Title: An Assessment of Statistical Methods for Non-independent Data in Ecological Meta-analyses

Running title: Non-independence in meta-analysis

Authors: Chao Song[1,*], Scott D. Peacor[1], Craig W. Osenberg[2], James R. Bence[1]

Affiliations:

[1]Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan 48824, USA

[2]Odum School of Ecology, University of Georgia, Athens, Georgia 30602, USA.

[*]Correspondence: chaosong@msu.edu

Abstract

In ecological meta-analyses, non-independence among observed effect sizes from the same source paper is common. If not accounted for, non-independence can seriously undermine inferences. We compared the performance of four meta-analysis methods that attempt to address such non-independence and the standard random-effect model that ignores non-independence. We simulated data with various types of within-paper non-independence, and assessed the standard deviation of the estimated mean effect size and type I error rate of each method. Although all four methods performed substantially better than the standard random-effects model that assumes independence, there were differences in performance among the methods. A two-step method that first summarizes the multiple observed effect sizes per paper using a weighted mean and then analyzes the reduced data in a standard random-effects model, and a robust variance estimation method performed consistently well. A hierarchical model with both random paper and study effects gave precise estimates but had a higher type I error rates, possibly reflecting limitations of currently available meta-analysis software. Overall, we advocate the use of the two-step method with a weighted paper mean and the robust variance estimation method as reliable ways to handle within-paper non-independence in ecological meta-analyses.

Keywords

meta-analysis, non-independence, pseudoreplication, random effect, hierarchical model, robust variance estimation.

Introduction

Meta-analysis is a quantitative synthesis method that combines individual studies to quantify the overall effect and the heterogeneity in effects among studies. Since its introduction to ecology (Jarvinen 1991, Arnqvist and Wooster 1995), meta-analysis has played an increasingly influential role in the field, such as testing ecological theories, identifying research directions, and informing conservation and management strategies (Stewart 2009, Cadotte et al. 2012, Gurevitch et al. 2018). Given its wide application and large impact, rigorous methodology is crucial (Osenberg et al. 1999, Lortie et al. 2015). Although statistical methods and specialized software for meta-analysis have advanced greatly over the past few decades, many statistical issues still remain (Gurevitch and Hedges 1999, Nakagawa and Santos 2012, Koricheva and Gurevitch 2014).

A prevalent statistical issue in meta-analysis is non-independence among observed effect sizes (Gurevitch and Hedges 1999, Nakagawa et al. 2017, Noble et al. 2017). A common type of non-independence structure arises when observed effect sizes come in identifiable groups, where they are non-independent within groups but independent across groups. This type of non-independence has been called pseudoreplication and can seriously undermine statistical inferences (Hurlbert 1984). Many mechanisms, such as shared experimental subjects, common experimental time/sites, or similar methodology, could lead to this type of non-independence and result in varying strength of correlation among observed effect sizes in the group (Noble et al. 2017). One of the most common way such a group arises is when single source papers consist of multiple studies, i.e., yield multiple observed effect sizes. Studies from the same source paper and the resulting observed effect sizes arising from them can be thought of as comprising a group. Here, we define a study as the experimental/observational procedures and the resulting set of data that lead to a single observed effect size. In this paper, we address non-independence within source papers, although our results likely apply to other types of group or hierarchical structures that may also generate non-independence, e.g., studies from the same lab group or geographic locations.

Within-paper non-independence is ubiquitous in ecological meta-analysis (Noble et al. 2017). For example, a source paper used in a meta-analysis may include multiple responses measured in the same experiment, such as biomass, growth rate, and fecundity. Observed effect sizes from this paper will be non-independent because they were observed in the same experiment or may have been based on the same subjects. Observed effect sizes from the same paper could also be non-independent even if they arose from separate experiments because experiments likely share common methods, contexts, or other characteristics that influences the effect size, e.g., studies from the same paper might all be done at the same geographic location or in the same time period. Because results of ecological research often depend strongly on the ecological and methodological context, we can expect non-independence among observed effect sizes from the same paper to be common. While non-independence does not lead to bias in parameter estimation in general, ignoring non-independence usually leads to incorrect estimates of uncertainty, which in turn can invalidate hypothesis tests (Kwok et al. 2007). The extent of the inferential problems resulting from ignoring non-independence of observed effect sizes within papers will depend on the nature of the non-independence and how studies are distributed among paper.

While it is ideal to explicitly incorporate the non-independence structure in the meta-analysis model, information necessary to model the exact non-independence structure is often unavailable from source papers. Thus, analysts typically use omnibus strategies for addressing within-paper non-independence. The first strategy is a two-step method. Analysts first derive a single summary effect size for each paper based on the multiple observed effect sizes in that paper and then analyze the summary effect sizes using standard meta-analysis methods that assume independence (Rosenthal and Rubin 1986, Marín-Martínez and Sánchez-Meca 1999). The summary effect size might be obtained by randomly choosing one of the observed effect sizes from each paper, or it could be derived as the mean of the observed effect sizes in a paper. The second strategy is to include a random paper effect in addition to the random study effect in the meta-analysis model, assuming such a hierarchical model can approximately model the actual pattern of non-independence. More recently, a third strategy, known as the robust variance estimation, was developed (Hedges et al. 2010, Tipton 2015). This method extends the work on robust variance estimators (Huber 1967, White et al. 1980) to meta-analysis and does not require knowledge of the non-independence structure among observed effect sizes within groups.

These methods make different assumptions about how observed effect sizes from the same source paper are correlated. For example, including a random paper and study effect (both commonly assumed to be independent and identically distributed random variables following normal distributions) is equivalent to assuming that the observed effect sizes within the same paper are positively correlated, with an equal correlation coefficient for each pair. However, one might expect the correlation to vary among pairs of studies. This could arise, for example, if some studies within a source paper were conducted closer in space and/or time and thus have more similar ecological settings (Noble et al. 2017). While methods exist to explicitly model spatial and temporal correlations, these methods require knowledge of the timing and spatial locations, which is often not reported. In these situations, a method that allows variable and unknown correlation among pairs of observed effect sizes, such as the robust variance estimation method, may perform better. To determine the best methods among those that are available, it is critical to evaluate these methods under different scenarios of within-paper non-independence.

Despite the ubiquity of non-independence within papers, the performance of various methods attempting to address this issue has not been comprehensively evaluated in the context of ecological meta-analysis. In this study, we performed simulation experiments to examine the effectiveness of different methods used to address non-independence with the intent of providing practical guidance on choosing appropriate methods for ecological meta-analysis. Specifically, we simulated datasets that had a hierarchical structure, with multiple studies within each source paper used in the meta-analysis. The simulated data ranged from no correlation to strong but unequal correlation among observed effect sizes within the same source paper, and allowed for plausible variation in the number of observed effect sizes per source paper. We applied five analytic methods to each simulated data set and assessed their performance: four methods that have been proposed to handle non-independence within papers as well as the all-too-common method of simply ignoring the issue.

Methods

We simulated meta-analyses consisting of data from 20 papers, each containing a number of studies (Fig 1). For each study, we simulated replicated control and treatment groups, with data from each source paper simulated to obtain various patterns of non-independence among observed effect sizes within the paper. For each study, we calculated a log response ratio and its estimated variance. The log response ratio is the most commonly used effect size metric in ecology (Nakagawa and Santos 2012), but our qualitative results should apply to other metrics as well. We estimated the overall mean effect size using alternative meta-analysis methods that differ in how they account for non-independence and compared the their performance. We conducted two sets of simulation experiments (Fig. A1). In the first experiment, observed effect sizes from the same source paper were correlated with the same correlation coefficients for all pairs. In the second experiment, we varied the correlation between pairs of observed effect sizes.

Simulation of data for an individual study

We simulated a response variable y in the control and treatment group for each study as

$$y_{cijk} = \mu\varepsilon_{cijk} \tag{1}$$

$$y_{tijk} = \alpha_{ij}\mu\varepsilon_{tijk}, \tag{2}$$

where y_{cijk} and y_{tik} are the response variables for the kth replicates in the control and treatment group of study j in paper i, ε_{cijk} and ε_{tijk} are random errors following log-normal distributions, and α_{ij} is the multiplicative treatment effect, which can be decomposed as

$$\alpha_{ij} = \alpha e_{ij}. \tag{3}$$

Here, e_{ij} represents the study-specific random deviation from the mean treatment effect and is assumed to follow a log-normal distribution. Once data for each study were simulated, we calculated a log response ratio for each study (θ_{ij}) as $\log(\overline{y_{tij}}/\overline{y_{cij}})$, and its variance as $\mathrm{var}(y_{tij})/n_{tij}\overline{y_{tij}}^2 + \mathrm{var}(y_{cij})/n_{cij}\overline{y_{cij}}^2$, where n_{cij} and n_{tij} are the number of replicates in the control and treatment groups for study j within source paper i (Hedges et al. 1999).

Patterns of non-independence in the simulations

Based on this simulation approach, the log response ratio for each study, θ_{ij}, follows a normal distribution asymptotically with mean $\log(\alpha) + \log(e_{ij})$ (Hedges et al. 1999). Thus, we can express θ_{ij} simulated by equations 1–3 as

$$\theta_{ij} = \log(\alpha) + \log(e_{ij}) + \epsilon_{ij}. \tag{4}$$

Equation 4 matches a random-effects meta-analysis model. Here, $\log(\alpha)$ is the mean effect size, $\log(e_{ij})$ is the random study effect, and ϵ_{ij} is the within-study error. Both $\log(e_{ij})$ and ϵ_{ij} are normally distributed.

Non-independence among θ_{ij} within papers may occur through correlations among $\log(e_{ij})$ and/or ϵ_{ij}. Typically, correlation among ϵ_{ij} arise from using shared control or measuring the same group of subjects in studies with multiple endpoints. The resulting correlation structure in ϵ_{ij} can be explicitly calculated and incorporated in the meta-analysis model (Gleser and Olkin 2009, Lajeunesse 2011). Therefore, we assumed independence among ϵ_{ij} and only considered non-independence arising from correlation among the random study effects, i.e. $\text{cov}(\theta_{ij}, \theta_{ik}) = \text{cov}(\log(e_{ij}), \log(e_{ik}))$ for study j and k in paper i.

We conducted simulation experiments with two patterns of non-independence. In the first experiment, the random study effects from the same source paper were equally correlated (i.e.,$\text{cov}(\theta_{ij}, \theta_{ik}) = \rho\tau^2$, where ρ is the correlation coefficient between each pair of observed effect sizes and τ^2 is the variance of $\log(e_{ij})$). We included a special case of zero correlation (independence) in the first experiment. In the second experiment, the random study effects within a source paper were correlated but the correlation was not equal ($\text{cov}(s_{ij}, s_{ik}) = \rho_{ijk}\tau_i^2$). Here, the correlation coefficient, ρ_{ijk}, was allowed to vary among pairs of observed effect sizes. We subscripted the among-study variance τ^2 because we sometimes allow this quantity to vary among papers in the second experiment. In both experiments, we only considered non-negative correlations, given that similar contexts and shared data for studies from the same source paper would be expected to lead to similar, rather than dissimilar, observed effect sizes within a paper.

Patterns of non-independence in θ_{ij} were simulated by drawing $\log(e_{ij})$ for each source paper from multivariate normal distributions with appropriate covariance matrices. The covariance matrix was generated as **KRK**, where **K** is matrix with τ as the diagonal elements and **R** is the correlation matrix. In the first experiment, **R** was a matrix with 1 at the diagonal positions and the same correlation coefficient ρ at all other positions. In the second experiment, **R** was a symmetric matrix with 1 at the diagonal positions and different correlation coefficients at all others to reflect the fact that studies were unequally correlated. We used the C-vine method proposed in Lewandowski et al. (2009) to generate such correlation matrices.

Details of experimental design

In the simulation experiments, we did not systematically vary the parameters that were not expected to influence how well methods address non-independence. Specifically, we set μ at 10 and α at 1. We simulated $\log(\varepsilon_{tijk})$ from a normal distribution with mean 0 and standard deviation randomly chosen between 0.1–0.3 for each study. The number of replicates for each study, equal for both the control and treatment groups, was chosen with equal probability from integers between 3–20. Finally, we set the number of papers at 20, a relatively low number in ecological meta-analysis. Methods that perform well in this situation are expected to perform at least as well if the meta-analysis contains more papers.

We systematically varied the parameters that we expected to influence the efficacy of methods used to address non-independence, including the number of studies per paper, the magnitude of correlation among observed effect sizes from the same source paper, and the among-study variability. Below, we provide the levels of these parameters and the rationales for these choices.

Number of studies per paper: We examined the frequency distribution of the number of studies per paper (n_i) in 15 published ecological meta-analyses chosen haphazardly (Appendix S1: Section S1). A shifted negative binomial distribution adequately described the frequency distribution of n_i (Appendix S1: Fig. S1): i.e., $n_i - 1$ followed a negative binomial distribution. We also found that the mean and standard deviation of $n_i - 1$ were correlated on the logarithmic scale (Appendix S1: Fig. S2). Therefore, we chose three levels for the mean of $n_i - 1$ spanning the observed range (0.5, 4.5 and 14.5) and calculated the corresponding standard deviation (1.1, 7.0, and 18.7) based on the linear regression. We then drew $n_i - 1$ for each paper from a negative binomial distribution with the mean and standard deviation specified above.

Correlation coefficient: In the first experiment, in which the correlation coefficient (ρ) was the same for all pairs of study within a paper, we set ρ at 0, 0.1, 0.5, and 0.9, ranging from independence to quite strong correlation. In the second experiment where ρ varied among pairs of studies within the same paper, we set the range of ρ for each paper. Because the C-Vine method (Lewandowski et al. 2009) generates a correlation matrix from user-specified partial correlation coefficients, we randomly chose partial correlation coefficients from two uniform ranges: 0.1–0.4 and 0.6–0.9. The resulting ranges for pairwise correlations generally matched these specified ranges for the partial correlations.

Among-study variance: In the first experiment where correlation between pairs of observed effect sizes is equal, we set levels of the standard deviation for among-study variability, τ, at 0.1, 0.5, and 1. The chosen levels represent plausible ranges in ecological meta-analyses. For example, a τ of 1 led to the treatment effect for a particular study, α_{ij}, ranging between 14% to 710% of the mean treatment effect for 95% of the studies. A τ of 0.1 led to a range of 82% to 122%. In addition, our choice of among-study variance and among-replicate variance within a study are also consistent with the typical proportion of within- and among-study heterogeneity in ecological meta-analyses (Senior et al. 2016).

In the second experiment where correlation between pairs of observed effect sizes is unequal, we used the same three levels of τ and added a new scenario in which τ varied among papers. This scenario is plausible in ecological meta-analyses. For example, some papers included in the meta-analysis may contain studies from more diverse environments than others (Hillebrand and Gurevitch 2014). For this scenario, we chose τ for each paper from a uniform distribution between 0.1 and 1.

Methods of meta-analysis for non-independent data

We evaluated five methods commonly used in ecological meta-analysis. The five methods are: 1) the standard random-effects meta-analysis model that ignores non-independence; 2) a two-step method in which we analyze the weighted mean effect size for each paper in a random-effect meta-analysis model. This is equivalent to performing a fixed-effect meta-analysis for each paper and using the resulting means and their standard errors from the fixed effect model in the second step; 3) a two-step method in which we analyze one randomly chosen observed effect size from each paper in a random-effects meta-analysis model; 4) a hierarchical model that included a random paper effect and a random study effect; and 5) a robust variance estimation method for meta-analysis (Hedges et al. 2010).

All random-effects meta-analysis models were implemented using the function “rma” in R (version 3.6.2) package “metafor” (version 2.1) (Viechtbauer 2010) with the variance of the random effects estimated by restricted maximum likelihood (Veroniki et al. 2016). We constructed confidence intervals based on the adjustment proposed by Hartung and Knapp (2001) and Sidik and Jonkman (2002). We implemented the method with a random paper effect using function “rma.mv” in “metafor”, and constructed confidence intervals based on the t-distribution of the Wald statistic. Finally, we implemented the robust variance estimation method using function “robu” in R package “robumeta” with the default weights and adjustment for small sample size (Fisher et al. 2017). Code for the simulation experiments are provided in Data S1.

Metrics for model performance

We evaluated the performance of methods by the precision of the estimated mean effect size and the type I error rate. We calculated the standard deviation of the estimated mean effect size over the 5000 iterations of simulations as the measure of precision. We calculated type I error rate as the percentage of times in the simulations when the 95% confidence interval for the mean effect size did not cover the true value. The confidence interval for the estimated error rate was calculated based on the binomial distributions for the number of falsely significant results in the simulations. None of the methods produced appreciable bias in the estimated mean effect sizes and we therefore do not present results about bias.

Results

Precision of estimated mean effect size

For clarity of presentation, the figures contain a representative subset of results. Full results can be found in Appendix S1: Fig. S3 and S4. The standard random-effects meta-analysis model that assumes independence among observed effect sizes had a low standard deviation when the observed effect sizes were actually independent(Fig. 2), but resulted in a higher standard deviation when observed effect sizes were non-independent (Fig. 2). The loss of precision was more pronounced when the correlation was strong (Fig. 2). Among the four methods that account for non-independence, the two-step method using one randomly chosen study from each paper had a consistently high standard deviation. This problem of low precision, however, was less severe when correlations among observed effect sizes were strong(Fig. 2). The methods that included a random paper effect performed consistently well in terms of precision under all scenarios considered in the simulations. Finally, the two-step method using a weighted paper mean and the robust variance estimation method gave low standard deviations except when observed effect sizes from the same source paper were independent.

Type I error rates

The standard random-effects model that ignored non-independence substantially inflated the type I error rates (Fig. 3), sometimes to over 70%, unless observed effect sizes were independent or only mildly non-independent (i.e., low correlation and very few observed effect sizes per paper) (Fig. 3). Under all scenarios of non-independence, all four methods that accounted for non-independence offered substantial improvement in error rates. Surprisingly, including a random paper effect led to error rates consistently above the correct level of 5% (between 5% and 8%) in the presence of non-independence (Fig. 3). The two-step method that used one study from each paper gave correct error rates consistently. The two-step method using a weighted paper mean and the robust variance estimation method both gave correct error rates when observed effect sizes were non-independent. However, these two methods sometimes generated error rates significantly lower than the correct level of 5% when observed effect sizes were independent (Fig. 3).

Discussion

Non-independence among observed effect sizes from the same source paper is common in ecological meta-analyses and can arise through a variety of mechanisms, such as shared experimental subjects, common experimental sites, or similar methodology (Noble et al. 2017). The variety of mechanisms leading to within-paper non-independence gives rise to different patterns and strength of correlations among observed effect sizes from the same source paper. Our simulations, using ecologically realistic parameter values, represent a broad range of scenarios. We found that treating non-independent data as if they were independent caused error rates that were substantially higher than the correct level (5% for a 95% confidence interval) unless the observed effect sizes were only mildly non-independent (i.e., low mean n_i and ρ) (Fig. 3). Even for the lowest non-zero level of non-independence, the error rate was still non-negligibly above the correct level of 5%. All four methods that accounted for non-independence offered considerable improvements with regard to error rates. In addition, ignoring non-independence led to imprecise estimates of the mean effect size when the correlation among studies was strong. Because meta-analyses in ecology are still often done using methods that ignore non-independence (Gurevitch and Hedges 1999, Nakagawa and Santos 2012, Noble et al. 2017), our study demonstrates an urgent need for meta-analysts to adopt methods that account for this.

The two step method with one randomly chosen study from each paper consistently produced less precise estimates compared to other methods that accounted for non-independence (Fig. 2), presumably because valuable information was discarded using this method. The decrease in precision was sometimes substantial. For example, when observed effect sizes from the same source paper were equally correlated with $\rho = 0.1$, $\tau = 1$, and $\mathrm{E}(n_i) = 15.5$, the standard deviation of the estimated mean effect size based on this method was 0.225 compared to 0.104 based on the method with a random paper effect. The response ratio would be between 0.64–1.55 95% of the time using this method compared to 0.82–1.23 using the method with a random paper effect. Although error rates based on this method were consistently correct (Fig. 3), other methods offered comparable performance in error rate but substantially better performance in precision. As a result, we do not recommend this method as a general way to handle non-independence within papers.

Including a random paper effect consistently inflated the type I error rates. Surprisingly, this method inflated error rates even when it was the correct model (i.e., when observed effect sizes from the same source paper had equal correlation). We speculate that the consistently higher than correct error rates arose from the limitation of the methods for statistical inference in hierarchical models currently implemented in "metafor". Confidence intervals for parameter estimates were constructed based on a t-distribution for the Wald statistic, which is known to cause high error rates, primarily because uncertainty in the standard error estimates is not fully accounted for (Pinheiro and Bates 2000). Although we do not recommend this specific method due to the higher error rates, the issue causing this problem could likely be resolved. In the general mixed-effect model literature, this issue is addressed by adjusting the degrees of freedom for a t- or F-test (Kenward and Roger 1997). Implementation of these inferential methods in hierarchical meta-analysis models could be valuable, considering the high precision (Fig. 2) and the unique advantage of partitioning sources of variation using hierarchical models. To date, no meta-analysis software has these methods implemented. Improvements to metafor could be extremely useful since it has become the most versatile and widely used software for meta-analysis using the frequentist approach.

Both the two-step method with a weighted mean for each paper and the robust variance estimation method controlled error rates well and had similar standard deviations over the range of conditions we explored. There was, however, some cost in terms of low statistical power, as evidenced by error rates significantly lower than 5%, when these methods were applied to data that were actually independent; but that cost disappeared in the presence of non-independence. There are potential shortcomings with these two methods (Appendix S1: Section S2). For example, the robust variance estimation method requires user-specified weights (Hedges et al. 2010). Because weights proportional to the inverse covariance matrix give the most efficient estimates but the user does not know the covariance matrix, this method may be far from optimal and result in less efficient estimates. Additionally, the robust variance estimation method is asymptotic and thus requires sufficient number of papers to be effective. The two-step method using the mean for each paper also has potential limitations. For example, the variance of the true mean for each paper will generally vary among papers, with means from papers with many studies having lower variance. This heterogeneity is not accounted for. These issues, however, do not appear to substantially influence the performance of these two methods. Taken together, we suggest either using the robust variance estimation method or the two-step method starting with a weighted mean for each paper to handle non-independence within papers, at least when conditions (i.e., number of papers, number of studies per paper, levels of variation, and degrees of non-independence) are expected to be similar to the conditions we explored in the simulations.

Acknowledgements

This work was supported by the U.S. National Science Foundation (DEB–1655394 and DEB–1655426). JRB and SDP received support from AgBioResearch of Michigan State University. The Institute for Cyber-Enabled Research at Michigan State University provided computational resources. This is publication 2020-XX of the Quantitative Fisheries Center at Michigan State University.

References

Arnqvist, G. and D. Wooster. 1995. Meta-analysis: synthesizing research findings in ecology and evolution. Trends in Ecology & Evolution, **10**:236–240.

Cadotte, M. W., L. R. Mehrkens, and D. N. Menge. 2012. Gauging the impact of meta-analysis on ecology. Evolutionary Ecology, **26**:1153–1167.

Fisher, Z., E. Tipton, and Z. Hou. 2017. robumeta: Robust variance meta-regression. R package version 2.1.

Gleser, L. J. and I. Olkin. 2009. Stochastically dependent effect sizes. In H. Coopers, L. V. Hedges, and J. C. Valentine, editors, The handbook of research synthesis and meta-analysis, pages 357–376. Russell Sage Foundation, New York, NY, 2nd edition.

Gurevitch, J. and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. Ecology, **80**:1142–1149.

Gurevitch, J., J. Koricheva, S. Nakagawa, and G. Stewart. 2018. Meta-analysis and the science of research synthesis. Nature, **555**:175–182.

Hartung, J. and G. Knapp. 2001. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Statistics in Medicine, **20**:1771–1782.

Hedges, L. V., J. Gurevitch, and P. S. Curtis. 1999. The meta-analysis of response ratios in experimental ecology. Ecology, **80**:1150–1156.

Hedges, L. V., E. Tipton, and M. C. Johnson. 2010. Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, **1**:39–65.

Hillebrand, H. and J. Gurevitch. 2014. Meta-analysis results are unlikely to be biased by differences in variance and replication between ecological lab and field studies. Oikos, **123**:794–799.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In L. M. Le Cam and J. Neyman, editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 221–233.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs, **54**:187–211.

Jarvinen, A. 1991. A meta-analytic study of the effects of female age on laying-date and clutch-size in the Great Tit *Parus major* and the Pied Flycatcher *Ficedula hypoleuca*. Ibis,

133:62–67.

Kenward, M. G. and J. H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, **53**:983–997.

Koricheva, J. and J. Gurevitch. 2014. Uses and misuses of meta-analysis in plant ecology. Journal of Ecology, **102**:828–844.

Kwok, O.-m., S. G. West, and S. B. Green. 2007. The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. Multivariate Behavioral Research, **42**:557–592.

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

Lewandowski, D., D. Kurowicka, and H. Joe. 2009. Generating random correlation matrices based on vines and extended onion method. Journal of Multivariate Analysis, **100**:1989–2001.

Lortie, G., Christopher J. ad Stewart, H. Rothstein, and J. Lau. 2015. How to critically read ecological meta-analysis. Research Synthesis Methods, **6**:124–133.

Marín-Martínez, F. and J. Sánchez-Meca. 1999. Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. The Spanish Journal of Psychology, **2**:32–38.

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

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

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

2425.

Osenberg, C. W., O. Sarnelle, S. D. Cooper, and R. D. Holt. 1999. Resolving ecological questions through meta-analysis: goals, metrics, and models. Ecology, **80**:1105–1117.

Pinheiro, J. and D. Bates. 2000. Mixed-effects models in S and S-PLUS. Springer, New York, NY.

Rosenthal, R. and D. B. Rubin. 1986. Meta-analytic procedures for combining studies with multiple effect sizes. Psychological Bulletin, **99**:400–406.

Senior, A. M., C. E. Grueber, T. Kamiya, M. Lagisz, K. O'dwyer, E. S. Santos, and S. Nakagawa. 2016. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. Ecology, **97**:3293–3299.

Sidik, K. and J. N. Jonkman. 2002. A simple confidence interval for meta-analysis. Statistics in Medicine, **21**:3153–3159.

Stewart, G. 2009. Meta-analysis in applied ecology. Biology Letters, **6**:78–81.

Tipton, E. 2015. Small sample adjustments for robust variance estimation with meta-regression. Psychological Methods, **20**:375.

Veroniki, A. A., D. Jackson, W. Viechtbauer, R. Bender, J. Bowden, G. Knapp, O. Kuss, J. P. Higgins, D. Langan, and G. Salanti. 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods, **7**:55–79.

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

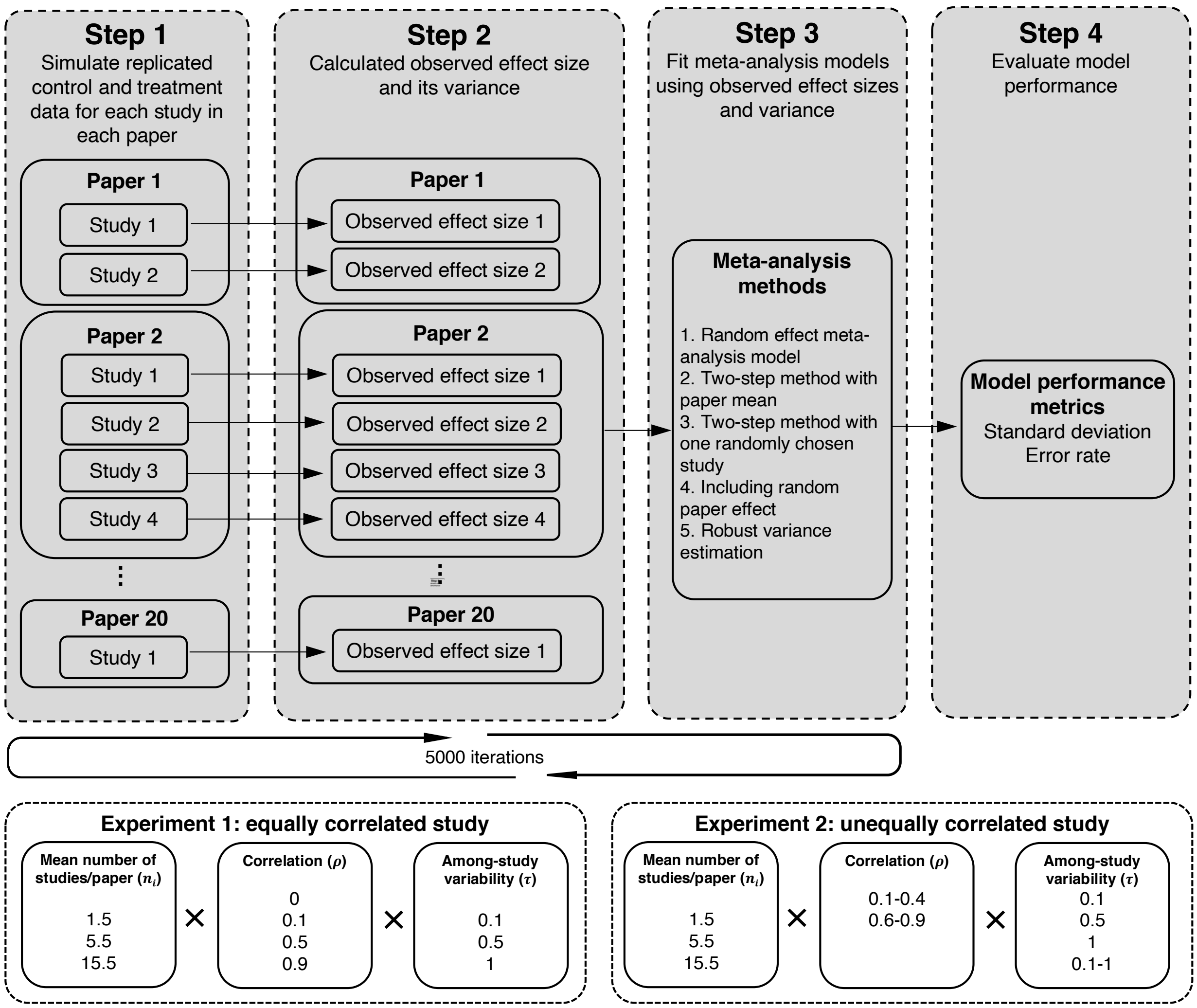
White, H. et al. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, **48**:817–838.

Figure 1 Diagram of the experimental design. In experiment 1, observed effect sizes from the same source paper were correlated with the same correlation coefficient for all pairs. In experiment 2, the magnitude of correlation varied for pairs of observed effect sizes. We systematically varied mean number of studies per paper (n_i), correlation among observed effect sizes within the same paper (ρ), and among-study variability (τ). For each combination of the experimental factor levels, steps 1–3 were repeated 5000 times. The estimated mean effect sizes for each of the five methods over the 5000 iterations were used to quantify the standard deviation of the estimates and the type I error rate in step 4.

Figure 2: Standard deviations of the estimated mean effect size based on the five meta-analysis methods. The mean and standard deviation of the distribution for the number of studies per paper, the among-study standard deviation (τ), and the correlation coefficient among observed effect sizes from the same paper (ρ) are noted on each panel. Methods 1–5 are 1) random-effect meta-analysis model, 2) two-step method using a weighted mean from each paper, 3) two-step method with one randomly chosen observed effect sizes from each paper, 4) meta-analysis with random paper and study effects, and 5) robust variance estimation method.

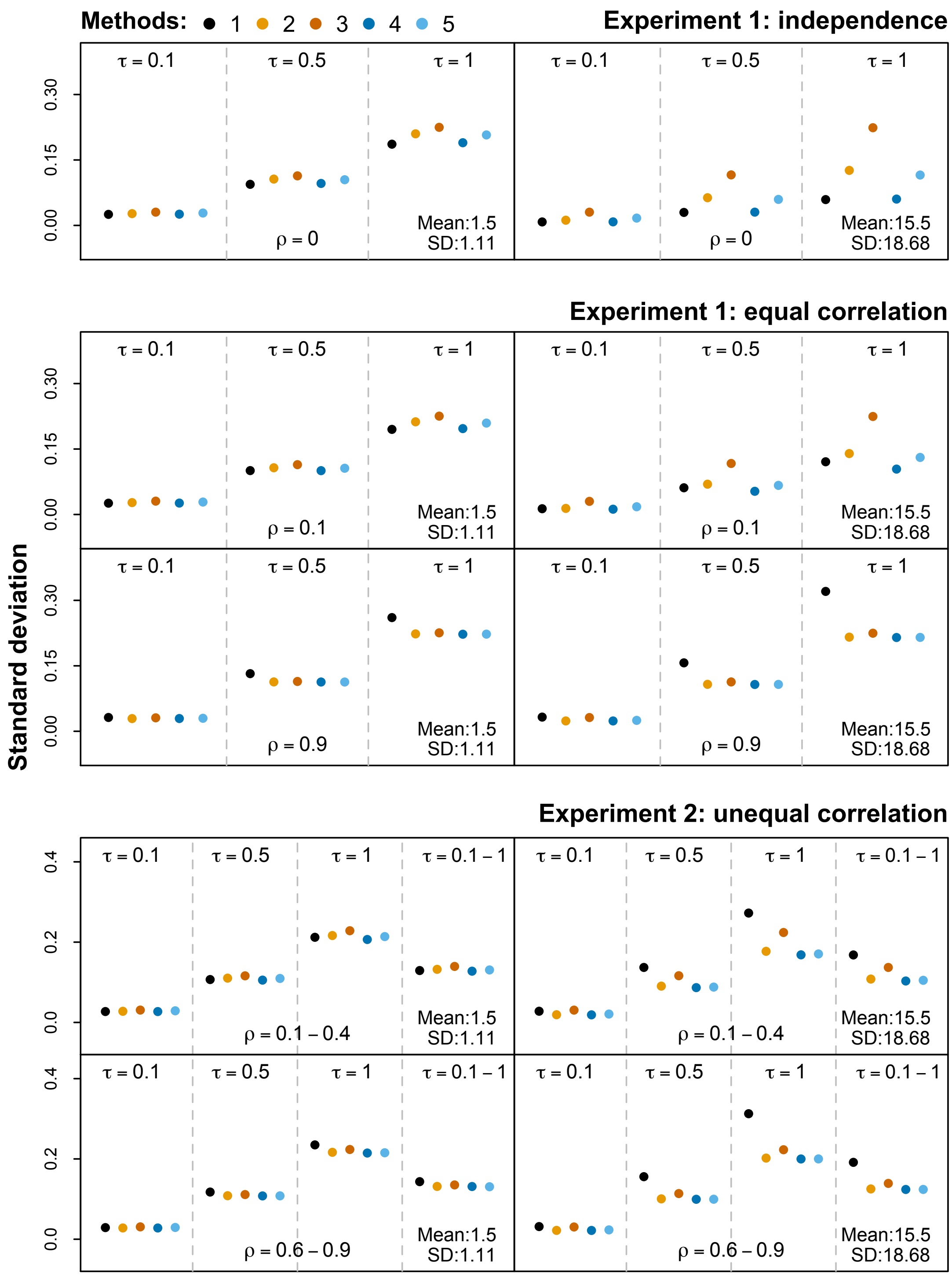
Figure 3: Type I error rates based on the five meta-analysis methods. Error bars are 95% confidence intervals. Error rates exceeding 10% are indicated with the actual error rates. The mean and standard deviation of the distribution for the number of studies per paper, the among-study standard deviation (τ), and the correlation coefficient among observed effect sizes within the same paper (ρ) are noted on each panel. Methods 1–5 are 1) random-effect meta-analysis model, 2) two-step method using a weighted mean from each paper, 3) two-step method with one randomly chosen observed effect sizes from each paper, 4) meta-analysis with random paper and study effects, and 5) robust variance estimation method.

Figure 1



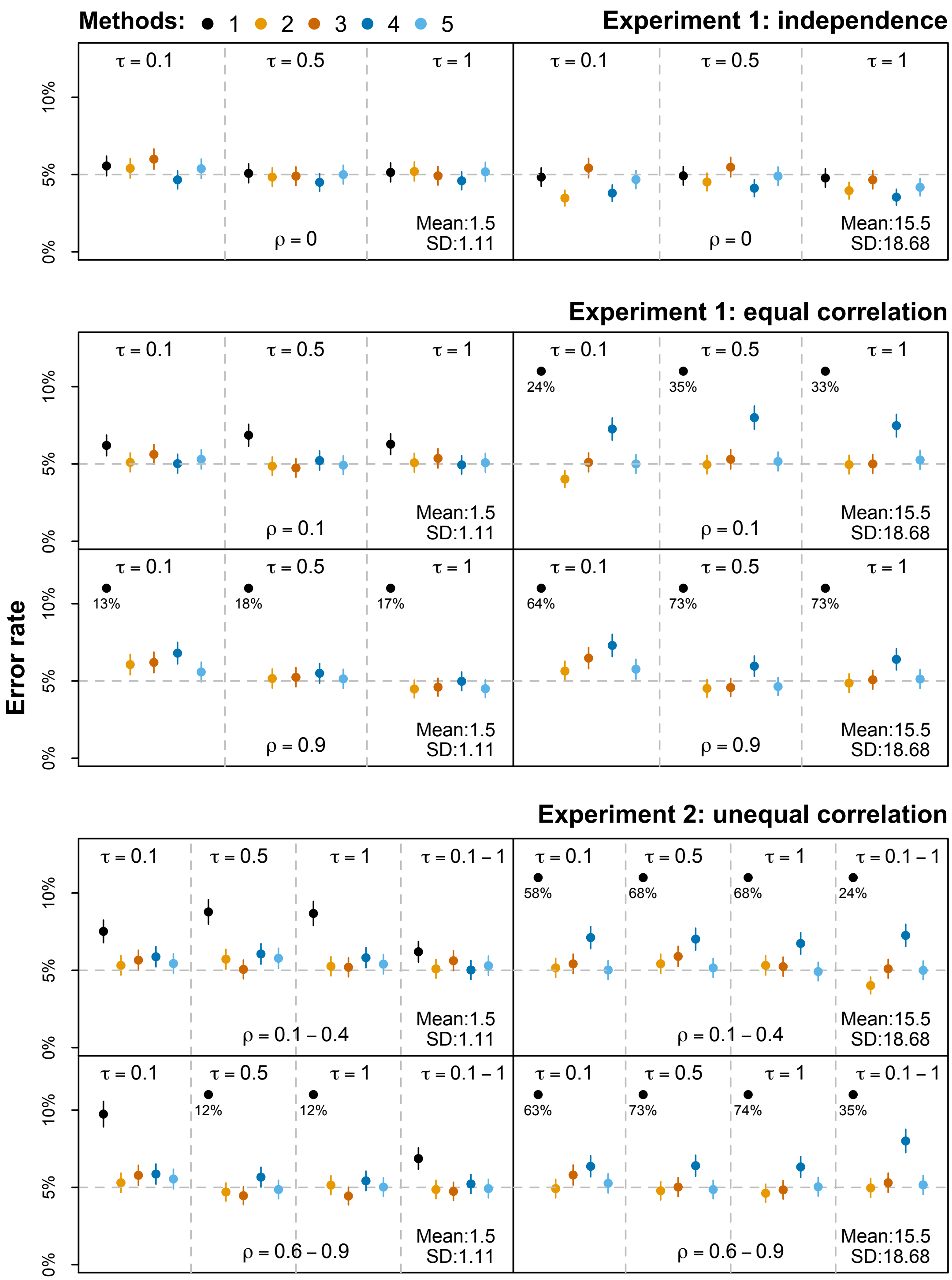
Step 1
Simulate replicated control and treatment data for each study in each paper
Paper 1
Study 1
Study 2
Paper 2
Study 1
Study 2
Study 3
Study 4
Paper 20
Study 1
Step 2
Calculated observed effect size and its variance
Paper 1
Observed effect size 1
Observed effect size 2
Paper 2
Observed effect size 1
Observed effect size 2
Observed effect size 3
Observed effect size 4
Paper 20
Observed effect size 1
Step 3
Fit meta-analysis models using observed effect sizes and variance
Meta-analysis methods
1. Random effect meta-analysis model
2. Two-step method with paper mean
3. Two-step method with one randomly chosen study
4. Including random paper effect
5. Robust variance estimation
Step 4
Evaluate model performance
Model performance metrics
Standard deviation
Error rate
5000 iterations
Experiment 1: equally correlated study
Mean number of studies/paper (n_i)
1.5
5.5
15.5
Correlation (ρ)
0
0.1
0.5
0.9
Among-study variability (τ)
0.1
0.5
1
Experiment 2: unequally correlated study
Mean number of studies/paper (n_i)
1.5
5.5
15.5
Correlation (ρ)
0.1-0.4
0.6-0.9
Among-study variability (τ)
0.1
0.5
1
0.1-1

Figure 2



Methods: 1 2 3 4 5
Experiment 1: independence
Experiment 1: equal correlation
Experiment 2: unequal correlation
Standard deviation
τ = 0.1
τ = 0.5
τ = 1
τ = 0.1 – 1
ρ = 0
ρ = 0.1
ρ = 0.9
ρ = 0.1 – 0.4
ρ = 0.6 – 0.9
Mean:1.5
SD:1.11
Mean:15.5
SD:18.68
0.00
0.15
0.30
0.0
0.2
0.4

Figure 3



Methods: 1 2 3 4 5
Experiment 1: independence
τ = 0.1
τ = 0.5
τ = 1
ρ = 0
Mean:1.5
SD:1.11
Mean:15.5
SD:18.68
Experiment 1: equal correlation
ρ = 0.1
ρ = 0.9
24%
35%
33%
13%
18%
17%
64%
73%
73%
Experiment 2: unequal correlation
τ = 0.1 – 1
ρ = 0.1 – 0.4
ρ = 0.6 – 0.9
58%
68%
68%
24%
12%
12%
63%
73%
74%
35%
Error rate
0%
5%
10%