***Review prepared for Methods for Ecology and Evolution***

**Methods for testing publication bias in ecological and evolutionary meta-analyses**

Shinichi Nakagawa1, Malgorzata Lagisz1, Michael D. Jennions2, Julia Koricheva3, Daniel W.A. Noble2, Timothy H. Parker4, Alfredo Sánchez-Tójar5, Yefeng Yang1, & Rose E. O’Dea1

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

2 XXXXX

3 XXXX

4 XXXXX

5 Evolutionary Biology, Bielefeld University, Bielefeld, 33615, Germany

\* Correspondence: S. Nakagawa

e-mail: [s.nakagawa@unsw.edu.au](mailto:s.nakagawa@unsw.edu.au)

Short title: Publication bias tests for ecology & evolution

# ORCID

Shinichi Nakagawa: 0000-0002-7765-5182

Malgorzata Lagisz: 0000-0002-3993-6127

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

…

Alfredo Sánchez-Tójar: 0000-0002-2886-0649

Rose E. O’Dea: 0000-0001-8177-5075

# Abstract

Aim: 1) the results of a survey, 2) a review of methodologies for publication bias and 3) introducing the most practical method

To do

1. Figures 1-7 = Rose, Shinichi, Losia, Alfredo and Yefeng
2. Supporting information (Appendix S1-S5) = Alfredo and Yefeng and Shinichi
3. Abstract = Shinichi
4. Read several simulation papers to double-check = Shinichi

**KEYWORDS**

Meta-regression, *p*-curve, selective reporting, selection bias, Egger’s regression, trim and fill….

# 1 | INTRODUCTION

Evidence from meta-analyses often lead future research, and sometimes to changes in policy and practice ({Gurevitch, 2018 #1;Nakagawa, 2017 #2}). Therefore, it is essential for meta-analytic evidence to be unbiased. However, the validity of meta-analytic results gets compromised by publication bias ({Marks-Anglin, 2021 #14}). Publication bias occurs when a subset of studies, such as those with statistically non-significant results, are less likely to be disseminated (e.g., the file drawer problem; {Rosenthal, 1979 #49}). In a wider sense, publication bias could encompass many different types of bias relating dissemination of evidence (see {Moller, 2001 #27;Jennions, 2013 #33;Marks-Anglin, 2021 #14}). Here two types are most relevant: 1) outcome reporting bias where selective reporting occurs within published studies ({Marks-Anglin, 2020 #13;, 2021 #14}) and 2) time-lag bias where positive results are published faster than negative results ({Koricheva, 2013 #31;Trkalinos, 2005 #34;Koricheva, 2019 #26}). Regardless of the process of how bias occurs, publication bias leads to an unrepresentative sample of available evidence and thus, meta-analytic results would be distorted.

To tackle this very issue, statisticians have proposed numerous methods for testing publication bias. These tests can be broadly categorised into two: the ones detecting publication bias and the others also assessing the impact of publication bias ({Sutton, 2009 #17}). Both of these categories have been routinely used in meta-analyses in medical and social sciences ({Rothstein, 2005 #3}). However, in a survey of 100 meta-analyses in ecology and evolution, only 49% tested for publication bias, with only 22% conducting both types of the tests ({Nakagawa, 2012 #19}). In another survey, only 31% of 322 ecological meta-analyses reported at least one test of publication bias ({Koricheva, 2014 #18}). The observed low uptake may result from a practical reason that many currently available tests for publication bias are unsuitable for ecological and evolutionary meta-analyses ({Nakagawa, 2012 #19}).

Two features, common to meta-analytic datasets in ecology and evolution, pose problems for publication bias tests: high levels of heterogeneity and non-independence. Ecologists and evolutionary biologists are interested in all types of ecosystems and species on earth so that their meta-analytic data are highly heterogeneous. Also, non-independence is pervasive because many studies produce multiple effect sizes and, if a meta-analytic dataset includes multiple species, then effect sizes will also be correlated due to phylogenetic relatedness ({Noble, 2017 #20}). Importantly, many currently available tests for publication bias fail with high levels of heterogeneity (e.g., {Moreno, 2009 #11;Sterne, 2001 #81;Macaskill, 2001 #84}). Furthermore, Nakagawa and Santos ({, 2012 #19}) noted that at the time, there were no statistical methods to test for publication bias that could explicitly account for non-independent effect sizes. If a publication bias test were to be useful in ecology and evolution, we would need a test that could adequately handle both heterogeneity and non-independence (cf. {Fernandez-Castilla, 2019 #37;Rodgers, 2020 #40}).

Our aim for this article is two-fold. First, we review classic and emerging methods for publication bias by conducting a new survey using 102 meta-analyses in ecology and evolution. Second, we introduce a method that both detects and adjusts for publication bias, while dealing with heterogeneity and non-independence among effect sizes. To make our article widely accessible, we start by revising key statistical concepts in meta-analysis such sampling variance, weights, and heterogeneity (readers who are familiar with these concepts can, therefore, skip the next section).

# 2 | KEY STATSTICAL CONCEPTS

## 2.1 | Sampling variance, standard error, precision and weight

In ecology and evolution, three standardised effect statistics are most common ({Nakagawa, 2012 #19;Koricheva, 2014 #18}). The first effect statistic is the standardised mean difference, SMD (also known as Cohen’s *d* or Hedges’ *g*), whose point estimate and sampling variance can be written as ({Hedges, 1985 #72}):

where the *i*th effect size (SMD) and sampling variance (Var) are a function of the means (), standard deviations (SD of sample) and sample size (*n*) of the two groups (1 & 2). The second effect statistic is the logarithm of response ratio ({Hedges, 1999 #71}; also known as the ratio of means; {Friedrich, 2008 #73}), which can be written as:

where the notations are the same as above (see also {Lajeunesse, 2015 #117;Senior, 2020 #118}). The last one is Fisher’s transformation of the correlation coefficient, *Zr* (unbounded and normally distributed), which can be written as ({Hedges, 1985 #72}):

where *ni*is the *i*th sample size used to obtain the correlation coefficient, *ri*. Incidentally, the variance of the correlation coefficient is: , although a meta-analysis using *r* – bounded at -1 and 1 – is not recommended (see a relevant point in Section 4.2).

Sampling variance is the heart of meta-analysis as this quantity, which is always a function of sample size, indicates (un)certainty for the point estimate of each effect size (see equations above). Confusingly, sampling variance, (sampling) standard error, precision, and weight are often used interchangeably in the meta-analytic literature; for example, a point estimate with high certainty has low standard error and variance, but high precision and weight (Figure 1). Even more confusingly, ‘standard error’ (SE) can be referred to as ‘standard deviation’ (SD). This is not incorrect because standard error is ‘standard deviation of a (parameter) estimate’ – not to be confounded with ‘standard deviation of sample’.

## 2.2 | Heterogeneity

Ecologists and evolutionary biologists have been predominately using a ‘random-effects model’ of meta-analysis rather than a ‘fixed-effect model’ ({Nakagawa, 2012 #19;Koricheva, 2014 #18}). A fixed-effect model assumes homogeneity or that a common overall mean exists among the population of effect sizes. A random-effects model and its extensions, on the other hand, assume each study has its own mean estimate (for an extension, see Section 4.1; {Nakagawa, 2012 #19}; see also Figure 4 in {Nakagawa, 2017 #2}). A random-effects model can be written as:

where *s*i is the between-study (effect-size) effect for the *i*th effect size, normally distributed with a mean of zero and a variance of (which is more commonly referred to as ; note when = 0, this model reduces to a fixed-effect model), and *mi*is the sampling error for the *i*th effect size, distributed with the *i*th sampling variance (note that *i* = 1, 2, …, *Neffect-size*, the number of effect sizes, effect sizes are independent from each other so *Neffect-size* = *Nstudy*, the number of studies). The proportion of against the total variance is often quantified as where is referred to as the ‘typical’ within-study (sampling) variance, which can be considered as a mean value of ({Higgins, 2002 #8}). In ecological and evolutionary meta-analyses, *I*2 is around 90%, on average, meaning only ~10% of variation among effect sizes is due to sampling variance ({Senior, 2016 #74}). Therefore, publication bias tests assuming homogeneity (*I*2 or = 0) are unlikely to be useful for ecology and evolution.

# 3 | PUBLICATION BIAS TESTS

The primary goal of this section is to provide a non-exhaustive but up-to-date overview of publication bias tests, both classic and emerging (cf. {Moller, 2001 #27;Jennions, 2013 #33}; for thorough reviews see {Marks-Anglin, 2020 #13;Marks-Anglin, 2021 #14;Rothstein, 2005 #3;Vevea, 2019 #16}). We summarise different methods referring to the results of our new survey of publication bias tests used in 102 ecology and evolutionary meta-analyses; Figure 2 shows the results of the survey (for the details of survey procedure see Supporting Information, Appendix S1). Further, we describe these tests, especially whether they can handle heterogeneity and can be extended to model non-independence.

Following Sutton ({, 2009 #17}) (see also {Vevea, 2019 #16}), we categories publication bias tests into two: 1) detecting publication bias (e.g., funnel plots, Egger’s regression; Section 3.1), and 2) assessing the impact of publication bias (e.g., Fail-safe *N*, trim-and-fill method, and selection models; Section 3.2). Publication bias, including outcome reporting bias, creates patterns of missing data (known as ‘funnel asymmetry’; see the next section), resulting in an overestimation of the magnitude of an overall effect. Effect sizes, which are small in magnitude and based on small sample sizes (i.e. large sampling variances), are more likely to be missing (unpublished) because these effect sizes are statistically non-significant. Unlike publication and outcome reporting bias, time-lag bias is moderated more by publication year than by statistical significance (or sampling variance) so that it requires different tests (see Section 3.1.3).

## 3.1 | Detecting publication bias

### 3.1.1 | Funnel plots and radial plots

In the absence of publication bias and heterogeneity, plotting effect sizes against a measure of certainty (or uncertainty; see Figure 1) should produce a symmetrical funnel shape around the overall effect, referred to as funnel plots. These graphs are the most popular method in our survey (Figure 2). Funnel plots are also the most preferred graphical tool for publication bias in medical and social sciences ({Marks-Anglin, 2020 #13;Sterne, 2005 #4;Vevea, 2019 #16;Sutton, 2009 #17}), even although many graphical methods were proposed such weighted histograms and normal quantile plots of effect sizes (as in Figure 2; for other graphical methods, see {Rothstein, 2005 #3;Marks-Anglin, 2020 #13}).

The original funnel plot used sample size as the measure of uncertainty ({Light, 1984 #75}; Figure 3a). Yet, more recent recommendations are to use either SE, precision, variance or the inverse of variance (Figure 1; {Sterne, 2005 #4}; but see Section 4.3). For these four quantities, unlike for sample size, we can draw 95% confidence intervals (based on the *y*-axis; 1.96 x SE) that can create a funnel, showing the degree of heterogeneity among effect sizes (if data are homogeneous all dots will be inside the 95% confidence interval region, e.g., Figure 3b & c). This confidence region can also aid visually detecting funnel asymmetry caused by the lack of statstically non-significant effect sizes with high uncertainties (see Figure 3b & c). In a similar vein, the contour-enhanced funnel plot shows different statistical significance regions (around 0) to help detecting asymmetry ({Peters, 2008 #41}; Figure 3d). Lastly, Kossmeier, and colleagues ({, 2020 #77}) have recently proposed a sunset funnel plot, a type of contour-enhanced plot, which adds visual indicators of statistical power (Figure 3e).

A limitation of funnel plots is that funnel asymmetry could show a pattern different from the one expected from publication bias (as in Figure 3f, missing large effect sizes of high uncertainties) since there are several different sources of asymmetry. The most important source is heterogeneity among effect sizes, which can create asymmetries of any kinds (Figure 3c & g); the other sources of asymmetry are data irregularities (e.g., mistakes, frauds, unique observations; cf. {Nakagawa, 2016 #78}), artefacts (see Section 4.3), and chance ({Egger, 1997 #6}). As mentioned above, high heterogeneity is normal in ecological and evolutionary meta-analyses ({Senior, 2016 #74}). Therefore, a standard funnel plot is unlikely informative on publication bias. To account for some of the heterogeneity, several researchers recommended to plot residuals from a meta-regression model (Figure 3g; e.g., {Roberts, 2005 #9}). In practice, however, no meta-regression model would explain all the heterogeneity. Thus, remaining heterogeneity could still generate asymmetry in a residual funnel plot. The funnel plot should, therefore, be seen as a tool to explore small-study effects where effect sizes based on small sample sizes tend to be larger. Small-study effects may indicate publication bias, but not exclusively so ({Sterne, 2005 #4}). Although extensive work exists on funnel plots and heterogeneity, no systematic studies exist on how funnel plots would perform under the scenario where effect sizes are correlated (but see Section 4.1).

Before moving to the next section where we introduce inferential tests to statistically detect funnel asymmetry (or small-study effects), the radial plot proposed by Galbraith ({, 1988 #79}) is worth mentioning, even though our survey found none. An idea of a radial plot is similar to that of a funnel plot. The radial plot shows effect sizes divided by their SEs (essentially, *z* scores) on the *y*-axis and corresponding precisions on the x-axis. The plot also has a line extending from zero on the *x*-axis to an arc showing an overall effect and its 95% confidence interval, while the 95% ‘rectangle’ confidence region is based on lines drawn from ±1.96 values (Figure 3h). The radial plot is useful for visually detecting heterogeneity because data are completely homogeneous when all the data are inside this rectangle (analogous to a funnel shape in funnel plots). These axes of the radial plot (note those of the funnel plot) help us better understand the original inferential test for observed funnel asymmetry, so-called, Egger’s regression ({Egger, 1997 #6}), which is our next topic.

### 3.1.2 | Regression- and correlation-based methods

Egger’s or Egger regression in its original form can be written as:

where *zi*is the *i*th *z* score obtained from dividing an effect size by its SE (*yi/sei*), is the intercept, is the slope for the precision (*prec* or 1/ *se*) and *e* is residuals, normally distributed with a variance of . When (not ) is significantly different from zero, then we statistically detected funnel asymmetry (Figure 4a); the more deviates from zero, the more severe the asymmetry.

Although Egger’s regression checks for asymmetry in a funnel plot, Equation 8 does not have effect sizes as a variable, while a funnel plot does (Figure 3). We intuitively like to draw a regression slope () not the intercept () in a funnel plot, but this is incorrect. It is possible to reformulate Egger’s regression (Equation 8) into a weighted regression, as follows ({Thompson, 1999 #80}):

where *yi* is the *i*th effect size and is the residuals, normally distributed with a variance of , which is sampling variance (*v*) and the multiplicative parameter () estimated in the weighted regression. Notably, Equation 8’s is identical to Equation 9’s (we demonstrate this in Supplementary Information, Appendix S2). Therefore, we can now look at the statistical significance of the slope of SE (*sei*in Equation 9), whose magnitude indicates the severity of asymmetry, and also are able to put a regression line through a funnel plot (Figure 4b).

Given that Equation 9 is very similar to a meta-regression, later versions of Egger’s regression variants have taken the same form as a meta-regression ({Moreno, 2009 #11}), for example:

which is the same as Equations 7 (the random-effects model) plus the slope of SE () (note that different variants have precision, variance o the inverse of variance instead of SE; {Moreno, 2009 #11}).

According to simulation studies ({Moreno, 2009 #11;Sterne, 2001 #81;Macaskill, 2001 #84}), Egger’s regression and its variants suffer from low power and poor performance when there are fewer than 20 effect sizes, or when the effect is large. However, meta-analyses in ecology and evolution often include over 20 effect sizes and our overall effect is usually small ({Senior, 2016 #74}). Therefore, the regression-based method for publication bias is likely to be of use, at least to detect small-study effects. Furthermore, in this meta-regression formulation, two things are possible: 1) adding more moderators to account for heterogeneity, and 2) using multilevel meta-analysis to account for non-independence among effect sizes. We expand these possibilities in Section 4.

According to our survey (Figure 2), the correlation-based method is relatively popular in ecology and evolution (9.4%) like the regression-based method (11.7%). These two classes of methods are closely related because they both statistically examine a relationship between effect sizes and corresponding uncertainties (e.g. sampling variance). All the correlation methods are based on a version of the rank correlation test first proposed by Begg and Manzumdar ({, 1994 #43}). This method essentially calculates a Kendall’s rank correlation between effect sizes and their sampling variance; a statistically significant correlation can indicate a small-study effect. Thus, it is a very simple to implement, but it seems that the rank correlation is not as powerful as Egger’s regression under many circumstances ({Macaskill, 2001 #84}). Given the rank correlation is a non-parametric method, there is, theoretically, no assumption of the independence of data. Yet, a recent simulation shows that a rank correlation method performs poorly when effect sizes are correlated ({Fernandez-Castilla, 2019 #37}).

### 3.1.3 | Time-lag bias tests

Time-lag bias occurs when larger or significant effects are published more quickly than smaller or non-significant effects, but this bias is often used to mean that overall effects decline over time (i.e., a decline effect ({ Koricheva, 2019 #26}). According to our survey (Figure 2), fewer than 5% of meta-analyses in ecology and evolution tested for this type of publication bias. This is concerning, given that time-lag bias is likely to be prevalent in ecology and evolution ({Jennions, 2002 #113;Sanchez-Tojar, 2018 #119}). There are two recommended ways to investigate time-lag bias (or a decline effect): 1) using a cumulative meta-analysis, and 2) using a regression-based method (see {Koricheva, 2013 #31;Trkalinos, 2005 #34;Koricheva, 2019 #26}). We do not recommend the correlation-based method, for example, quantifying a rank correlation between effect size and publication year (e.g., {Barto, 2012 #85}), because this approach does not account for different precisions of effect sizes.

Cumulative meta-analysis is where a meta-analytic model (e.g., random-effects model) is applied to a set of effect sizes, which is increased by one effect size at a time iteratively (starting from the oldest effect size). Then, the results are displayed as a forest plot (see Figure 4c). One can easily see when significance of overall effect sizes changes over time. Again, cumulative meta-analysis is only possible when we employ a simple model (fixed-effect and random-effects models). When multiple effect sizes are obtained from each study, adding one study (one or more effect sizes) rather than one effect size is more practical. Yet, a complex but more appropriate meta-analytic model (see Section 4.1) is unlikely to run with a few studies (effect sizes).

The second method is based on regression and is easy to fit, for example (cf. Equation 10):

where *yeari* is the publication year for the *i*th study (effect size). As with Equation 8, this method can accommodate other moderators (i.e. potential confounding variables) and also can be extendable to model non-independent effect sizes (see Section 4.2).

## 3.2 | Assessing the impact of publication bias

### 3.2.1 | Fail-safe *N*

We now move to publication bias methods that can assess the impact of publication bias rather than merely detecting publication bias. Fail-safe *N* (also known as the ‘file-drawer number’) represents the number of negative unpublished results needed to make a statistically significant overall effect non-significant (e.g., {Rosenthal, 1979 #49;Rosenberg, 2005 #46}) or negligible (e.g., {Owrin, 1983 #59}). The original fail-safe approach by Rosenthal ({, 1979 #49}) is the oldest publication bias method and probably the simplest:

where *zi* is the *i*th z value (*yi/sei*) as in Equation 7 and 1.645 is the z value for α = 0.05 (the one-tailed test). The method by Owrin ({, 1983 #59}) relies on the effect size rather than statistical significance; one version of this method can be written as:

where is the overall mean (i.e. an estimate from a fixed-effect model) and *yn* is the effect size value that is considered to be small or negligible. Although Rosenthal’s and Orwin’s fail-safe numbers ignore sample sizes (uncertainty) of effect sizes in the dataset, the method proposed by Rosenberg ({, 2005 #46}) explicitly includes such information. An equation that assumes a fixed-effect model can be written as:

where *wi* is the inverse of sampling variance (1/*vi*; note that *wi* can be modified for a random-effects model) and denotes the *t* value with the α level of 0.05 with the number of studies (effect sizes) as the degrees of freedom, DF (for the use of a different DF, see {Rosenberg, 2005 #46}).

Although fail-safe approaches are the most popular after the funnel plot in our survey (14.1%), Becker ({, 2005 #5}) has called for abandoning all the fail-safe approaches, given the other methods for publication bias available. She has argued that the fail-safe *N* is difficult to interpret (e.g., no criterion on what constitutes a small or large *N)*, and also that depending on methods, a variety of fail-safe numbers can be obtained for the same data set. For example, the *R* package *metafor* implements the three methods above ({Viechtbauer, 2010 #87}); its example dataset shows *NRosenthal* = 598, *NOwrin* = 13, and *NRosenberg* = 370 (see Supporting Information, Appendix S3). Unfortunately, none of the proposed methods adequately control for heterogeneity (e.g., by incorporating moderators), nor non-independence among effect sizes. Furthermore, all methods of fail-safe *N* are not inferential methods.

### 3.2.2 | Trim-and-fill tests

The trim-and-fill test provides a non-parametric method that statistically both detects and corrects for funnel asymmetry and that can visualize potentially missing data ({Duval, 2000 #61;Duval, 2000 #62}). A recent survey showed that the number of studies using the trim-and-fill method is increasing every year (in 2018, over 2000 meta-analyses used this method; {Shi, 2019 #88}), and this method is not uncommon in ecology and evolution (7.5% of the meta-analyses in our survey). In short, this method uses an iterative process to determine how many effect sizes are missing (say, *Nmissing*) from a funnel, using an initial overall estimate and one of three estimators (*R*0, *L*0, & *Q*0; see an accessible account, see ). Then, it ‘trims’ off *Nmissing* effect sizes to suppress funnel asymmetry, and estimates a new overall mean to see whether it can trim more effect sizes till the value *Nmissing* stabilizes. Subsequently, *Nmissing* effect sizes are ‘filled’ as mirror images (Figure 4e & f). Finally, an overall effect is re-estimated with filled values. We note that Duval ({, 2005 #60}) has recommended the use of *R*0 and *L*0, and that the estimator *R*0 can provide a significance test for whether the number of missing values is zero or not.

The problem of the trim-and-fill test is that the original method assumes homogeneity (i.e. a true mean for all effect sizes). In practice, the trim-and-fill method seems to tolerate some heterogeneity, but performs worse as heterogeneity increases ({Moreno, 2009 #11;Peters, 2007 #116}). Although trim-and-fill tests have been extended for meta-regressions ({Weinhandl, 2012 #86}), the implementation of this extension is limited to one moderator. Further, recent simulation work by Rogers and Pustejovksy ({, 2020 #40}) shows that ignoring non-independence and fitting a trim-and-fill method (using *R*0) increases Type I error rates, especially when a large effect exists.

### 3.2.3 | *P*-value-based methods and selection models

Ecologists and evolutionary biologists have hardly used the methods based on *p*-values and selection methods (*p*-value-based: 1.4%, selection models: 0%, Figure 2), although both types of methods can provide adjusted overall means. The same researchers who popularized the terms ‘researcher degrees of freedom’ ({Simmons, 2011 #23}) and ‘*p*-hacking’, have introduced the *p*-curve method ({Simonsohn, 2014 #21}). The *p*-curve method relies on the distribution of *p* values of effect sizes in a dataset (Figure 6a,b). The *p*-uniform method is a similar method, which also exploits the distribution of *p* values ({van Assen, 2015 #65}). Interestingly, McShane et al. ({, 2016 #64}) has pointed out that both *p*-curve and *p*-uniform tests are versions of a selection model first suggested by Hedges ({, 1984 #67}); all of these methods, unfortunately, do not perform well with heterogeneity as they assume one true effect, i.e. homogeneity (see also, {van Aert, 2016 #63}). Clearly, in ecology and evolution where high levels of heterogeneity are commonplace (Senior et al. 2016), these methods may be of limited use, especially compared to more advanced selection models.

Selection model-based methods represent the most sophisticated and complex class of publication bias methods (reviewed in {Marks-Anglin, 2020 #13;Rothstein, 2005 #3;Vevea, 2019 #16}). Although few ecologists and evolutionary biologists, if any, seem to use selection models, there are probably as many selection models as all other methods combined ({Marks-Anglin, 2020 #13}). One property common to all selection models is that they model how effect sizes are missing (or selected to be published), based on, for example, *p* values, effect sizes and/or sampling variance (Figure 5d-f). Importantly, selection models can tolerate and model heterogeneity. Indeed, one recent model by Citkowicz and Vevea ({, 2017 #91}) can statistically test for publication bias, incorporate moderators, tolerate substantial heterogeneity, provide an adjusted overall effect, and even correct estimates for small sample sizes. Yet, no selection methods are implemented for dependent effect sizes, as far as we are aware (cf. {Kirkham, 2018 #92}).

# 4 | METHODS FOR DEPENDENT EFFECT SIZES

In this section, we first define a multilevel model that explicitly incorporates non-independence among effect sizes. Then, we consider how to best visualize such datasets as a funnel plot. Then, we build upon a regression-based method introduced above to propose a new publication bias method. This new method can both detect and correct for funnel asymmetry or small-study effects while modelling heterogeneity and complex non-independence involving correlation and variance-covariance matrices.

## 4.1 | A multilevel meta-analysis and forest plots

The simplest multilevel meta-analytic model can be written as ({Nakagawa, 2012 #19}):

where is the overall estimate (or meta-analytic mean), and *sj* is the between-study effect for the *j*th study, normally distributed with the variance of , and *ui* is the between-effect-size effect, or within-study effect, for the *i*th effect size, distributed with a mean of zero and the variance of , and *mi* is as in Equation 7 (but note that *j* = 1, 2, …, *Nstudy*, the number of studies, and *i* = 1, 2, …, *Neffect-size*, the number of effect sizes; *Neffect-size* > *Nstudy*). Equation 15 explicitly models multiple effect sizes per study. Also, in Equation 7, the term is the only source of heterogeneity, while in Equation 15, both and are each contributing to effect size heterogeneity.

Now we can easily extend this to a meta-regression model. For example, a meta-regression with two moderators can be written as:

where is the slope for *x*1, a study-level moderator (characteristics of different studies, *j*;e.g.,experimental *vs.* observational) and is the slope for *x*2, an effect-size-level moderator (characteristics of effect sizes, *i*; different measurements or sexes). We have mentioned that we can draw a funnel plot with residuals rather than the observed effect sizes (Figure 5a). A complication is that, given Equation 15, we can extract at least 3 different residuals, which are:

where *residm* represents marginal residuals (subtracting only fixed effects from the observations; Figure 5b), whereas *residc*1 and *residc*2 are conditional residuals (Figure 5c & d; {Nobre, 2007 #95}). As shown in Figure 5, marginal residuals still show the patterns due to originating from the same studies (i.e. sample sizes are the same or similar). Contrastingly, conditional residuals no longer show these patterns as we took out a clustering factor (*sj*). Therefore, funnel plots with conditional residuals seem like a useful exploratory tool for publication bias when effect sizes are correlated.

Notably, the residuals from Equation 17 are analogous to marginal residuals from a random-effects model, whereas those from Equation 18 are analogous to marginal residuals from a fixed-effect model. Therefore, Nakagawa and Santos ({, 2012 #19}) suggest using the conditional residuals along with corresponding sampling variance or standard error (*vi* or *sei*) in publication bias tests (e.g., the original Egger’s regression and trim-and-fill tests). However, there are at least two assumptions in this approach. First, all such residual analyses assume that sampling SE (*sei*) does not covary with moderators in meta-regression (e.g., *x*1 and *x*2 in Equation; see {Freckleton, 2002 #97}). Second, sampling SE is the same as the SE of the residuals; they are not, although they are strongly correlated (see {Doleman, 2020 #99}). Furthermore, in the presence of non-independent data, Equation 15’s sampling variances are often correlated; that is, where **M** is a variance-covariance matrix. For example, when *Neffect-size* = 3 and the first two effect sizes’ sampling variance are correlated, then we can write **M** as:

where is the correlation between the sampling effects of the first two effect sizes ( is the covariance). Whenever sampling (error) effects are correlated, neither *residc*1 nor *residc*2 are independent. Then, none of publication bias tests, reviewed in Section 3, should be used.

## 4.2 | Multilevel meta-regression and Egger’s regression

As an alternative to using residual analysis, we can directly model sampling SE in Equation 15 (cf. Equation 10; {Fernandez-Castilla, 2019 #37;Rodgers, 2020 #40}):

By examining Equation 21, we may realise that represents a conditional estimate of an overall effect when SE is 0, which means, theoretically, there is no uncertainty (Figure 5e). Then, does provide an adjusted estimate of an overall effect, when is statistically significant? This question has been examined by Stanley and Doucouliagos ({, 2012 #101;, 2014 #12}). They have shown that is downwardly biased when a true positive effect exists (Figure 5e). If the slope of SE () is statistically significant, fitting sampling variance instead of SE, as follows:

This is equivalent to fitting , which is a quadratic term. Stanley and Doucouliagos ({, 2012 #101;, 2014 #12}) has shown that in Equation 22 is still downwardly biased, but much less so, although Equation 21 is more powerful (i.e. an adjustment is biased toward a null effect) when there is a positive effect (Figure 5f). While this two-step approach may seem simplistic (see also {Stanley, 2017 #10}), it provides an easy-to-implement publication bias test which explicitly models non-independent data.

Further, this regression approach can be used to test time-lag bias (a decline effect) by modelling the publication year (*yearj*):

When heterogeneity exists, it is best to combine Equation 21 and 23 with moderators, for example:

where is the slope for the *k*th moderator (*k* = 3, 4,…, *Nmod*; the number of moderators), the other parameters are as above. However, simulation studies have shown Egger’s regression variants with sampling standard error as a moderator (e.g., Equations 10 & 21) perform poorly, even when adequately powered ({Macaskill, 2001 #84;Deeks, 2005 #100}). This is especially true under two scenarios: 1) when there is a relationship between effect size and sampling SE, and 2) when SE is estimated poorly.

## 4.3 | Multilevel meta-regression using sample size

To understand how a correlation between effect size and SE can come about, and when SE can be estimated poorly, we now go back to comparing sampling variance among the three commonly used effect sizes (Equations 2, 4 & 6). We may realize that SMD’s variance has the square of the point estimate (i.e. SMD; Equations 2). This can lead to a correlation between SMDs and sampling SE, resulting in ‘artefactual’ funnel asymmetry (Section 3.2). Further, we also notice that in Equation 4 (i.e. lnRR’s variance), when sample sizes (*n*1 and *n*2) are small, (sample mean) and especially SD (sample standard deviation) will be poorly estimated, resulting in an unreliable estimate of sampling variance (this is the case for Equation 2 too). These issues do not affect the sampling variance of *Zr*, which is a function only of sample size (*n*; Equation 6). Therefore, the sample size (*n*1 + *n*2) has been suggested as a moderator instead of SE (e.g., Equation 19) when we use effect size statistics such as SMD and lnRR (also correlation, *r*; see Section 2.1); this approach is known as the funnel plot test ({Macaskill, 2001 #84}). Simulations suggest using the sample size as a moderator outperforms SE in the cases of both independent and dependent effect sizes ({Macaskill, 2001 #84;Deeks, 2005 #100}) and dependent effect sizes ({Fernandez-Castilla, 2019 #37}).

Instead of the sample size (*n*1 + *n*2), however, we propose using the ‘effective sample size’ () that accounts for unbalanced sampling, for SMD and lnRR, and effective sample size can be given by ({Bakbergenuly, 2020 #103;, 2020 #105}; also see; {Deeks, 2005 #100;Bakbergenuly, 2020 #104}):

When *n* = *n*1 = *n*2, the formula reduces to 2*n*. Indeed, the inverse of is a part of sampling variance in both SMD and lnRR (Equations 4 & 6):

where the middle part of the formula corresponds to Equation 2 when setting SMD = 0, while the righthand side corresponds to Equation 4 when setting CV (SD/) = 1.

Taken together, we can rewrite Equations 21 and 22, respectively, as ({Deeks, 2005 #100}):

where is a replacement of *sei* in Equation 21, and is a replacement of *vi* in Equation 22 (note that, at the intercept, is infinitely large). We recommend using Equation 27 (more power) to check the statistically significance of funnel asymmetry (small-study effects). Then, Equation 28 (less biased) to obtain an overall estimate adjusted for publication bias, when effect sizes are SMD or lnRR (with *Zr*, we should use Equations 21 and 22). This adjusted estimate should not be taken as a true estimate, however. We should treat it as a possible overall estimate as a part of sensitivity analysis in which we run alternative statistical models to test the robustness of results from the original analysis ({Noble, 2017 #20}).

In practice, multilevel meta-analytic models are often more complex. For example, Nakagawa and Santos ({, 2012 #19}) have proposed a phylogenetic multilevel model with a phylogenetic random factor and a non-phylogenetic random factor as a theoretically sound model when effect sizes are obtained from different species (see also {Hadfield, 2010 #114}). The major benefit of our proposed meta-regression approach to publication bias is that we can easily extend these models to incorporate other sources of heterogeneity. An example of a meta-regression model testing publication bias and time-lag bias, but that also includes phylogenetic and non-phylogenetic random effects, can be written as:

where *ah* is the phylogenetic effect for the *h*th species, considered multivariate normally distributed with a covariance of (**A** is a correlation matrix derived from a phylogeny), and *qh* is the between-effect-size effect, or within-study effect, for the *h*th species, distributed with the variance of (*h* = 1, 2, …, *Nspecies*, the number of species; ), and the other notations are the same as above. Relevantly, when using SMD or lnRR, we may be better off using and residuals for drawing funnel plots (see Section 4.1; {Doleman, 2020 #99}). We use two datasets and the three effect sizes to illustrate the proposed methods in Supporting Information (Appendix S4).

## 4.4 | Alternative approaches: averaging or sampling

Many of the methods we introduced in Section 3 are still useful, even in the presence of non-independent data, if we aggregate effect sizes per study or sample one effect size per study. Relevantly, overall means will generally not be biased using aggregated or sampled effect sizes ({Song, 2020 #94}). Also, Rodgers and Pustejovsky ({, 2020 #40}) showed that all Egger’s regression (similar to Equation 10), the trim-and-fill test (using *R*0 estimator) and the three-parameter selection model (as in {Vevea, 1995 #110}) all performed well having the appropriate level of Type 1 error, although the three-parameter selection model was noticeably more powerful than the others. However, this is not a general solution when we have a phylogenetic signal ( > 0; Equation 29). In such a case, averaging or sampling per species will not eliminate non-independence as effect sizes are still correlated via phylogeny (i.e. **A** in Equation 29). Furthermore, even when there is no phylogenetic signal ( = 0), or we do not have the species-level structure in a dataset, these alternative approaches could be problematic. For example, if we average effect sizes, we will lose all effect-size-level moderators (e.g., one cannot average categorical moderators such as types of measurements, evaluation methods or sex). Similarly, although the sampling approach would allow controlling for effect-size-level moderators, this approach also reduces the information content of the dataset. Incidentally, when sampling variances are correlated (i.e. **M** as in Equation 29), ‘average’ sampling variance needs to be calculated by using the following formula (not by simple weighted averaging as for the mean; {Borenstein, 2009 #115}):

where *yg* is the *g*th effect size in a study (*g* = 1, …, *Nwithin* and *l* = 1, …, *Nwithin* where *Nwithin* is the number of effect size within a paper or a species to be combined), and are the sampling error variances for *yg* and *yl*, and *rgl* is the correlation between the sampling errors of *yg* and *yl*. Despite these limitations, under some circumstances, averaging and sampling could be useful (examples and implementations for the trim-and-fill test and a selection model in Supporting Information, Appendix S5).

# 5 | CONCLUSIONS

Given high levels of heterogeneity and non-independence in ecological and evolutionary meta-analytic datasets, our choice of suitable tests for publication bias is limited. Indeed, our proposed regression-based method appears to be the only practical method that could fulfil statistical assumptions under most circumstances. Nonetheless, we have described the main methods for testing publication bias alongside our recommendations, as summarised in Figure 7. Although using averaging or sampling are not a universal solution, they may be useful in supplementing our regression-based method. This is because all publication bias tests should be seen as a part of sensitivity analysis (see above; {Noble, 2017 #20}), meaning that we should run more than one publication bias test. We have noticed that only two simulation studies exist explicitly investigating the performance of publication bias tests with non-independent data, both of which supported similar models to the regression-based method we proposed here ({Fernandez-Castilla, 2019 #37;Rodgers, 2020 #40}). Clearly, we need more methodological and simulation-based work in the future.

Finally, we repeat that the results of publication bias tests should be always cautiously interpreted because no methods could ever verify the actual number of missing effect sizes. Indeed, a recent study compared the results of 15 meta-analyses and pre-registered replication projects on the same topics ({Kvarven, 2020 #112}). The overall effects from the replication projects are smaller than those of the meta-analyses. More importantly, the replication projects’ estimates are, in general, also smaller than adjusted effects from the trim-and-fill method, the three-parameter selection model and the 2-step regression model (the method by {Stanley, 2012 #101;, 2014 #12}). Nonetheless, as long as we acknowledge the limitations and assumptions of these methods, publication bias tests are an essential part of meta-analysis. All future meta-analyses in ecology and evolution should test for publication bias and related biases.

# ACKNOWLEDGEMENTS

SN, REO and ML were supported by an ARC (Australian Research Council) Discovery grant (DP200100367). AST was funded by the German Research Foundation (DFG) as part of the SFB TRR 212 (NC3) – Project no. 316099922 and 396782608. Thanks Wolfgang Viechtbauer.

# AUTHORS’ CONTRIBUTIONS

# DATA AVAILABILITY

We have relevant data and code available at the GitHub repository (https://github.com/itchyshin/publication\_bias).

# REFERENCES

# FIGURE LEGENDS

**FIGURE 1.** SE, V, Prec and W (Losia)

**FIGURE 2.** Results of survey (Rose)

**FIGURE 3.** Funnel plots x 8 (Shinichi) = one could simulation data for this – use the random-effects model with low hetero

**FIGURE 4.** 6 panels**:** A: radial, B: Egger, C: cumulative, D, bubble, E: trim-and-fill 1 and F: trim-and-fill 2 (Shinichi) – again use simulated data

**FIGURE 5.** A-B – p values (p hacking) and C, D, E & F – selection models (Shinichi)

**FIGURE 6.** A-D: funnel raw, residual 1, 2, 3 (we need correlated data here) and E: SE and F: Vi (Shinichi) – it may be good to get a real data but you can also simulate using the mulilevel model

**FIGURE 7.** Schematics – key stuff, heterogeneity, power, non-independent – recommended or not (Rose + Shinichi)

## FIGURE 1

**Diagram

Description automatically generated**

\*Weight here is weight for the fixed-effect model

## FIGURE 2

****

## FIGURE 3

**Chart

Description automatically generated with medium confidence**

## FIGURE 4

**Chart, scatter chart

Description automatically generated**

## FIGURE 5

P curves etc etc….Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Selection models

**Chart, histogram

Description automatically generated**

Chart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generated

## FIGURE 6

**A picture containing text, whiteboard

Description automatically generated**

## FIGURE 7

A schematic or table?

I think something like this but better as a figure?

A picture containing diagram

Description automatically generated