenon. This is an important consideration since publication bias is a common problem in ecology (e.g., Yang *et al.* 2022). Second, a formal meta-analytic model can also quantify heterogeneity (i.e., variation among effect sizes not due to sampling variance) while unweighted models cannot. Quantifying heterogeneity is essential because the overall mean effect size can only be appropriately interpreted in the context of the level of heterogeneity (Hedges & Olkin 1985; Nakagawa *et al.* 2017; Gurevitch *et al.* 2018; Spake *et al.* 2022).

Here, we propose four new methods for handling studies with missing SDs when the lnRR is the effect size of choice (Nakagawa & Santos 2012; Koricheva & Gurevitch 2014; Kambach *et al.*

2020). We note here that our methods do not readily extend to the SMD because the point estimate of SMD is extremely sensitive to the SD, which adds complexity. However, our methods readily integrate with formal meta-analytic models, including traditional random-effects models and more complex multilevel models that are typically more appropriate in ecology (see Fig. 1). We start with an adjusted sampling variance formula for lnRR developed by Doncaster and Spake (2018), which we improve and extend to provide two methods for handling missing SDs: using this adjustment only for effect sizes with missing SDs (the 'missing-cases' method) and using this adjustment for all effect sizes regardless of missingness (the 'all-cases' method). We then describe a third method that extends traditional weighted regression (the 'multiplicative' method). Finally, we combine the missing-cases and multiplicative methods, to give a 'hybrid' method. To compare the performance of these four methods, we have carried out a simulation study including a standard meta-analytic model without missing SDs as a reference. Under a very broad range of simulated conditions, the all-cases method performs best. Surprisingly, even with missing SDs, the all-cases method outperforms the reference method with complete data. Finally, we make recommendations for future meta-analyses. Importantly, we implement and illustrate these new methods *via* the widely used R package, *metafor* (Viechtbauer 2010; all relevant data and code are available at a GitHub repository; see below).

NEW STATISTICAL METHODS

More precise sampling variances: the missing-cases and all-cases methods

The effect size statistic, lnRR, was first proposed by Hedges and colleagues (1999) as follows:

$$\mathrm{lnRR}_1 = \ln\left(\frac{m_1}{m_2}\right), \qquad (1)$$

$$v(\mathrm{lnRR}) = \frac{sd_1^2}{n_1 m_1^2} + \frac{sd_2^2}{n_2 m_2^2} = \frac{\mathrm{CV}_1^2}{n_1} + \frac{\mathrm{CV}_2^2}{n_2}, \qquad (2)$$

where m_1 and m_2 are the means of groups 1 and 2, respectively (e.g., experimental and control groups), v represents the sampling variance, sd and n are the corresponding SDs and sample sizes, respectively, and CV (sd/m) is the coefficient of variation.

However, when the sample size (n; i.e., number of replicates) per effect size is small, the CVs in Equation 2 are often imprecise. This is because the CV is based on sd and m, which are themselves estimates that become less precise with small sample sizes. If we assume the CV values for group 1 and group 2 are reasonably homogeneous across effect sizes (studies), we can obtain a single more precise estimate of CV^2 by averaging across all values in the dataset (Doncaster & Spake 2018; see also Hedges & Olkin 1985; Hunter & Schmidt 1990; Berkey *et al.* 1995):

$$v^*(\text{lnRR}) = \frac{\sum_{i=1}^{K}\left(\text{CV}_{1i}^2\right)/K}{n_1} + \frac{\sum_{i=1}^{K}\left(\text{CV}_{2i}^2\right)/K}{n_2}, \qquad (3)$$

where CV_{1i}^2 and CV_{2i}^2 are the CVs from the ith study (study; i = 1, 2, …, K; we assume the number of effect sizes = the number of studies = K). Indeed, Doncaster and Spake (2018) have demonstrated that the use of Equation 3 over Equation 2 improves the accuracy and precision of the overall (meta-analytic) mean estimate, especially when n is small (e.g., n = 3–10 observations, with $n_1 + n_2$ = 6–20). Notably, they also suggested this formula could be used when SDs are missing from some studies, although this application was not investigated by simulation.

Here we propose two improvements to Equation 3. Using simulations, Lajeunesse (2015) showed that Equations 1 and 2 are biased when sample sizes are small to moderate, and that the following estimators – based on the second-order Taylor expansion – can reduce these biases (see also Senior *et al.* 2020):

$$\text{lnRR}_2 = \ln\left(\frac{m_1}{m_2}\right) + \frac{1}{2}\left(\frac{\text{CV}_1^2}{n_2} - \frac{\text{CV}_2^2}{n_1}\right), \qquad (4)$$

$$v(\mathrm{lnRR}) = \frac{\mathrm{CV}_1^2}{n_1} + \frac{\mathrm{CV}_2^2}{n_2} + \frac{\mathrm{CV}_1^4}{2n_1^2} + \frac{\mathrm{CV}_2^4}{2n_2^2}. \quad (5)$$

Therefore, unifying Equations 3 and 5, and using the square of the weighted average CV (rather than average of CV^2, which is more sensitive to the assumption of normality; see Section "The accuracy and limitation of lnRR") gives the following new estimators for the effect size and sampling variance:

$$\mathrm{lnRR}_3 = \ln\left(\frac{m_1}{m_2}\right) + \frac{1}{2}\left(\frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{1i}) / \sum_{i=1}^{K} n_{1i}]^2}{n_1} - \frac{[\sum_{i=1}^{K}(n_{2i}\mathrm{CV}_{2i}) / \sum_{i=1}^{K} n_{2i}]^2}{n_2}\right), \quad (6)$$

$$\tilde{v}(\mathrm{lnRR}) = \frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{1i}) / \sum_{i=1}^{K} n_{1i}]^2}{n_1} + \frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{2i}) / \sum_{i=1}^{K} n_{2i}]^2}{n_2} + \frac{[\sum_{i=1}^{K}(n_{1i}\mathrm{CV}_{1i}) / \sum_{i=1}^{K} n_{1i}]^4}{2n_1^2} + \frac{[\sum_{i=1}^{K}(n_{2i}\mathrm{CV}_{2i}) / \sum_{i=1}^{K} n_{2i}]^4}{2n_2^2}. \quad (7)$$

We can use Equations 6 and 7 to calculate effect sizes and sampling variances when SDs are missing by simply imputing the pooled CV from the subset of studies that do report SDs. We call this approach as the 'missing-cases' method because we only apply Equations 6 and 7 in studies with missing SDs, while the standard approach of Equations 4 and 5 are applied in studies that report SDs (see Fig. 1 and Table 1 where we consolidated information about the different methods and their assumptions).

Alternatively, one may use Equation 7 for all effect sizes/studies regardless of the missingness of SDs; we call this approach the 'all-cases' method (Table 1). The key difference between the missing- and all-cases methods is that the former assumes that Equation 5 (which bases sampling variances on the study-specific CVs) provides the best estimate of a given effect size's sampling variance, reverting to Equation 7 in cases where SDs are not available. In contrast, the all-cases method assumes that Equation 7 always gives more precise estimates of the sampling variance. Two issues to note are: 1) it is important to use the square of the weighted average CV (Equations 6 & 7) rather than a weighted average of CV^2. CV^2 is very sensitive to non-normally distributed effect sizes

(with large outlying CVs) which might be generated by count data (see Section "The accuracy and limitation of lnRR" below), and 2) when we have multiple effect sizes per study (most meta-analytic datasets in ecology; Nakagawa & Santos 2012), we need to first calculate a weighted average of CVs within studies before taking the weighted average of these cross-study CVs. Alternatively, we could such a weighed average using a multilevel meta-analysis of lnCV (Nakagawa *et al.* 2015; cf. Vachon *et al.* 2019).

A weighted-regression-like approach: the multiplicative method

In the absence of SDs, it has been suggested that information on sample sizes, which are more commonly available, can be used to approximate the sampling variances for lnRR (or SMD), using the inverse of the following (e.g., Lajeunesse 2013; Kambach *et al.* 2020):

$$\tilde{n} = \frac{n_1 n_2}{n_1 + n_2}. \quad (8)$$

However, treating Equation 8 (originally proposed in Hedges & Olkin 1985) as an estimate of the 'exact' sampling variance is erroneous because it ignores the other terms in Equations 2 & 5 (i.e., mean and SD) (see the review by Kambach *et al.* 2020). A more realistic assumption is to treat $1/\tilde{n}$ as proportional to the sampling variance; indeed, Equation 2 reduces to the inverse of Equation 8 (i.e., $1/\tilde{n}$) when we set both CVs to 1. Weighted regression models, commonly used to correct for heteroscedasticity, make this assumption of proportionality. Note that this differs from the classical random-effects meta-analytical model, which assumes that the exact sampling variances are known (and not just up to a proportionality constant). Many ecologists are likely to be familiar with weighted regression models that specify sample sizes as weights (Fletcher & Dixon 2012).

The simplest random-effects meta-analytic model using lnRR can be written as follows:

$$\mathrm{lnRR}_i = \beta_0 + s_i + m_i, \quad (9)$$

$$s_i \sim \mathcal{N}(0, \sigma_s^2), \; m_i \sim \mathcal{N}(0, v_i),$$

where β_0 is the overall/average effect (or meta-analytic mean); s_i is the between-study effect for the ith effect size, sampled from a normal distribution with a mean of zero and variance σ_s^2 (sometimes referred to as τ^2), m_i is the sampling error for the ith effect size, which is also normally distributed with variance equal to the ith sampling variance (note that i = 1, 2, …, K, the number of effect sizes = the number of studies). As mentioned earlier, this model assumes that the sampling variance of lnRR is known (i.e., either Equation 2 or 5 = v_i in Equation 9). The ratio between σ_s^2 and the total variance is often used to quantify heterogeneity (I^2):

$$I^2 = \frac{\sigma_s^2}{\sigma_s^2 + \bar{v}}, \qquad (10)$$

where $\bar{v}$ is known as the 'typical' (or 'average') sampling variance (originally referred to as 'typical within-study variance'; *sensu* Higgins & Thompson 2002), which can be estimated in several ways (Xiong *et al.* 2010).

Unlike the meta-analytic model above, in a weighted regression, the following is assumed:

$$v_i = \phi\left(\frac{1}{\tilde{n}_i}\right), \quad (11)$$

where ϕ, which is estimated by the model, functions as a 'multiplicative' parameter fulfilling the assumption of proportionality (i.e., $1/\tilde{n}_i \propto v_i$). The key point here is that the missing- and all-cases methods both assume that Equations 5 and/or 7 provide an accurate estimate of a study's sampling variance (Table 1). However, Doncaster and Spake's simulation suggests that the sampling variance (using Equation 3) is likely to be imprecise when sample sizes are small (e.g., $n_1 + n_2$ = 6 – 20). Therefore, it may instead be advisable to assume that v_i^* (Equation 3) is proportional to the true sampling variance. In the case that we have missing data, we can extend the assumption of proportionality to Equation 7 to estimate the sampling variance as:

$$v_i = \phi\tilde{v}_i. \quad (12)$$

Practically, this can be implemented as a version of a weighted-regression model that estimates ϕ and assumes proportionality for the sampling variance as in Equation 12 (Fig. 1). We refer to this as the 'multiplicative' method. This method also assumes that Equation 12 provides the best estimate of sampling variance for all studies/effect sizes regardless of SD missingness (Table 1).

Combining missing-cases and the multiplicative method: the hybrid method

In the multiplicative method, Equation 12 is used regardless of whether SDs are missing or not. We can, however, combine the missing-cases and multiplicative methods together into a 'hybrid' method (Fig. 1). In this case, when SDs are available, we can use Equation 5 to obtain the sampling variance of lnRR (along with Equation 4 for the point estimate). When SDs are missing, we can use the multiplicative method (Equation 12, for the sampling variance and Equation 6 for the point estimate). The hybrid method assumes that Equation 5 gives the best estimate of the sampling variances like the missing-case method, but that Equation 12 is an acceptable substitute when SDs are missing. We can write the hybrid method, using a multilevel meta-analysis (including modelling multiple effect sizes per study) as follows:

$$\mathrm{lnRR}_{ij} = \beta_0 + s_i + u_{ij} + m_{ij}, \quad (13)$$

$$s_i \sim \mathcal{N}(0, \sigma_s^2), \; u_{ij} \sim \mathcal{N}(0, \sigma_u^2), \; m_{ij} \sim \mathcal{N}(0, \mathbf{V})$$

where s_i is the between-study effect for the ith study (i = 1, 2, …, K), normally distributed with a mean of 0 and variance σ_s^2 (often referred to as τ^2), u_{ij} is the between-effect-size effect (or within-study effect) for the jth effect size in the ith study, distributed with a mean of zero and variance σ_u^2 (j = 1, 2, …, L_i, where L_i denotes the number of effect sizes within the ith study), $\mathbf{V}$ is a diagonal matrix with v_{ij} (Equation 5) when no SDs are missing and $\phi\tilde{v}_{ij}$ (Equation 12) for cases of missing SD. For example, when we have five effect sizes in three studies, $\mathbf{V}$ would be:

$$\mathbf{V} = \begin{bmatrix} v_{11} & 0 & 0 & 0 & 0 \\ 0 & v_{12} & 0 & 0 & 0 \\ 0 & 0 & \phi\tilde{v}_{21} & 0 & 0 \\ 0 & 0 & 0 & \phi\tilde{v}_{22} & 0 \\ 0 & 0 & 0 & 0 & v_{31} \end{bmatrix},$$

where 1st, 2nd and 5th effect sizes have SDs while the 3rd and 4th are without SDs, and as above, ϕ is estimated in the model. Because this model can account for non-independence, it is appropriate in ecological meta-analyses that include correlations among-effect sizes such as when there is more than one effect size per study or species (Nakagawa & Santos 2012; Noble *et al.* 2017; Nakagawa *et al.* 2022; but for a more complex model with **V** including covariances, or sampling variances with dependencies, see Appendix S1; https://alistairmcnairsenior.github.io/Miss_SD_Sim/). Importantly, all methods described in Table 1 can be used with multilevel meta-analysis making this approach comparable with others.

SIMULATION

Simulation design

We conducted a simulation study to compare the performance of the missing-cases, all-cases, multiplicative and hybrid methods on meta-analytic datasets with varying proportions of missing SDs. We also computed a meta-analytic model with full data, for reference (see Table 1 for the summary of which equations were used for each method; see also Fig. 1). To represent a typical dataset in ecology (and also evolutionary biology), we simulated a hierarchical structure where each study contained ≥1, correlated effect size; i.e., we simulated an intra-class correlation for each study; $ICC_s = \sigma_s^2 / (\sigma_s^2 + \sigma_u^2)$ using the terms in Equation 13. For each simulated dataset we analyzed the full dataset using the conventional approach, before deleting SDs for 5%, 15%, 25% 35%, 45%, or 55% of the studies. We treated 55% as the upper limit of missingness after consulting earlier surveys (e.g., Senior *et al.* 2016; Kambach *et al.* 2020; the latter found ecological meta-analyses had missing SDs for up to 30% of cases). Missingness was imposed at the study-level, rather than the effect size-level. We then analyzed each dataset with the four proposed methods for handling missing SDs.

Datasets were analyzed using models that included a study-level and an effect-size-level random effect, specified using the 'rma.mv' function in *metafor* (Viechtbauer 2010) with the default REML (restricted maximum likelihood) estimator, which has been shown to provide robust estimations of random effects or variance components (e.g., Langan *et al.* 2019). For each model, we calculated: i) bias (as the difference between the estimated and the true, parametrized value) for the meta-estimate of the overall mean effect size, ii) bias for the total amount of heterogeneity ($\tau^2 = \sigma_s^2 + \sigma_u^2$ in Equation 13, and $\tau^2 = \sigma_s^2$ in Equation 9; difference between estimated and parametrized value on the long scale) and the estimated ICC_s (difference between estimated and parametrized value), and iii) coverage of 95% confidence intervals (CIs) for the overall mean. CIs were calculated as the estimated effect ± *t*-value × SE, where for *t*-values the degrees of freedom were the number of effect sizes minus 1, when $ICC_s = 0$, and the number of studies minus 1 when $ICC_s > 0$ (cf. Nakagawa *et al.* 2022).

Each simulated dataset contained *K* studies. Values of $K = 12$, 30, and 100 were tested; these values were taken as representative of small, medium and large meta-analyses based on the survey in Senior *et al.* 2016 (see also Lajeunesse 2015). Because studies often vary in the number of effect sizes they contain, the number of effect sizes per study, *L*, was assigned as a random variable. We simulated *L* using a double Poisson distribution, which is a discrete probability distribution that can be under/over dispersed relative to a Poisson distribution via a multiplicative dispersion parameter. Using the 'rDPO' function in the *gamlss.dist* package, *L* was drawn from a double Poisson distribution with a mean of 2 and a multiplicative dispersion parameter of 2.88, before adding 1 (to prevent 0 values). This resulted in *L* having a minimum of 1, a mean of 3, and SD of 2.4 (i.e., dispersion of 1.92). We termed this set of parameters Set I. We also simulated a second set where *L* is fixed to 1 (i.e., each study had only one effect size; $L = 1$, dispersion = 0), which we called Set II. Set II is equivalent to a meta-analysis with just one effect size per study (i.e., no dependency), and

which we assessed using a standard random-effects meta-analysis (i.e., Equation 9) combined with the missing-cases, all-cases, multiplicative and hybrid methods to handle missing SDs.

To simulate effect sizes that were correlated in a hierarchical manner, we assumed an overall lnRR (θ) of 0.3 ($e^{0.3}$ = 1.35, or a 35% increase in the mean) with either negligible ($\tau^2 = 9\times10^{-6}$ or $\tau / \theta = 0.01$) or high total heterogeneity ($\tau^2 = 0.09$ or $\tau / \theta = 1$), referred to as the low and high heterogeneity settings, respectively. This heterogeneity was partitioned between among- and within-study level effects assuming a given intra-class correlation (ICC_S; values of 0 and 0.5 were tested) such that the jth effect size ($j = 1 \ldots L_i$) in the ith ($I = 1 \ldots K$) study, θ_{ij} (cf. Equation 13) was drawn from a hierarchical pair of random normal distributions ('rnorm' function in *base* R) as:

$$\theta_i \sim N\left(\theta, \sqrt{\tau^2 \times \mathrm{ICC}_s}\right),$$

$$\theta_{ij} \sim N\left(\theta_i, \sqrt{\tau^2 \times (1 - \mathrm{ICC}_s)}\right).$$

To simulate variation in the precision of the studies in the dataset we treated the sample size of the underlying studies as a random variable, N. We assumed N varied at the level of the study such that each group/effect size within the same study had the same sample size. In our experience it is common for experimental designs to vary among, more than within, studies. We drew the simulated sample size for study k by drawing a random value from a double Poisson distribution before adding a value of 3. The double Poisson distribution was parametrized with a mean of either 2 or 27 coupled with dispersion parameters of either 3.65 or 1.66. After adding the constant of 3, this resulted in two different distributions of N both with a minimum of 3, and (over) dispersion of 1.5, but with a mean (μ_N) of either 5 or 30. The smaller mean value of 5 is more typical in terrestrial/ecosystem ecology (or some pre-clinical biomedical studies), while the larger mean value is more like evolutionary/behavioural ecology studies. Note that under the large-mean condition, the sample size of an individual study can be ~250 per group, which matches very large studies in ecology and evolution biology (Senior *et al.* 2016).

The underlying data in control and treatment groups in each effect size were drawn from normal distributions 'shifted' to ensure both groups had a positive mean as is required for analysis using lnRR. From these individual simulated values, we calculated the mean and SD in each group for the calculation of lnRR and meta-analysis. The observations for the control group in effect size j in study i were drawn from the random normal distribution, $N(100, \sigma_i)$, and the paired treatment group from the random normal distribution, $N\left(100 \times e^{\theta_{ij}}, \sigma_i\right)$, where σ_i is the SD in the underlying individual observations in study i.

Because we assessed the performance of methods to deal with missing SD values, we chose to treat the within-group (among-observation) SD as a random variable, S. The SD for study i was drawn from a random Gamma distribution (the 'rgamma' function in *base* R) with shape $\frac{\mu_S^2}{\sigma_S^2}$ and scale $\frac{\sigma_S^2}{\mu_S}$, where μ_S is the mean of S (i.e., mean SD of studies; here 15), and σ_S is the SD in S. This latter parameter thus specifies how heterogeneous the within-study (among-observation) variances are; we tested values of 10^{-10} (~0), 3.75, and 7.5 (i.e., entirely homogeneous variances, or the CV for the SD among studies is 0.25 or 0.5). A summary of the key parameters and their values is given in Table S1. Each combination of parameter values was simulated 10,000 times for both Set I and Set II. For Set I presented in the main text, we used the multilevel meta-analytic model (Equation 13; with the missing and all-cases methods) and its variants (the multiplicative and hybrid methods). For Set II, we used the random-effects meta-analytic model (Equation 9) and its variants. The results from Set II are presented in the supplementary materials, and match those from Set I. For all four methods, we needed to calculate the average CV as in Equations 6 and 7. In Set I, this calculation was done by averaging CV within studies and then taking the weighted-average CV across studies (using mean n per study as the weight), disregarding rows containing missing SDs. For Set II, we calculated the weighted CV among studies (using n per study as the weight) as we only had one CV value per study (see also Fig. S1-S3).

Simulation results

Fig. 2A shows the distribution of the median bias in the estimated overall effect under each simulated condition with complete data and using the four different methods for handling missing SDs. Even with full data, both upward and downward biases were possible for the estimated effect size, and this was also observed in the analyses using the missing-cases and hybrid methods to handle missing SDs. Notably, even at its most extreme, this bias only amounted to a little over 2% of the true effect size and was usually ~0.5%, meaning all the proposed methods performed well (all methods had a median bias across conditions that was < 0.0001). Nonetheless, the all-cases and multiplicative methods, both of which use the weighted average CV to estimate the sampling variance for all effect sizes regardless of missingness, yielded the lowest bias on average and were considerably less variable than other methods (Fig. 2A). The all-cases and multiplicative methods were consistently less biased than the other approaches, regardless of the degree of missingness (Fig. S4A). The degree of bias across conditions in the full data analysis correlated very strongly with that of bias from the missing-cases and hybrid methods, while bias in the missing-cases and hybrid methods correlated strongly with each other (Fig 2B). This observation suggests that the methods fall into two classes that perform similarly across situations: the all-cases and multiplicative methods and the missing-cases and hybrid methods. Contrasting the missing-cases and all-cases methods directly, the absolute level of bias in the missing-cases method was almost always higher than that for the all-cases method (Fig. 2C). Further, where the all-cases method had a higher bias than the missing-cases method, this difference was small (Fig. 2C). Although the all-cases and multiplicative methods outperformed the other approaches on average, they yielded extremely biased estimates on rare occasions; Fig. 2D shows the range in bias among the individual replicates under each simulated condition as a function of the different methods. With the all-cases method, large ranges in bias only occurred when the SDs among different studies were highly heterogeneous, and within-study sample sizes were low (Fig. 2E).

All methods for handling missing data, and the full data analyses, could produce 95% CIs that were too narrow, or too wide under different scenarios (Fig. 3A). The full data, and the missing-cases and hybrid methods tended to typically produce CIs that were slightly too narrow, whereas the all-cases and multiplicative methods were prone to producing wider CIs (Figs 3A and S4B). Again, contrasting the missing-cases and all-cases method, the all-cases method's tendency to produce a CI that is too wide occurs when the total heterogeneity among studies is low (Figs 3B and 3C). However, where total heterogeneity is high, the all-cases method performs as well as the missing-cases method (Fig. 3B and 3C).

Fig. 4A shows the median bias in the estimated heterogeneity under each condition and method. Under most conditions, the missing-cases, all-cases and hybrid methods estimated heterogeneities with little bias, but could also overestimate the total heterogeneity, although to a similar degree to the full data analysis (Fig. 4A). The multiplicative method tended to slightly underestimate heterogeneity (Fig. 4A). Any bias in the estimation of heterogeneity was independent of the actual level of missingness (Fig S4C). Overestimation of heterogeneity occurred where the actual level of heterogeneity was low (Fig. 4B). On average most methods did a good job of partitioning heterogeneity between the within- and among-study levels, although the multiplicative method displayed a slight bias on average (Fig. 4C). Under some circumstances all methods could be biased in partitioning heterogeneity (Fig. 4C). As an example, the missing-cases and all-cases methods were prone to biased partitioning when the total heterogeneity was low; overestimating the ICC when the simulated study effect was absent and underestimating when it was present (Figs 4D and 4E).

In summary, although the all-cases method performed with the least bias under the broad range of simulated conditions tested, all the methods fared surprisingly well, compared with the full data analysis (see Discussion for more). The results presented here pertain to the performance of these

methods in the context of multilevel meta-analytic models (Equation 13, which explicitly model non-independence). However, these conclusions are mirrored for more traditional random-effects models (i.e., analyses without non-independence; contrast Figs 2-4 *vs* Figs S1-S3).

IMPLEMENTATION

The accuracy and limitation of lnRR

The accuracy of the sampling variance for lnRR depends on whether lnRR is normally distributed. Hedges et al. (1999) suggested a simple test to check the assumption of normality based on Geary (1930), who originally advocated screening for effect sizes with $\sqrt{n}/\mathrm{CV} \geq 3$. This test was improved by Lajeunesse (2015) as:

$$\frac{1}{\mathrm{CV}}\left(\frac{4n^{\frac{3}{2}}}{1+4n}\right) \geq 3. \quad (14)$$

If many effect sizes fail to fulfil this relationship, then, meta-analytic results are unlikely to be robust. Lajeunesse (2015) suggests a sensitivity analysis, which excludes effect sizes that fail to fulfill Equation 14. However, such tests are rarely used. Count data and related types (e.g., counts per a given time and space, which are standardized), which are extremely common in ecology (Spake *et al.* 2021), may often fail this test (Equation 14). This is because such data is usually over-dispersed, meaning CV > 1. For example, it is not uncommon for count data to have CV = 5, especially when the mean is close to zero (cf. Lajeunesse 2015). When CV = 5, the sample sizes need to be >226 for each group to pass Equation 14, which would be difficult for most ecological studies to attain.

All meta-analyses of lnRR are sensitive to the assumption of normality to some degree, but our proposed formulations may be more sensitive because the Taylor expansion used in Equations 4-7 also assumes normality. Therefore, it may be advisable to use Equation 1 for the point estimate and

the following estimator of the sampling variance (rather than Equation 7) when many effect sizes fail Geary's test (see also Table S2):

$$\tilde{v}(\text{lnRR}) = \frac{[\sum_{i=1}^{K}(n_{1i}\text{CV}_{1i}) / \sum_{i=1}^{K} n_{1i}]^2}{n_1} + \frac{[\sum_{i=1}^{K}(n_{1i}\text{CV}_{2i}) / \sum_{i=1}^{K} n_{2i}]^2}{n_2}. \quad (15)$$

This formula still relies on the first-order Taylor expansion, but not the second-order, and is therefore less sensitive than Equation 7 to violations of Geary's test. Other limitations (and advantages) of lnRR are discussed elsewhere (e.g., Spake *et al.* 2021; Yang *et al.* 2022).

Worked examples

Bird and colleagues (2019) conducted a meta-analysis exploring the impacts of competition on herbivorous insect fitness when occupying a host plant with another species or in isolation. In brief, they collected data on a series of fitness measurements (e.g., abundance, body size, development time, fecundity; see Table 2 in Bird *et al.* 2019) and quantified the impact of competition on those measures using phylogenetic multilevel meta-analyses (Cinar *et al.* 2022; Appendix S1).

For demonstration purposes, we focused on the largest dataset that used measures of abundance (population size). We restricted our analysis to data on the ratio scale (i.e., having true zero, which is a condition required for lnRR) and those effect sizes that passed the 'improved' Geary's test (Equation 14 above), giving a total of 173 effect sizes from 62 studies. We use a multilevel meta-analytic model (Equation 13) to estimate the overall impact of competition on focal insect fitness (i.e., intercept or overall meta-analytic mean) while controlling for phylogeny, research group, and research year (as per the analysis by Bird *et al.* 2019). We then introduced missing data at the study (article) level, so that a randomly selected ~20% of articles had effect sizes with missing SD in the control and experimental groups; a scenario that is typical of many meta-analyses (cf. Kambach *et al.* 2020).

An analysis of these data applying the different methods compared to the full data is provided in Table 2. We can see that the complete-case analysis (excluding all data with missing SDs) gives slightly larger confidence intervals that cross zero, and a reduction in the meta-analytic mean effect size, relative to most of the other methods. The missing-cases, multiplicative and hybrid methods all suggest the overall meta-analytic is slightly larger and result in greater precision around this estimated effect size than the complete-case analysis. The all-cases method had the smallest overall effect size magnitude, which was not significantly different from zero, while the other three methods yielded mean estimates that were significant (see Discussion). Using this example, we show how each approach is implemented in the supplement (Appendix S2) along with an additional example (McDonald *et al.* 2019; Appendix S3).

DISCUSSION

In this study, we have developed new methodological procedures to handle missing SDs in meta-analyses of lnRR. Our methods will enable ecologists to include studies with missing SDs in their meta-analyses, while also using appropriately weighted formal meta-analyses rather than unweighted counterparts. Our simulation suggested that the least biased estimates were obtained by the 'all-cases' method. This method uses the weighted average CV (estimated from those studies with SDs) to calculate point estimates and sampling variances for all effect sizes, regardless of missingness in SD (Table 1). In terms of implementation, this is also the easiest method of those that we describe (see Supporting Information).

The all-cases method effectively uses 'single imputation' (rather than 'multiple imputation'), and single imputations are generally believed to fare worse than meta-analysis with full data (using Equation 4 & 6, see Table 1; Nakagawa & Freckleton 2008; Nakagawa 2015; van Buuren 2018; Kambach *et al.* 2020; see also Fletcher & Dixon 2012). Yet, this is not what we found. In their previous simulation, Doncaster and Spake (2018) found that Equation 3, which uses the average CV for all effect sizes, performed better than analysis with Equation 2, which uses study-specific CVs.

476 Thus, on reflection, we might have expected the all-cases method to do well (see also Lin & Aloe
477 2021).
478
479 The all-cases method and Doncaster and Spake's procedure (i.e., using Equation 3 rather than
480 Equation 2) perform well because, even where they are reported, the CV values from individual
481 studies are often imprecise due to the small within-study sample size. This, in turn, results in
482 imprecise estimates of the sampling variance. However, using a pooled CV improves estimates of
483 the sampling variance, with benefits to the downstream analyses. Of relevance, another simulation
484 study by Bakbergenuly and colleagues (2020) suggests that sample size (more precisely, $\tilde{n}$ as in
485 Equation 8) is the most important component of weighting in the analysis of lnRR. This insight
486 explains why the all-cases and multiplicative methods do well even in simulations that violate the
487 assumption that CV is homogenous across studies, especially when the number of effect (K) is large
488 (see more for this point below).
489
490 It is important to note that our simulation built on those in Doncaster and Spake (2018) in at least
491 three respects. First, Doncaster and Spake (2018) never tested how their method fared with missing
492 data. Second, our simulation uses multilevel models that are now being applied to many ecological
493 datasets. Third, our simulation has shown that, as well as reducing bias in overall estimates, using a
494 pooled CV does not compromise the accuracy of heterogeneity estimates (i.e., variance
495 components). Between our work and the previous publication by Doncaster and Spake (2018), we
496 have established that using a cross-study averaged CV in the estimation of effect sizes can improve
497 ecological meta-analyses in a range of realistic scenarios.
498
499 Incidentally, Doncaster and Spake (2018) are not the first to use the 'averaging' method. For
500 example, Hedges and Olkin (1985) also proposed to use the average of the observed standardized
501 mean differences in the computation of their sampling variances when meta-analyzing a large

number of small studies. Also, Hunter and Schmidt (Hunter & Schmidt 1990) proposed to use the weighted average of correlations in the sampling variance for the correlation coefficient. Similarly, Berkey et al. (1995) showed that using averages of counts or proportions in the equations for computing the sampling variances of log relative risks and odds ratios led to less biased estimates.

There were two conditions where the all-cases method could result in biased estimates. The first scenario is when CVs are very different between studies, and within-study sample size is relatively small. As discussed below, parallel analysis with the missing-cases method (or alternatively the hybrid method, although the latter is more difficult to implement) could help establish the stability of meta-analytic results. In addition, a meta-analysis of lnCVR (log CV ratio) or lnCV (log CV) could help to evaluate how large the between-study variance in CV is (Nakagawa *et al.* 2015; Senior *et al.* 2020). Large variation in between-study CVs would violate our assumption that the CV is relatively constant (cf. Nakagawa *et al.* 2015). Note, however, that our simulation shows this assumption is less important when studies have larger sample sizes. The second scenario is when there is very low total heterogeneity ($\tau^2 = \sigma_s^2 + \sigma_u^2$, which usually translates to low I^2; see Higgins *et al.* 2003; Nakagawa & Santos 2012; also see Borenstein *et al.* 2017). As mentioned earlier, heterogeneity is typically high in meta-analyses in ecology (and evolutionary biology). Indeed, Senior et al (2016) showed that on average, ecological and evolutionary meta-analyses have high heterogeneity with I^2 of around 90%. Therefore, the second scenario may not be of concern to most ecologists.

Based on the simulation results alone it would be natural to recommend the use of the all-cases method as the default. While we believe the all-cases method is generally the most robust, we advocate that analysts take caution and adopt the following procedure: One should conduct a meta-analysis using both the missing-cases and all-cases methods in tandem, which is very straightforward (see Supplementary Information). If the results of the two methods are qualitatively

528 the same (e.g., both statistically significant, with similar effect size magnitudes), one can present the
529 all-cases method in confidence. If, however, the results are qualitatively different, both results
530 should be presented (e.g., our worked example: see Table 2). In such a case, one should conclude
531 carefully and emphasize uncertainty about their results. An analysis of the heterogeneity among
532 CVs may help guide the user to decide which results to favor; if the CVs are quite different across
533 studies, results from the missing-cases method may be more reliable (see above).
534
535 Notably, our simulation assumes that SDs are missing completely at random. Therefore, when cases
536 with missing SDs are non-random and have consistently higher or lower CVs than cases with SDs,
537 one could use the hybrid method. The hybrid method was shown to work as well as the all-cases
538 method, but this method also can adjust for higher or lower CVs via the multiplicative term ϕ (see
539 Equations 12). Yet, the issue is that we are unlikely to know what CVs missing cases have so that
540 one needs to try the hybrid method to find out (ϕ being more or less than 1). Here, we re-emphasize
541 that all the methods we proposed work well under many conditions (i.e., were not more/less biased
542 than an analysis of the full data). Regardless, it is important to report the % of missing SDs, and
543 which methods have been used to handle missing data, in accordance with the PRISMA-EcoEvo
544 (Preferred Reporting Items for Systematic reviews and Meta-Analyses in Ecology and Evolutionary
545 biology) reporting guidelines (O'Dea *et al.* 2021).
546
547 Finally, our proposed methods are easy to implement and readily extend to a host of complex
548 models. We hope that meta-analysts in ecology and evolution will adopt these two new approaches
549 to improve their meta-analytic estimation, especially the all-cases approach which performs well
550 even in the absence of missing data. Importantly, we should also all be aware of the limitations of
551 the lnRR for meta-analyses, for example, by more routinely evaluating the underlying assumptions
552 using the improved Geary's test.
553

ACKNOWLEDGEMENTS

SN, ML & DAWN were supported by an ARC (Australian Research Council) Discovery grant (DP200100367).

REFERENCES

1. Bakbergenuly, I., Hoaglin, D.C. & Kulinskaya, E. (2020). Estimation in meta-analyses of response ratios. *Bmc Med Res Methodol*, 20.
2. Berkey, C.S., Hoaglin, D.C., Mosteller, F. & Colditz, G.A. (1995). A random-effects regression model for meta-analysis. *Stat Med*, 14, 395-411.
3. Bird, G., Kaczvinsky, C., Wilson, A.E. & Hardy, N.B. (2019). When do herbivorous insects compete? A phylogenetic meta-analysis. *Ecol Lett*, 22, 875-883.
4. Borenstein, M., Higgins, J.P.T., Hedges, L.V. & Rothstein, H.R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Res Synth Methods*, 8, 5-18.
5. Buck, R.J., Fieberg, J. & Larkin, D.J. (2022). The use of weighted averages of Hedges' d in meta-analysis: Is it worth it? *Methods Ecol Evol*, 13, 1093-1105.
6. Cinar, O., Nakagawa, S. & Viechtbauer, W. (2022). Phylogenetic multilevel meta-analysis: A simulation study on the importance of modelling the phylogeny. *Methods Ecol Evol*, 13, 383-395.
7. Doncaster, C.P. & Spake, R. (2018). Correction for bias in meta-analysis of little-replicated studies. *Methods Ecol Evol*, 9, 634-644.
8. Ellington, E.H., Bastille-Rousseau, G., Austin, C., Landolt, K.N., Pond, B.A., Rees, E.E. *et al.* (2015). Using multiple imputation to estimate missing data in meta-regression. *Methods Ecol Evol*, 6, 153-163.
9. Fletcher, D. & Dixon, P.M. (2012). Modelling data from different sites, times or studies: weighted vs. unweighted regression. *Methods Ecol Evol*, 3, 168-176.
10. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, 555, 175-182.
11. Hedges, L.V., Gurevitch, J. & Curtis, P.S. (1999). The Meta-Analysis of Response Ratios in Experimental Ecology. *Ecology*, 80, 1150-1156.
12. Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, Orlando.
13. Higgins, J.P.T. & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat Med*, 21, 1539-1558.
14. Higgins, J.P.T., Thompson, S.G., Deeks, J.J. & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *Brit Med J*, 327, 557-560.

15.
Hunter, J.E. & Schmidt, F.L. (1990). *Methods of meta-analysis : correcting error and bias in research findings*. Sage, Newbury Park ; London.
16.
Kambach, S., Bruelheide, H., Gerstner, K., Gurevitch, J., Beckmann, M. & Seppelt, R. (2020). Consequences of multiple imputation of missing standard deviations and sample sizes in meta-analysis. *Ecol Evol*, 10, 11699-11712.
17.
Koricheva, J. & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *J Ecol*, 102, 828-844.
18.
Lajeunesse, M.J. (2013). Recovering missing or partial data from studies: a survey. In: *The handbook of meta-analysis in ecology and evolution* (eds. Koricheva, J, Gurevitch, J & Mengersen, K). Princeton University Press Princeton, pp. 195-206.
19.
Lajeunesse, M.J. (2015). Bias and correction for the log response ratio in ecological meta-analysis. *Ecology*, 96, 2056-2063.
20.
Langan, D., Higgins, J.P.T., Jackson, D., Bowden, J., Veroniki, A.A., Kontopantelis, E. *et al.* (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods*, 10, 83-98.
21.
Lin, L.F. & Aloe, A.M. (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. *Stat Med*, 40, 403-426.
22.
McDonald, S.E., Lawrence, R., Kendal, L. & Rader, R. (2019). Ecological, biophysical and production effects of incorporating rest into grazing regimes: A global meta-analysis. *J Appl Ecol*, 56, 2723-2731.
23.
Nakagawa, S. (2015). Missing data: mechanisms, methods and messages. In: *Ecological statistics: contemporary theory and application* (eds. Fox, GA, Negrete-Yankelevich, S & Sosa, VJ). Oxford University Press Oxford, pp. 81-105.
24.
Nakagawa, S. & Freckleton, R.P. (2008). Missing inaction: The dangers of ignoring missing data. *Trends in Ecology & Evolution*, 23, 592-596.
25.
Nakagawa, S. & Freckleton, R.P. (2011). Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behav Ecol Sociobiol*, 65, 103-116.
26.
Nakagawa, S., Noble, D.W., Senior, A.M. & Lagisz, M. (2017). Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC Biology*, 15, 18.
27.
Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M. *et al.* (2015). Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods Ecol Evol*, 6, 143-152.
28.
Nakagawa, S. & Santos, E.S.A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, 26, 1253-1274.
29.
Nakagawa, S., Senior, A.M., Viechtbauer, W. & Noble, D.W.A. (2022). An assessment of statistical methods for nonindependent data in ecological meta-analyses: Comment. *Ecology*, 103, e03490.

30.
Noble, D.W.A., Lagisz, M., O'Dea R, E. & Nakagawa, S. (2017). Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. *Molecular Ecology*, 26, 2410-2425.
31.
O'Dea, R.E., Lagisz, M., Jennions, M.D., Koricheva, J., Noble, D.W.A., Parker, T.H. *et al.* (2021). Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: a PRISMA extension. *Biol Rev*, 96, 1695-1722.
32.
Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York ; Chichester.
33.
Senior, A.M., Grueber, C.E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E.S.A. *et al.* (2016). Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. *Ecology*, 97, 3293-3299.
34.
Senior, A.M., Viechtbauer, W. & Nakagawa, S. (2020). Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Res Synth Methods*, 11, 553-567.
35.
Spake, R., Mori, A.S., Beckmann, M., Martin, P.A., Christie, A.P., Duguid, M.C. *et al.* (2021). Implications of scale dependence for cross-study syntheses of biodiversity differences. *Ecol Lett*, 24, 374-390.
36.
Spake, R., O'Dea R, E., Nakagawa, S., Doncaster, C.P., Ryo M, Callaghan, C.T. *et al.* (2022). Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology and Evolution*, Accepted.
37.
Vachon, H., Viechtbauer, W., Rintala, A. & Myin-Germeys, I. (2019). Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental Disorders: Meta-Analysis and Recommendations. *J Med Internet Res*, 21.
38.
van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition*. 2nd edn. CRC Press, Boca Raton.
39.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1-48.
40.
Wickham, H. (2009). *ggplot2 : elegant graphics for data analysis*. Springer, New York ; London.
41.
Xiong, C.J., Miller, J.P. & Morris, J.C. (2010). Measuring Study-Specific Heterogeneity in Meta-Analysis: Application to an Antecedent Biomarker Study of Alzheimer's Disease. *Stat Biopharm Res*, 2, 300-309.
42.
Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I. & Nakagawa, S. (2022). Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Glob Chang Biol*, 28, 969-989.

699 **Figure Legends**

700

701 **Figure 1**

702 Visual schematics of a hypothetical dataset with missing standard deviations (SDs) and five
703 different approaches used in this study, including 3 new methods. The symbols: $\ln RR_2$ (Equation 4),
704 $\ln RR_3$ (Equation 6), v (Equation 5), $\tilde{v}$ (Equation 7), and $\phi\tilde{v}$ (Equation 12). Note that, under some
705 circumstances, we could replace Equations 4 & 6 with Equation 1 while Equation 7 can be replaced
706 by Equation 15 (see the text for more details).

707

708 **Figure 2**

709 Results on overall meta-analytic mean from multi-level meta-analytic models: A) Violin plot
710 showing the distribution of median bias in the estimated effect under each simulated condition as a
711 function of the method used to handle missing data (distribution assuming full data shown for
712 reference). B) Pairwise correlations between the degree of bias under each simulated condition for
713 each method. C) Distribution of the difference between the missing-cases and all-cases methods in
714 the absolute degree of bias under each condition (positive values indicate greater median bias under
715 the missing-cases method). D) Violin plot showing the distribution of range bias ($\log_{10}$ transformed)
716 in the estimated effect under each simulated condition as a function of the method used to handle
717 missing data. E). Violin plot showing the distribution of range bias ($\log_{10}$ transformed) in the
718 estimated effect using the all-cases method under each simulated condition as a function of the
719 degree of heterogeneity in SDs among studies under two different (within-)study sample size
720 conditions. Our plots were drawn using the R package ggplot2 (Wickham 2009).

721

722 **Figure 3**

723 Results on coverage from multi-level meta-analytic models: A) Violin plot showing the distribution
724 of coverage of 95% CIs under each simulated condition as a function of the method used to handle

missing data (distribution assuming full data shown for reference). B) Violin plot showing the distribution of coverage under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using the missing-cases method to handle missing SDs. C) Violin plot showing the distribution of coverage under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using the all-cases method to handle missing SDs. In B and C, low heterogeneity is $\tau^2 = 9\times10^{-6}$ (or $\tau / \theta = 0.01$), and high heterogeneity is $\tau^2 = 0.09$ (or $\tau / \theta = 1$).

Figure 4

Results on heterogeneity from multi-level meta-analytic models: A) Violin plot showing the distribution of median bias in the estimated heterogeneity under each simulated condition as a function of the method used to handle missing data (distribution assuming full data shown for reference). Bias in heterogeneity is calculated as the log ratio of the estimated and parametrized value. B. Box plot showing the median bias in estimated heterogeneity under each simulated condition as a function of the method used to handle missing data (colours as in panel A), and the simulated level of heterogeneity. C) Violin plot showing the distribution of the median bias in the estimated ICC for study under each simulated condition as a function of the method used to handle missing data. Bias in the ICC was calculated as the difference between the estimated and parameterized value. D) Violin plot showing the distribution of the median bias in the estimated ICC for study under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using the missing-cases method to handle missing SDs. E) Violin plot showing the distribution of the median bias in the estimated ICC for study under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using the all-cases method to handle missing SDs. In D and E, low heterogeneity is $\tau^2 = 9\times10^{-6}$ (or $\tau / \theta = 0.01$), and high heterogeneity is $\tau^2 = 0.09$ (or $\tau / \theta = 1$).

Figure S1

Results on overall meta-analytic mean from random-effects meta-analytic models: A) Violin plot showing the distribution of median bias in the estimated effect under each simulated condition as a function of the method used to handle missing data (distribution assuming full data shown for reference). B) Pairwise correlations between the degree of bias under each simulated condition for each method. C) Distribution of the difference between the missing-cases and all-cases methods in the absolute degree of bias under each condition (positive values indicate greater median bias under the missing-case methods). D) Violin plot showing the distribution of range bias ($\log_{10}$ transformed) in the estimated effect under each simulated condition as a function of the method used to handle missing data. E. Violin plot showing the distribution of range bias ($\log_{10}$ transformed) in the estimated effect under each simulated condition using the all-cases method to handles missing SDs as a function of the degree of heterogeneity in SDs among studies under two different (within-)study sample size conditions.

Figure S2

Results on coverage from random-effects meta-analytic models: A) Violin plot showing the distribution of coverage of 95% CIs under each simulated condition as a function of the method used to handle missing data (distribution assuming full data shown for reference). B) Violin plot showing the distribution of coverage under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using the missing-case method to handle missing SDs. C) Violin plot showing the distribution of coverage under each simulated condition as a function of the simulated level of total heterogeneity and the ICC for study using the all-cases method to handle missing SDs. In B and C, low heterogeneity is $\tau^2 = 9\times10^{-6}$ (or $\tau / \theta = 0.01$), and high heterogeneity is $\tau^2 = 0.09$ (or $\tau / \theta = 1$).

Figure S3

Results on coverage from random-effects meta-analytic models: A) Violin plot showing the distribution of median bias in the estimated heterogeneity under each simulated condition as a function of the method used to handle missing data (distribution assuming full data shown for reference). Bias in heterogeneity is calculated as the log ratio of the estimated and parametrized value. B. Box plot showing the median bias in estimated heterogeneity under each simulated condition as a function of the method used to handle missing data (colours as in panel A), and the simulated level of heterogeneity.

Figure S4

Bias in A) overall meta-analytic mean estimation, B) coverage and C) heterogeneity, as function of the method used to handle missing SDs and the percentage of studies with missing SDs in the simulated dataset. Note that for the full data analysis no studies have missing SDs and thus no trend is expected. Random 'jitter' has been added to the *x*-axis to make overlaying points visible. Fitted lines are based on a generalised additive model (GAM) implemented using the 'geom_smooth' function in ggplot2.

794 **Table 1** Equations and assumptions for different methods, including the case with no missing data (see also Figure 1).

Method	**Point estimate**[1]	**Sampling variance (SD not missing)**	**Sampling variance (SD missing)**	**Assumptions in relation to sampling variance**
Reference (No missing data)	Equation 4	Equation 5	Not applicable	Equation 5 estimates sampling variance well (observed mean and SD values are reasonable estimates of true values)
Missing cases	Equations 4 & 6	Equation 5	Equation 7	When SD values are missing, Equation 7 can estimate sampling variance for these missing cases well
All cases	Equations 4 & 6	Equation 7	Equation 7	Equation 7 estimates sampling variance better than Equation 5 regardless of missing SD
Multiplicative	Equations 4 & 6	Equation 12	Equation 12	Equation 12 estimates sampling variance better than Equation 5 or 7 regardless of missing SD
Hybrid	Equations 4 & 6	Equation 5	Equation 12	When SD is missing, Equation 12 can estimate sampling variance for these missing cases well (better than Equation 7)

795 [1] Applying both Equations 4 & 6 (the latter for observations/rows with missing SD) or applying only Equation 6 (even for all studies where SDs are not
796 missing) would make little difference for (effect size) point estimates, unless effect sizes fulfill Equation 14.

797 **Table 2** Results from the re-analyses of a subset of data from Bird et al. (2019) using the methods
798 we propose to deal with missing SD data estimating the overall effects of competition on focal
799 insect abundance (LCI = lower, or 2.5%, confidence limit; UCI = upper, or 97.5%, confidence
800 limit).

Method	Est.	SE	95% LCI	95% UCI
Full data	0.202	0.085	0.036	0.369
Complete case	0.176	0.102	-0.024	0.377
Missing cases	0.186	0.091	0.008	0.364
All-cases	0.146	0.096	-0.043	0.334
Multiplicative	0.192	0.083	0.03	0.354
Hybrid	0.185	0.086	0.017	0.353

801
802

803

804 **Table S1.** Variables/parameters in simulations.

Variable (Notation)	Description and details	Value(s)
% Studies Missing SD	Percentage of studies that have missing SDs	5, 15, 25, 35, 45 or 55
Overall Effect Size (θ)	The overall mean lnRR effect size	0.3
Number of Studies (K)	Total number of studies within the meta-analytic dataset	12, 30, 100
Standard Deviation in Study (S)	The within-study SDs. Individual within-study SDs were randomly distributed following a Gamma distribution	Random with a mean (μ_S) of 15 and a SD of either 10^{-10}, 3.75 or 7.5

805

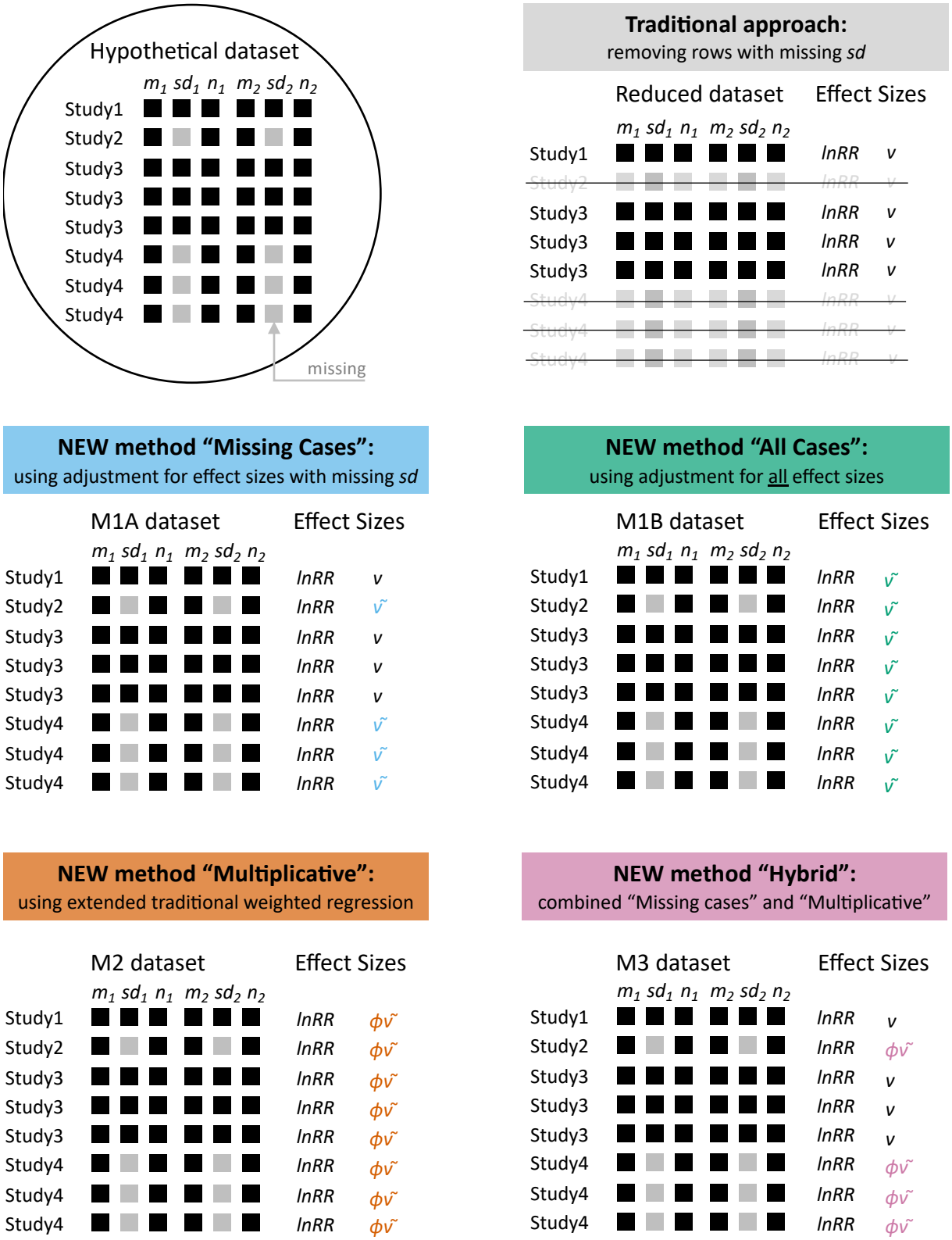
806

807 **Table S2** Recommendations for the use of equations when observations (effect sizes) fail Geary's
808 test; note ϕ is a multiplicative factor as in Equation 12 (cf. Table 1).

Method	**Point estimate**	**Sampling variance (SD not missing)**	**Sampling variance (SD missing)**
Missing cases	Equation 1	Equation 2	Equation 15
All cases	Equation 1	Equation 15	Equation 15
Multiplicative	Equation 1	Equation 15 x ϕ	Equation 15 x ϕ
Hybrid	Equation 1	Equation 2	Equation 15 x ϕ

809
810
811

812 **Figure 1**



813

814

815

Figure 2

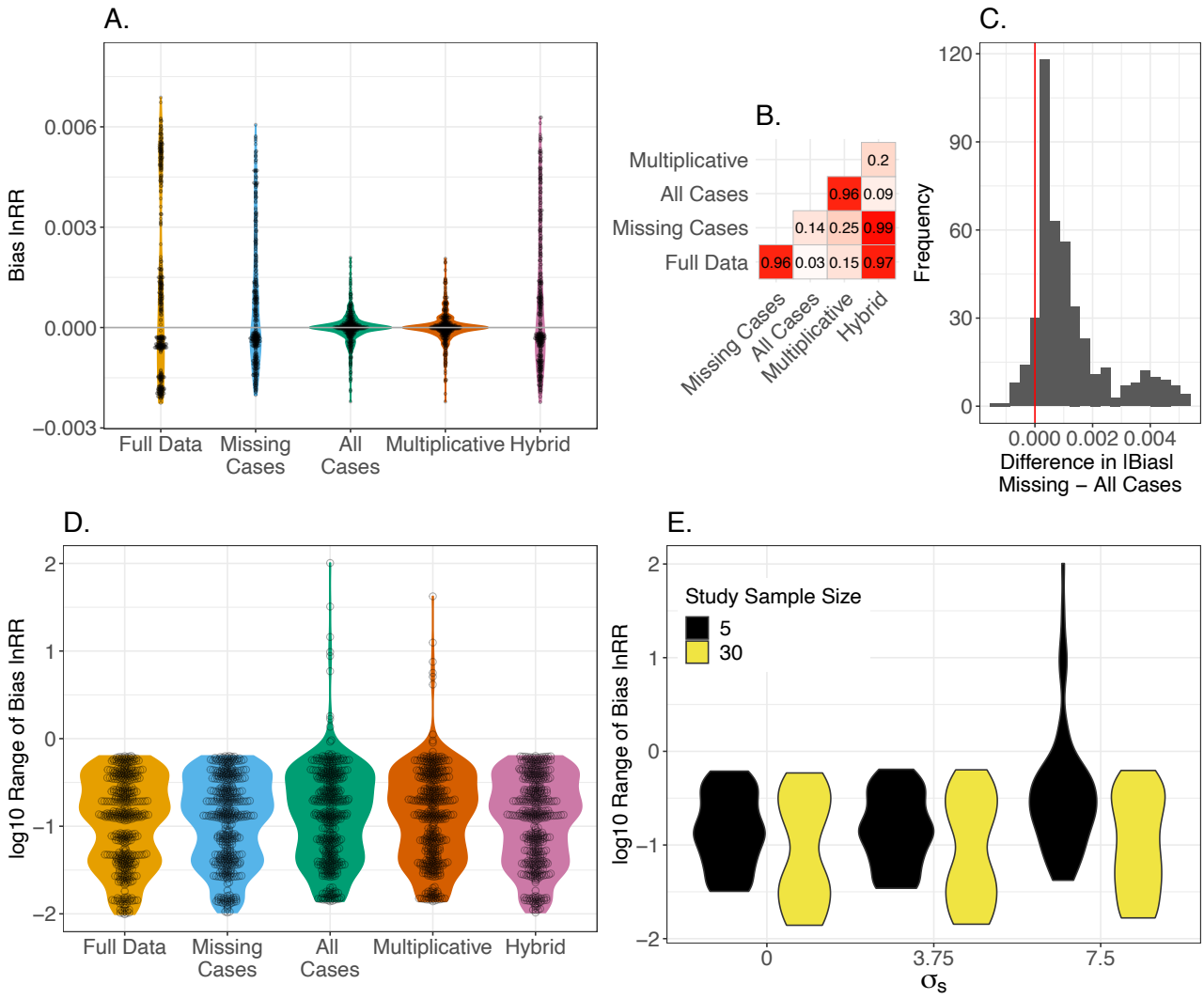

A.
Bias lnRR
0.006
0.003
0.000
−0.003
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
B.
Multiplicative
All Cases
Missing Cases
Full Data
0.2
0.96 0.09
0.14 0.25 0.99
0.96 0.03 0.15 0.97
Missing Cases
All Cases
Multiplicative
Hybrid
C.
Frequency
120
90
60
30
0
0.000 0.002 0.004
Difference in |Bias|
Missing – All Cases
D.
log10 Range of Bias lnRR
2
1
0
−1
−2
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
E.
log10 Range of Bias lnRR
Study Sample Size
5
30
0
3.75
7.5
σs

Figure 3

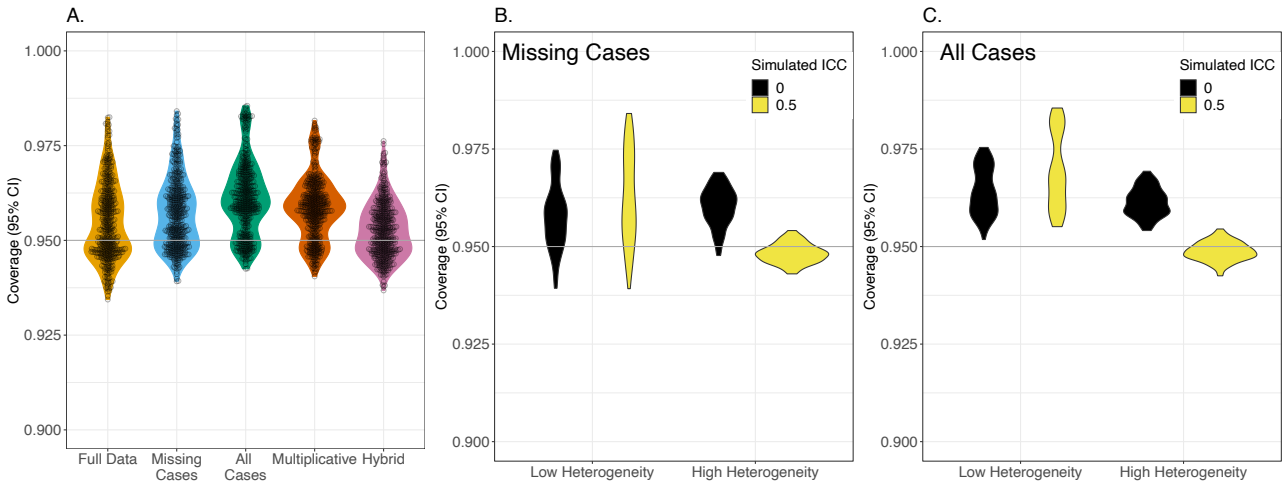

A.
B.
C.
Missing Cases
All Cases
Simulated ICC
0
0.5
Coverage (95% CI)
1.000
0.975
0.950
0.925
0.900
Full Data
Missing Cases
All Cases
Multiplicative Hybrid
Low Heterogeneity
High Heterogeneity

Figure 4

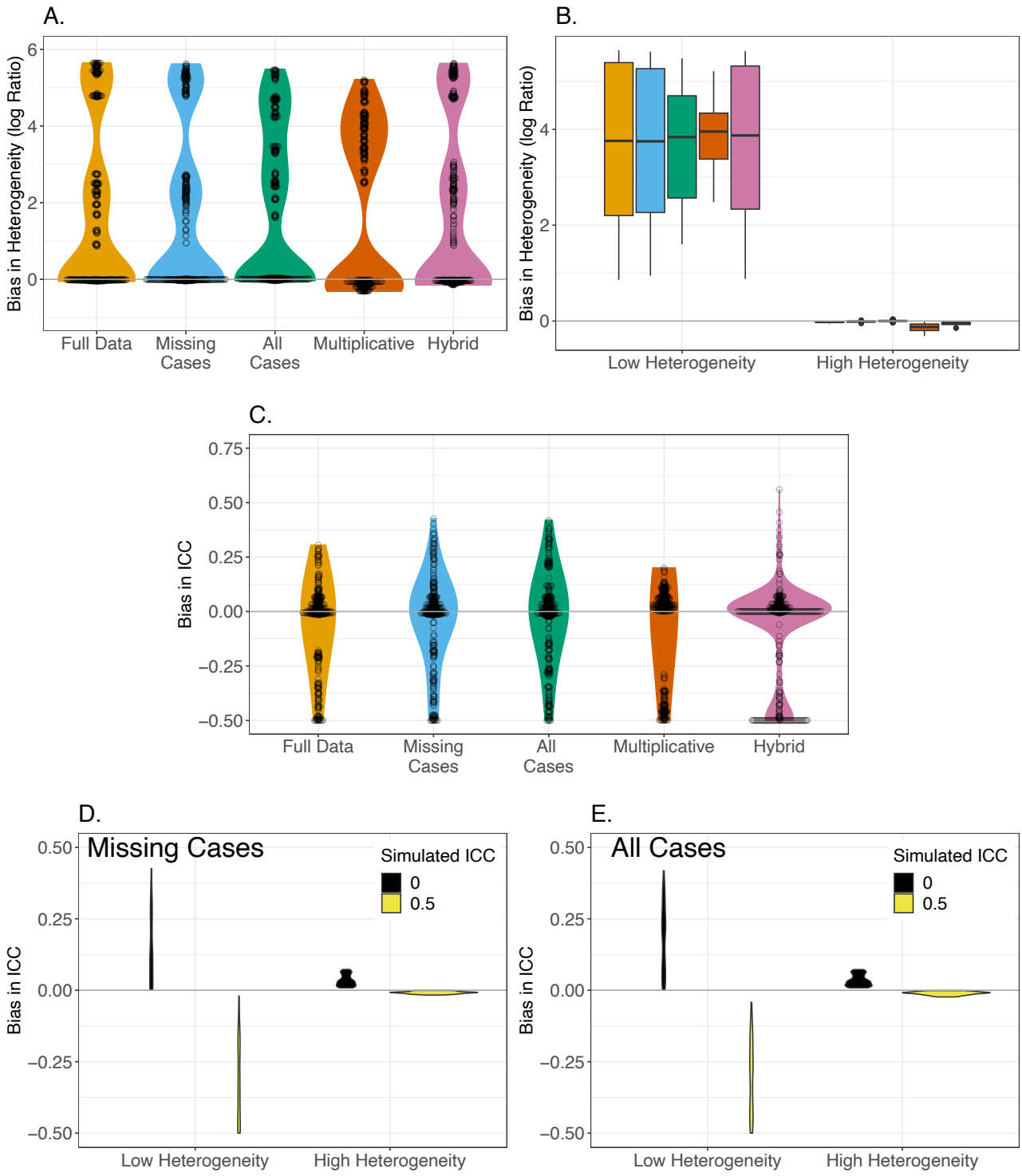

A.
Bias in Heterogeneity (log Ratio)
6
4
2
0
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
B.
Bias in Heterogeneity (log Ratio)
4
2
0
Low Heterogeneity
High Heterogeneity
C.
Bias in ICC
0.75
0.50
0.25
0.00
−0.25
−0.50
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
D.
Missing Cases
Simulated ICC
0
0.5
Bias in ICC
0.50
0.25
0.00
−0.25
−0.50
Low Heterogeneity
High Heterogeneity
E.
All Cases
Simulated ICC
0
0.5
Bias in ICC
0.50
0.25
0.00
−0.25
−0.50
Low Heterogeneity
High Heterogeneity

Figure S1

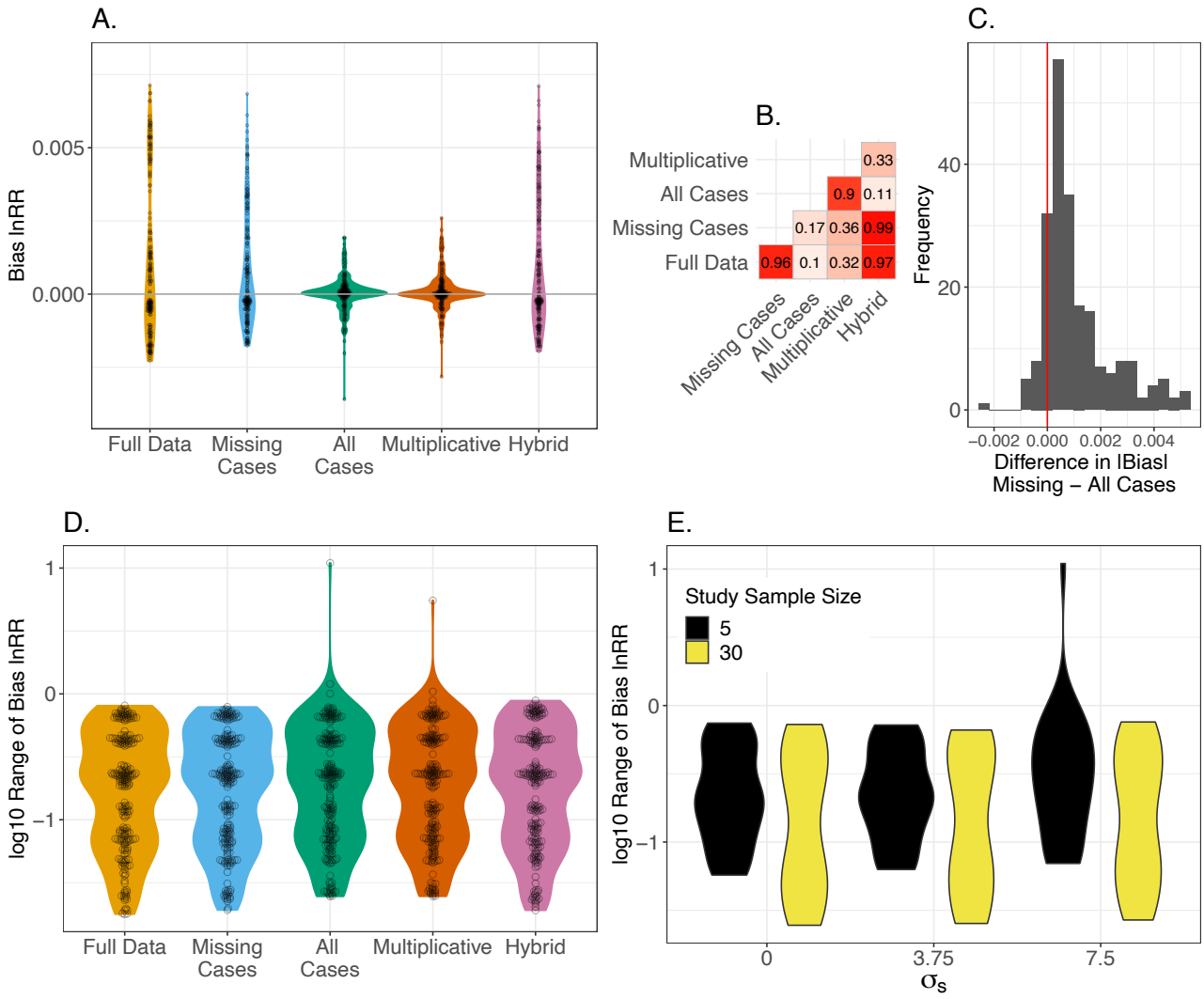

A.
Bias lnRR
0.005
0.000
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
B.
Multiplicative
All Cases
Missing Cases
Full Data
0.33
0.9
0.11
0.17
0.36
0.99
0.96
0.1
0.32
0.97
Missing Cases
All Cases
Multiplicative
Hybrid
C.
Frequency
40
20
0
−0.002
0.000
0.002
0.004
Difference in |Bias|
Missing – All Cases
D.
log10 Range of Bias lnRR
1
0
−1
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
E.
log10 Range of Bias lnRR
Study Sample Size
5
30
0
3.75
7.5
σs

Figure S2

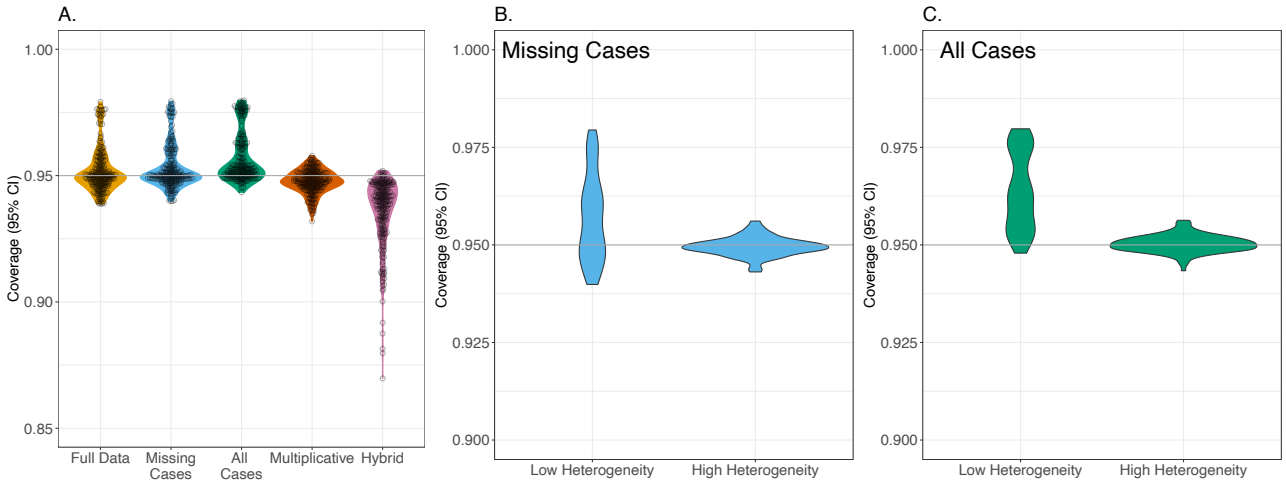

A.
B.
C.
Missing Cases
All Cases
Coverage (95% CI)
1.00
0.95
0.90
0.85
1.000
0.975
0.950
0.925
0.900
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
Low Heterogeneity
High Heterogeneity

Figure S3

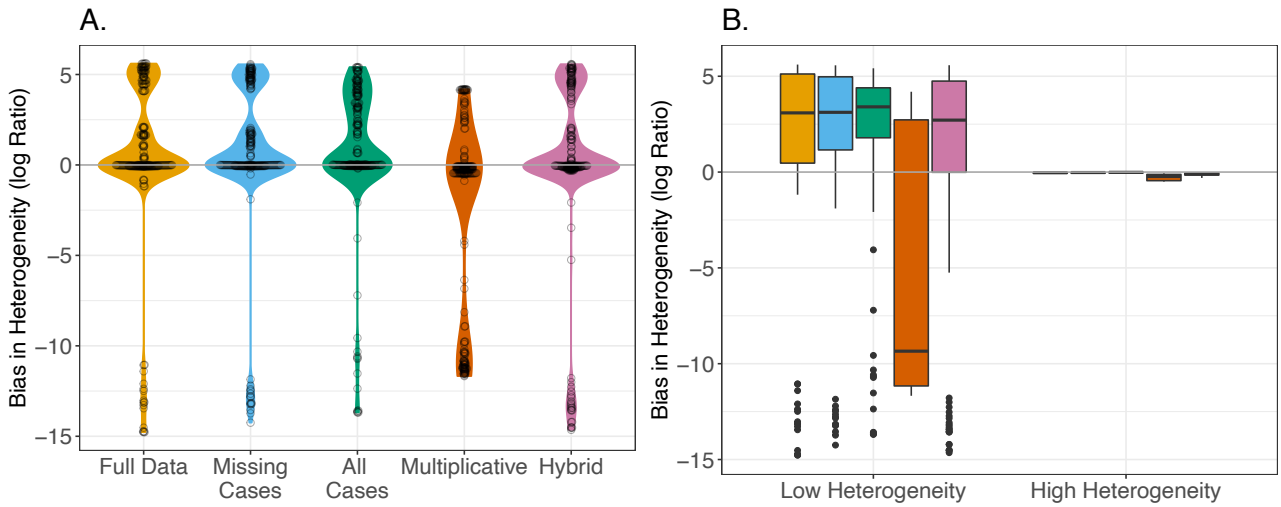
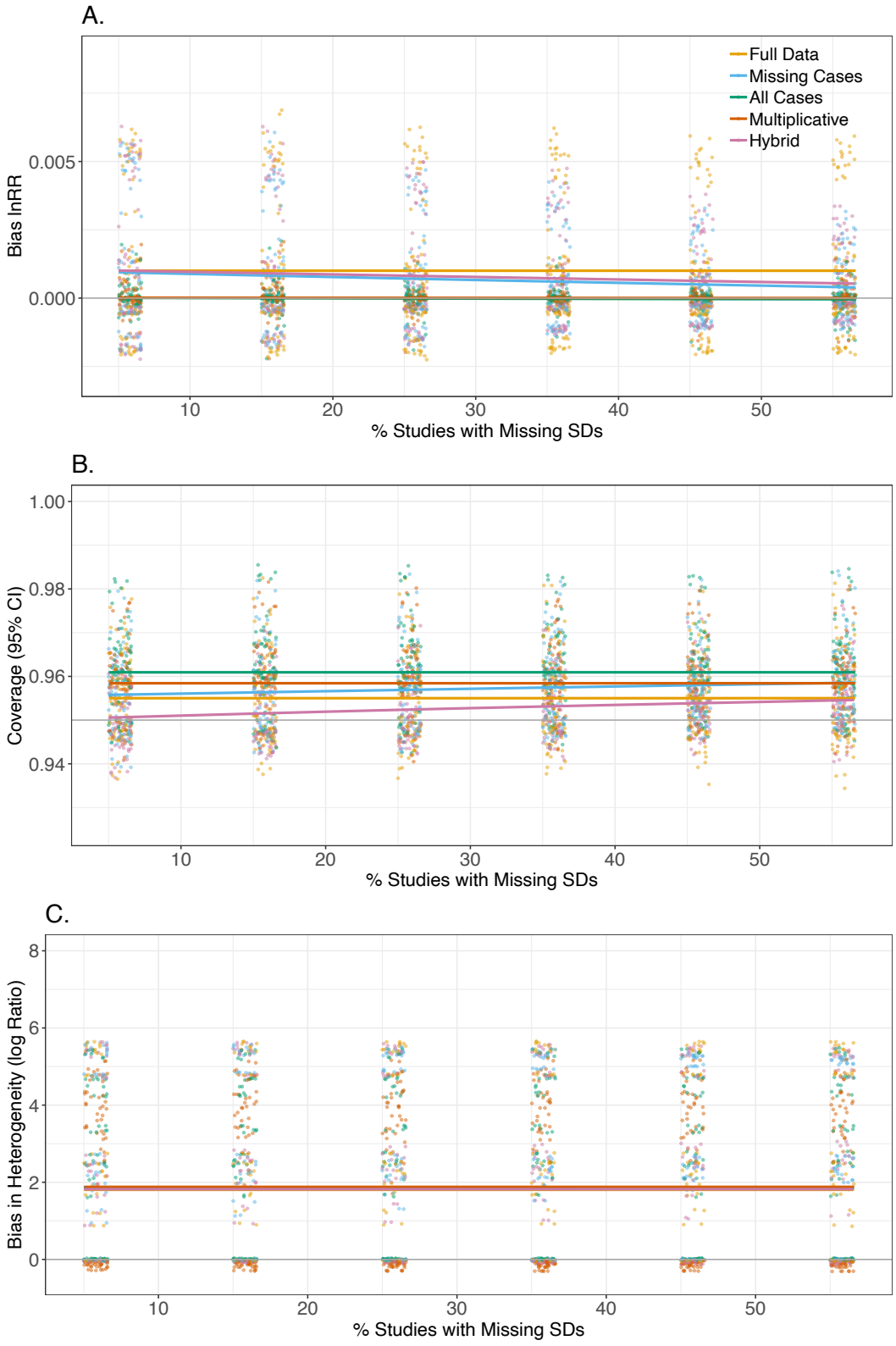

A.
B.
Bias in Heterogeneity (log Ratio)
5
0
−5
−10
−15
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
Low Heterogeneity
High Heterogeneity

Figure S4


A.
Full Data
Missing Cases
All Cases
Multiplicative
Hybrid
Bias lnRR
0.005
0.000
10
20
30
40
50
% Studies with Missing SDs
B.
Coverage (95% CI)
1.00
0.98
0.96
0.94
10
20
30
40
50
% Studies with Missing SDs
C.
Bias in Heterogeneity (log Ratio)
8
6
4
2
0
10
20
30
40
50
% Studies with Missing SDs