An assessment of statistical methods for non-independent data in ecological meta-analyses: Comment

Shinichi Nakagawa[1*], Alistair M. Senior[2], Wolfgang Viechtbauer[3], and Daniel W. A. Noble[4]

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

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

3. 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

4. Division of Ecology and Evolution, Research School of Biology, The Australian National University, Canberra, ACT, Australia

* correspondence: s.nakagawa@unsw.edu.au

Running title: Avoid averaging effect sizes per paper

Shinichi Nakagawa: 0000-0002-7765-5182

Alistair M. Senior: 0000-0002-7765-5182

Wolfgang Viechtbauer: 0000-0003-3463-4063

Daniel W. A. Noble: 0000-0001-9460-8743

Key words: meta-regression, Bayesian statistics, Satterthwaite approximation, multilevel modeling, hierarchical models, degrees of freedom

(Introduction)

Recently, Song et al. (2020) conducted a simulation study using different methods to deal with non-independence resulting from effect sizes originating from the same paper – a common occurrence in ecological meta-analyses. The main methods that were of interest in their simulations were: 1) a standard random-effects model used in combination with a weighted average effect size for each paper (i.e., a two-step method), 2) a standard random-effects model after randomly choosing one effect size per paper, 3) a multilevel (hierarchical) meta-analysis model, modelling paper identity as a random factor, and 4) a meta-analysis making use of a robust variance estimation method. Based on their simulation results, they recommend that meta-analysts should either use the two-step method, which involves taking a weighted paper mean followed by analysis with a random-effects model, or the robust variance estimation method.

Song et al.'s simulation results are an important and valuable contribution to the ecological community. However, we disagree with their primary recommendation of calculating a weighted average effect size for each study within a paper for two reasons. First, as we have stated elsewhere (Nakagawa & Santos 2012, Noble et al. 2017), we recommend the use of multilevel meta-analytic models because of improved power and the ability to answer richer biological questions about the drivers underlying variation in published effects. Second, we do not recommend the use of the two-step method with a weighted paper mean because other types of within-study non-independence often co-occur that need to be considered but that are not completely dealt with by Song et al. (2020)'s simulation. We fully agree that a robust variance estimation method is useful, but from Song et al. (2020) paper it would appear to be limited in applicability. However, we show that this method can easily be extended to multilevel meta-analysis, making the best of both worlds.

In this Comment, we overview a previous simulation study with different conclusions to that of Song et al. (2020) and put forward a strong case for why we need to make use of multilevel meta-analysis in the field of ecology. We discuss how the results of this previous simulation, along with our updated simulation results from Song et al. (2020), make different conclusions that show multilevel meta-analysis can perform well when non-independence exists. In our simulations, we demonstrate how a number of additional methods can provide solutions for any increase in Type I error when fitting multilevel meta-analysis models (an issue noted by Song et al., 2020).

Similar Simulations, Different Conclusions

Moeyaert et al. (2017) conducted a similar simulation study to Song et al. (2020) with some minor differences. First, Moeyaert et al. (2017) did not include a condition involving randomly choosing one effect per paper and used the standardized mean difference (aka Cohen's *d* or Hedges' *g*) as their effect size, instead of the log response ratio used by Song et al. (2020) (Hedges et al. 1999). Second, Moeyaert et al. (2017) did not model different correlations within papers (they referred to papers as studies) and heteroscedasticity among papers (different between-paper variances). Finally, Moeyaert et al. (2017) used PROC GLM in SAS 9.3 (SAS Institute Inc, 2011-2014) while Song et al. used R's *metafor* (Viechtbauer 2010) and *robumeta* (Fisher et al. 2017) for multilevel meta-analysis and robust variance estimation, respectively.

Like Song et al. (2020), Moeyaert et al. (2017) found that all of the three methods examined produced unbiased estimates of the overall (meta-analytic) mean. A striking difference was that in Moeyaert et al. (2017), multilevel meta-analyses performed as well as a robust variance estimation method in terms of 95% confidence interval (CI) coverage. In contrast, Song et al. (2020) reported consistently higher Type I error rates (i.e., greater than 5%) for multilevel meta-analyses. The highest Type I error rate of multi-level meta-analysis models achieved across all scenarios was about 8.2% [Mean (Median) Error Rates: 6.42% (6.42%)], which seems marginal in absolute terms, but relative to the nominal rate of 5% constitutes an increase of 64%. Further, Moeyaert et al. (2017) noted that when effect sizes from the same studies are not correlated, the two-step method with a weighted paper mean provided confidence intervals that were too wide (inefficient), which was also the case in Song et al.'s simulation. Based on their results, Moeyaert et al. (2017) recommend both multilevel meta-analysis and robust variance estimation methods but advised against the averaging method.

The differences in recommendations between Moeyaert et al. (2017) and Song et al. (2020) may have originated from a well-known issue in linear mixed-effects models, of which multilevel meta-analysis is a special type (Nakagawa & Santos 2012); that is, for linear mixed-effects (multilevel) models, it is difficult to determine the appropriate degrees of freedom, which is required for CI calculations. The SAS procedure used by Moeyaert et al. (2017) implements a method for calculating the degrees of freedom that is more appropriate for smaller sample sizes, while R's *metafor* used by Song et al. (2020) is yet to do so (at the moment, it simply sets the degrees of freedom equal to the total number of effect sizes minus 1). Indeed, Song et al. (2020) suspected this shortcoming by stating "this issue is addressed by adjusting the degrees of freedom... ", but did not explore the possible corrections. If this is the issue, there are several solutions that already exist to correct Type I error rates toward the nominal value. Another potential cause of the difference between Song et al. (2020) and Moeyaert et al. (2017) is that the former modeled heteroscedasticity among papers, while the latter did not. Here, we expand Song et al. (2020)'s simulations to show how currently available tools can resolve many of the issues they identified without the need to resort to averaging methods. Before doing so, we would first like to review the reasons for why, in the past, we have strongly recommend the use of multilevel/hierarchical meta-analytic models over a method that averages multiple effect sizes per paper.

Why Multilevel/Hierarchical Models over Averaging?

Nakagawa and Santos (2012) recommended the use of the following meta-analytic model for datasets which can include effect sizes across different species as well as different papers:

$$y_i = \mu + a_k + s_k + p_j + e_i + m_i, \tag{1}$$

where:

A. y_i is the ith effect size ($i = 1, \ldots, N_{effect\text{-}size}$; the number of effect sizes),

B. μ is the meta-analytic mean,

C. a_k is the phylogenetic effect for the kth species ($k = 1, \ldots, N_{species}$), which is distributed with $\mathrm{N}(0, \sigma_a^2\mathbf{A})$, where **A** is a correlation matrix derived from a phylogenetic tree for species included in a meta-analysis,

D. s_k is the non-phylogenetic (species) effect for the kth species, distributed according to $\mathrm{N}(0, \sigma_s^2)$,

E. p_j is the jth paper effect ($j = 1, \ldots, N_{paper}$), distributed according to $\mathrm{N}(0, \sigma_p^2)$,

F. e_i is the ith effect-size specific effect, distributed according to $\mathrm{N}(0, \sigma_e^2)$, and

G. m_i is the ith sampling error effect, distributed according to $N(0, \sigma_i^2)$ where σ_i^2 is the sampling error variance for the ith effect size (note when sampling errors are correlated, a variance-covariance matrix can replace σ_i^2; see below).

Although Song et al. (2012) did not mention phylogenetic non-independence (shown in Equation 1), this issue commonly arises in ecological meta-analysis, and is similar in manner to non-independence due to effects being derived from the same source paper. What is more, it is often important to appropriately take phylogeny into consideration in a meta-analysis (Chamberlain et al. 2012). If we follow the logic of averaging, and we want to avoid using multilevel meta-analysis, we need to average per species. Nakagawa and Santos (2012) put forward three main arguments against averaging (similar arguments were independently put forward by Cheung 2014): 1) the potential loss of statistical power and needlessly large standard errors for the overall effect, 2) the loss of information resulting from not being able to estimate within-paper (within-study) variance, and 3) perhaps most importantly, not being able to estimate ecologically important moderator effects given that aggregation will reduce the information content dramatically (i.e. removes within-species variation in estimated effects).

Both simulation studies suggest the first argument may not apply unless correlations among effect sizes are close to zero. Importantly, Song et al. (2012) discuss two scenarios where dependence among effect sizes could arise: 1) “because they were observed in the same experiment or may have been based on the same subjects” and 2) “even if they arose from separate experiments because experiments likely share common methods, contexts, or other characteristics that influence the effect size". However, Song et al. (2020) only focused on the latter scenario in their simulation not for the former. For the former, where sampling errors are correlated, we need to use the following formula (Borenstein et al. 2009) to obtain a sampling error variance (or a sampling standard error) to accompany a weighted mean, rather than the fixed-effect model used by Song et al. (2012):

$$\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \left(\frac{1}{n}\right)^2 \left(\sum_{i=1}^{n} \sigma_i^2 + \sum_{i \neq g}^{n} r_{ig}\sqrt{\sigma_i^2 \sigma_g^2}\right) \qquad (2),$$

where y_i is the ith effect size ($i = 1, \ldots, n$ and $g = 1, \ldots, n$, where n is the number of effect size within a paper to be combined), σ_i^2 and σ_g^2 are the sampling error variances for y_i and y_g, and r_{ig} is the correlation between the sampling errors of y_i and y_g (note that one can use the function, *aggregate* in *metafor* to calculate a weighted mean and accompanying sampling variance as in Equation 2). We believe that both types of non-independence frequently co-occur and need to be accounted for. For the multilevel meta-analysis, we can model the variance-covariance matrix of the sampling errors for the former type of non-independence as well as model a random effect for paper, although r_{ig} is often not known and needs to be assumed (detailed in Noble et al. 2017; see also, Lajeunesse, 2009; 2011. Further, beyond these two types of multilevel meta-analytic models we can model different sources of non-independence (e.g., phylogenetic relatedness and species relatedness not due to phylogeny; see Equation 1) simultaneously and flexibly, although more data is required for more complex models (Nakagawa and Santos 2012).

The loss of information is a more serious issue, especially the loss of moderator information. The high heterogeneity observed in ecological meta-analyses (Senior et al. 2016) often implies that ecologists must use meta-regression models, which use moderators (or 'predictors') to explain variation among effect sizes. In many cases, meta-regression models are likely to be more useful and informative in ecology than simple meta-analytic models (Gurevitch et al. 2017). Indeed, meta-regression can provide us with review- or synthesis-generated evidence which cannot be obtained via single studies (Nakagawa et al. 2017). If we extend Song et al. (2020)'s recommendation of not using multilevel meta-analyses to 'multilevel meta-regression', this would severely limit our ability to test moderator effects. For example, it is common to obtain separate effect sizes for males and females from one paper. If we aggregate these effect sizes per paper then we would not be able to test sex-specific effects, which runs counter to recent movements to test ubiquitous sex effects (Tannenbaum et al. 2019; Zajitschek et al. 2020).

Solutions for Type I Errors in Multilevel Meta-analysis without the Need for Averaging

Alongside the methods (referred to as Methods 1-5) used by Song et al. (2020), we explored four further methods that are known to overcome the slight excess in Type I error rates observed when using multi-level meta-analytic models. Our simulations reproduced the simulations by Song et al. (2020) but added: 1) a simple correction to the degrees of freedom used to calculate the overall effect size confidence intervals. This involved simply using one less than the total number of papers instead of the typical degrees of freedom that uses the total number of effect sizes (i.e., df = total papers – 1); 2) a Satterthwaite approximation to the effective degrees of freedom, which is commonly applied in the linear mixed effect model literature (Satterthwaite, 1946); 3) a second cluster-robust estimation method implemented in the *clubSandwich* package in R (Pustejovsky, 2020) that uses a bias-reduced linearization method (Pustejovsky and Tipton, 2018). The R package *clubSandwich* uses a similar robust-variance estimation method as *robumeta* (Fisher et al. 2017) used in Song et al. (2020), but can be applied to *metafor's rma.mv* model objects; and 4) a Bayesian modelling approach that uses an MCMC algorithm (using the R package *MCMCglmm* – Hadfield, 2010), instead of restricted maximum likelihood (REML) estimation, as MCMC algorithms are known to have robust coverage, albeit are slightly conservative with small sample sizes (Pappalardo et al. 2020). We also explored other modelling approaches, but present these four as they are simple solutions that can be easily implemented. We focus exclusively on coverage / error rates given that bias was unaffected by the different modelling approaches in Song et al. (2020)'s simulations. For each method (the five existing methods from Song et al. 2020) plus the four new approaches we describe above, we ran 5,000 iterations across all the scenarios detailed in Song et al. (2020). An updated set of scripts from Song et al. (2020), including a coding correction, that implements these new methods can be found at https://(Fisher et al. 2017).

Our new simulation results (Figure 1) show that the four proposed solutions perform quite well across all the scenarios described by Song et al. (2020). The overall performance of each method for each specific simulation scenario is provided in Figure 1B, which reproduces Figure 3 (Experiment 1 and 2) from Song et al. (2020). Overall, the simple approaches we implemented corrected the excess in Type I error rates in the multi-level meta-analytic models implemented in *metafor* (Figure 1A). In particular, Bayesian methods, while having inflated Type II error under small sample situations, perform extremely well across a variety of conditions (Figure 1A and B), with average Type I error rates converging on the 5% level but being slightly conservative overall [Mean (SD) = 4.82% (0.0053)]. A Satterthwaite approximation to the effective degrees of freedom also performs quite well under a variety of conditions as expected [Figure 1A & B – 5.02% (0.0046)]; even the simplest degrees of freedom correction that uses total papers minus one performs quite well [Figure 1A & B – 5.39% (0.0059]. Considering these results above, and the ease of implementation, we recommend fitting a multilevel model with a robust variance estimator because it can easily be applied to multilevel-meta-analytic models in *metafor*. Also, one can certainly use Bayesian modelling, as long as the dataset is large enough (e.g., > 100 effect sizes). A step-by-step guide to implement both of these methods can be found at https://github.com/daniel1noble/ecology_comment.

Conclusion

We appreciate the thorough simulations conducted by Song et al. (2020) in an attempt to better understand the ways in which meta-analysts can overcome one of the most common challenges of meta-analysis; dealing with non-independent effect sizes. While we agree with their recommendation of using robust variance estimation methods (with caveats), we disagree with their recommendation that averaging effect sizes within studies is a solution. While we recognise that there may be times when averaging effect sizes is easier (e.g., when there are very few studies with repeated effects), one most likely needs to use Equation 2 above, not the method of averaging suggested by Song et al. (2020). Regardless, averaging effect sizes within studies comes with a number of significant disadvantages that include: 1) not being able to control for additional sources of non-independence, such as phylogenetic non-independence, which will be commonplace in ecological meta-analyses and 2) not being able to understand the drivers of effect size heterogeneity given that moderator information, which could be included in meta-regression models, is lost. As we have shown, there are a number of very simple, and easily implemented solutions to correct any inflated Type I error rates to their nominal level. Indeed, even robust variance estimators can readily be incorporated into multilevel meta-analytic models, which we recommend ecologists employ. Ignoring these elements prevents meta-analysts from answering a richer set of biologically relevant questions about the drivers underlying effect size variability. As such, we argue strongly against averaging effect sizes within a paper whenever possible.

Acknowledgement

SN and DWAN are supported by an ARC (Australian Research Council) Discovery grant (DP200100367). AMS is supported by an ARC fellowship (DE180101520).

Literature Cited

Borenstein, M., Hedges, L.V., Higgins, J.P.T. & Rothstein, H.R. 2009. Introduction to meta-analysis. Wiley, Oxford.

Chamberlain, S. A., S. M. Hovick, C. J. Dibble, N. L. Rasmussen, B. G. Van Allen, B. S. Maitner, J. R. Ahern, L. P. Bell-Dereske, C. L. Roy, M. Meza-Lopez, J. Carrillo, E. Siemann, M. J. Lajeunesse, and K. D. Whitney. 2012. Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. Ecology Letters **15**:627-636.

Cheung, M. W. L. 2014. Modeling Dependent Effect Sizes With Three-Level Meta-Analyses: A Structural Equation Modeling Approach. Psychological Methods **19**:211-229.

Fisher, Z., E. Tipton, and H. Zhipeng. 2017. robumeta: Robust variance meta-regression. R package version 2.0.

Hadfield, J. D. 2010. MCMC methods for multi-response Generalised Linear Mixed Models: the MCMCglmm R package. Journal of Statistical Software **33**:1-22.

Lajeunesse, M. J. 2009. Meta-Analysis and the Comparative Phylogenetic Method. American Naturalist **174**:369-381.

Lajeunesse, M. J. 2011. On the meta-analysis of response ratios for studies with correlated and multi-group designs. Ecology **92**:2049-2055.

Moeyaert, M., M. Ugille, S. N. Beretvas, J. Ferron, R. Bunuan, and W. Van den Noortgate. 2017. Methods for dealing with multiple outcomes in meta-analysis
a comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. International Journal of Social Research Methodology **20**:559-572.

Nakagawa, S., D. W. Noble, A. M. Senior, and M. Lagisz. 2017. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. BMC Biology **15**:18.

Nakagawa, S., and E. S. A. Santos. 2012. Methodological issues and advances in biological meta-analysis. Evolutionary Ecology **26**:1253-1274.

Noble, D. W. A., M. Lagisz, E. O'Dea R, and S. Nakagawa. 2017. Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. Molecular Ecology **26**:2410-2425.

Pappalardo, P., K. Ogle, E. A. Hamman, J. R. Bence, B. A. Hungate, and C. W. Osenberg. 2020. Comparing traditional and Bayesian approaches to ecological meta-analysis. Methods in Ecology and Evolution **11**:1286-1295.

Pustejovsky, J. 2017. clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. R package version 0.2. 3. R Found. Stat. Comput., Vienna.

Pustejovsky, J. E., and E. Tipton. 2018. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. Journal of Business & Economic Statistics **36**:672-683.

Satterthwaite, F. E. 1946. An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin **2**:110-114.

Song, C., S. D. Peacor, C. W. Osenberg, and J. R. Bence. 2020. An assessment of statistical methods for nonindependent data in ecological meta-analyses. Ecology **n/a**:e03184.

Tannenbaum, C., R. P. Ellis, F. Eyssel, J. Zou, and L. Schiebinger. 2019. Sex and gender analysis improves science and engineering. Nature **575**:137-146.

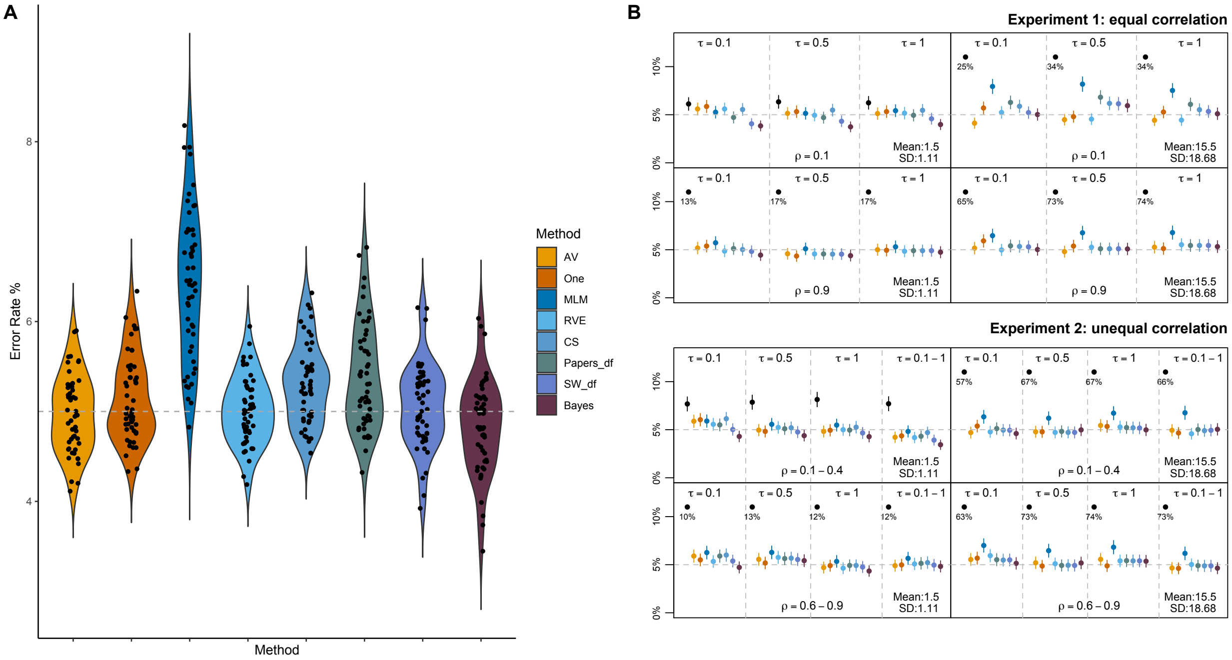
Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software **36**:1-48.

Zajitschek, S. R., F. Zajitschek, R. Bonduriansky, R. C. Brooks, W. K. Cornwell, D. S. Falster, M. Lagisz, J. Mason, A. M. Senior, D. W. A. Noble, and S. Nakagawa. 2020. Sex and Power: sexual dimorphism in trait variability and its eco-evolutionary and statistical implications. eLife.

Figure captions

Figure 1 – A) Density distribution of average Type I error rates (%) across all 51 scenarios simulated by Song et al. (2020). The first four methods reproduce Song et al. (2020)'s simulations and include: "One" = Choosing a single effect size; "AV" = two-step method that averages effects within a study; "MLM" = Multi-level meta-analytic model; "RVE" = Robust variance estimation method with *robumeta*. In addition to these, we implemented four new methods to correct the slight increase in Type I error rates for the MLM method. These included: "CS" = club sandwich robust variance estimation; "Papers_df" = degrees of freedom equal to the total number of papers minus one to adjust confidence intervals from MLM; "SW_df" = Satterthwaite degrees of freedom to adjust the confidence intervals of the MLM, and "Bayes" = Bayesian estimation methods. See details in text. Raw error rates across all simulated scenarios described by Song et al. (2020) are depicted by black points. Grey dashed line represents the nominal 5% error rate. Note that the method ignoring non-independence is not included here (see Figure S1). B) Average Type I error rates (%) across a sub-sample of scenarios simulated by Song et al. (2020). Note that the sub-sample of simulation scenarios matches those presented in Song et al. (2020) and does not include all 51 simulation scenarios presented in panel A. Colors match methods described in panel A, except we also present the original Method 1 as denoted in black, which completely ignores non-independence.

Figure 1


A
Error Rate %
8
6
4
Method
Method
AV
One
MLM
RVE
CS
Papers_df
SW_df
Bayes
B
Experiment 1: equal correlation
τ = 0.1
τ = 0.5
τ = 1
10%
5%
0%
25%
34%
34%
13%
17%
17%
65%
73%
74%
ρ = 0.1
ρ = 0.9
Mean:1.5
SD:1.11
Mean:15.5
SD:18.68
Experiment 2: unequal correlation
τ = 0.1 – 1
57%
67%
67%
66%
10%
13%
12%
12%
63%
73%
74%
73%
ρ = 0.1 – 0.4
ρ = 0.6 – 0.9