REVIEW ARTICLE

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

Shinichi Nakagawa[1] | Malgorzata Lagisz[1] | Michael D. Jennions[2] |
Julia Koricheva[3] | Daniel W. A. Noble[2] | Timothy H. Parker[4] |
Alfredo Sánchez-Tójar[5] | Yefeng Yang[1] | Rose E. O'Dea[1]

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

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

[3]Department of Biological Sciences, Royal Holloway University of London, Egham, UK

[4]Department of Biology, Whitman College, Walla Walla, WA, USA

[5]Department of Evolutionary Biology, Bielefeld University, Bielefeld, Germany

Correspondence
Shinichi Nakagawa
Email: s.nakagawa@unsw.edu.au

Alfredo Sánchez-Tójar
Email: alfredo.sanchez-tojar@uni-bielefeld.de

Funding information
Australian Research Council, Grant/Award Number: DP200100367; Deutsche Forschungsgemeinschaft, Grant/Award Number: 316099922 and 396782608

Handling Editor: Robert Freckleton

Abstract

1. Publication bias threatens the validity of quantitative evidence from meta-analyses as it results in some findings being overrepresented in meta-analytic datasets because they are published more frequently or sooner (e.g. 'positive' results). Unfortunately, methods to test for the presence of publication bias, or assess its impact on meta-analytic results, are unsuitable for datasets with high heterogeneity and non-independence, as is common in ecology and evolutionary biology.

2. We first review both classic and emerging publication bias tests (e.g. funnel plots, Egger's regression, cumulative meta-analysis, fail-safe N, trim-and-fill tests, p-curve and selection models), showing that some tests cannot handle heterogeneity, and, more importantly, none of the methods can deal with non-independence. For each method, we estimate current usage in ecology and evolutionary biology, based on a representative sample of 102 meta-analyses published in the last 10 years.

3. Then, we propose a new method using multilevel meta-regression, which can model both heterogeneity and non-independence, by extending existing regression-based methods (i.e. Egger's regression). We describe how our multilevel meta-regression can test not only publication bias, but also time-lag bias, and how it can be supplemented by residual funnel plots.

4. Overall, we provide ecologists and evolutionary biologists with practical recommendations on which methods are appropriate to employ given independent and non-independent effect sizes. No method is ideal, and more simulation studies are required to understand how Type 1 and Type 2 error rates are impacted by complex data structures. Still, the limitations of these methods do not justify ignoring publication bias in ecological and evolutionary meta-analyses.

1 | INTRODUCTION

Evidence from meta-analyses often drives future research, and sometimes leads to changes in policy and practice (Gurevitch et al., 2018; Nakagawa et al., 2017). Therefore, it is essential for meta-analytic evidence to minimize bias. However, the validity of meta-analytic results can be compromised by publication bias (Marks-Anglin et al., 2021). Publication bias occurs when a subset of research findings, such as statistically non-significant results, are less likely to be published (e.g. the file drawer problem; Rosenthal, 1979). In a wider sense, publication bias could encompass many different types of bias relating to dissemination of evidence (see Jennions et al., 2013; Marks-Anglin et al., 2021; Moller & Jennions, 2001). In this article, the following two types are most relevant: (a) outcome reporting bias, where selective reporting occurs within published studies (Marks-Anglin & Chen, 2020a, 2020b) and (b) time-lag bias, where positive results are published earlier than negative results (Koricheva et al., 2013; Koricheva & Kulinskaya, 2019; Trkalinos & Ioannidis, 2005). Regardless of underlying causes of publication bias, if published findings are unrepresentative of all available evidence, meta-analytic results can be distorted.

Numerous methods have been developed to test for publication bias. These tests can be broadly categorized into two types: those that detect publication bias, and those that also assess the impact of publication bias on the results of the meta-analysis (Sutton, 2009). Both of these types of tests have been routinely used in meta-analyses in the medical and social sciences (Rothstein et al., 2005). However, in a survey of 100 meta-analyses in ecology and evolution, only 49% tested for publication bias, with just 22% conducting both types of tests (Nakagawa & Santos, 2012). In another survey, only 31% of 322 ecological meta-analyses reported at least one test of publication bias (Koricheva & Gurevitch, 2014). Low uptake might reflect that many currently available tests for publication bias are unsuitable for ecological and evolutionary meta-analyses (Nakagawa & Santos, 2012), although the main cause probably is lack of widespread awareness of the importance of publication bias tests in meta-analysis in ecology and evolution (Koricheva & Gurevitch, 2014).

Two features common to meta-analytic datasets in ecology and evolution pose problems for publication bias tests: high levels of heterogeneity and non-independence. Importantly, many currently available tests for publication bias fail when there are high levels of heterogeneity (e.g. Macaskill et al., 2001; Moreno et al., 2009; Sterne et al., 2001). Furthermore, Nakagawa and Santos (2012) noted that, at the time, there were no statistical methods to test for publication bias that could explicitly account for non-independent effect sizes. Highly heterogeneous data are common in ecology and evolutionary biology, as research questions often span many types of ecosystems and species. Non-independence is pervasive because many studies produce multiple effect sizes and, if a meta-analytic dataset includes multiple species, then effect sizes might also be correlated due to phylogenetic relatedness (Noble et al., 2017). Therefore, for a publication bias test to be useful in ecology and evolution, it would need to adequately handle both heterogeneity and non-independence (cf. Fernandez-Castilla et al., 2021; Rodgers & Pustejovsky, 2021).

Our aim for this article is twofold. First, we review classic and emerging methods for detecting and adjusting for publication bias, and assess their usage by conducting a new survey of 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 revisiting key statistical concepts in meta-analysis such as effect sizes, sampling variance, weights and heterogeneity (readers who are familiar with these concepts can, therefore, skip the following section).

2 | KEY STATSTICAL CONCEPTS

2.1 | Common effect size statistics

Three types of standardized effect size statistics are most commonly used in meta-analyses in ecology and evolutionary biology (Koricheva & Gurevitch, 2014; Nakagawa & Santos, 2012). The first effect size statistic is the standardized mean difference, SMD (Cohen's *d* or Hedges' *g* are well-known estimators of SMD), whose point estimate and sampling variance can be written as (Cohen, 1988; Hedges & Olkin, 1985):

$$\text{SMD}_i = \frac{\overline{X}_{2i} - \overline{X}_{1i}}{\sqrt{\frac{(n_{1i} - 1)SD_{1i}^2 + (n_{2i} - 1)SD_{2i}^2}{n_{1i} + n_{2i} - 2}}}, \tag{1}$$

$$\text{Var}\left(\text{SMD}_i\right) = \frac{n_{1i} + n_{2i}}{n_{1i}n_{2i}} + \frac{\text{SMD}_i^2}{2\left(n_{1i} + n_{2i}\right)}, \tag{2}$$

where the *i*th effect size (SMD) and sampling variance (Var) are a function of the means ($\overline{X}$), standard deviations (*SD* of sample) and sample size (*n*) of the two groups (1 and 2); Equations 1 and 2 could be modified to add a small sample-size correction factor (see Borenstein et al., 2009). Second, the logarithm of response ratio, lnRR (Hedges et al., 1999; also known as the ratio of means, or RoM; Friedrich et al., 2008) can be written as:

$$\text{lnRR}_i = \ln\left(\frac{\overline{X}_{2i}}{\overline{X}_{1i}}\right), \tag{3}$$

$$\text{Var}\left(\ln RR_i\right) = \frac{SD_{1i}^2}{n_{1i}\overline{X}_{1i}^2} + \frac{SD_{2i}^2}{n_{2i}\overline{X}_{2i}^2}, \qquad (4)$$

where the notations are the same as above (see also Lajeunesse, 2015; Senior et al., 2020). Finally, Fisher's transformation of the Pearson's correlation coefficient, Zr (unbounded and normally distributed), can be written as (Hedges & Olkin, 1985):

$$Zr_i = \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right), \qquad (5)$$

$$\text{Var}\left(Zr_i\right) = \frac{1}{n_i - 3}, \qquad (6)$$

where n_i is the ith sample size used to obtain the correlation coefficient, r_i. Incidentally, the variance of the correlation coefficient is: $\text{Var}(r_i) = (1-r_i^2)^2/(n_i-1)$, although a meta-analysis using r, which is bounded at −1 and 1, is generally not recommended (see a relevant point in Section 4.2).

These frequently used equations show that sampling variance is at the heart of meta-analysis. As one can see, sampling variance is always a function of sample size, indicating (un)certainty around the point estimate of each effect size (see equations above). It is important to note that sampling variance, (sampling) standard error, precision and weight are often used interchangeably in the meta-analytic literature to refer to (un)certainty of a point estimate (Figure 1). For example, a point estimate with high certainty has low 'standard error' and 'variance', but high 'precision' and 'weight'.

2.2 | Heterogeneity

Ecologists and evolutionary biologists predominately use a 'random-effects model' meta-analysis rather than a 'fixed-effect model' (Koricheva & Gurevitch, 2014; Nakagawa & Santos, 2012). A fixed-effect model assumes that a common overall mean exists among the population of effect sizes (i.e. homogeneity). A random-effects model and its extensions, on the other hand, assume that each study has its own mean estimate (for an extension, see Section 4.1; Nakagawa & Santos, 2012; see also figure 4 in Nakagawa et al., 2017). A random-effects model can be written as:

$$y_i = \beta_0 + s_i + m_i, \qquad (7)$$

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

where β_0 is the overall estimate (or meta-analytic mean), s_i is the between-study (effect size) effect for the ith effect size, normally distributed with a mean of zero and a variance of σ_s^2 (which is more commonly referred to as τ^2; note when $\sigma_s^2 = 0$, this model reduces to a fixed-effect model), and m_i is the sampling error for the ith effect size, distributed with the ith sampling variance (v_i; note that $i = 1, 2, ..., N_{\text{effect}}$
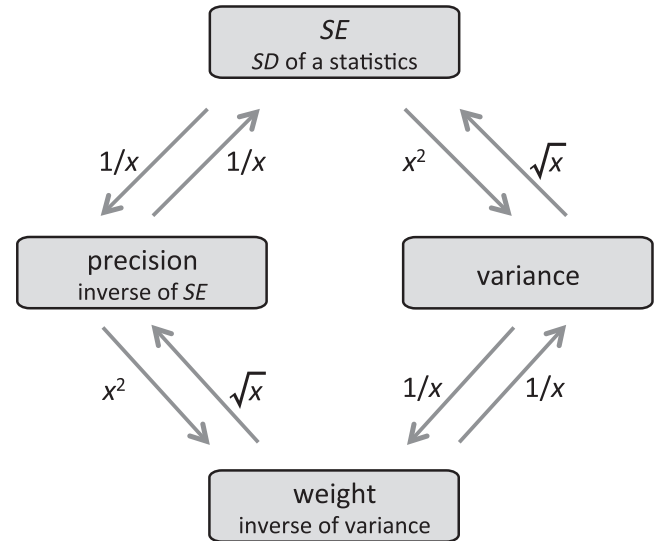

FIGURE 1 A schematic showing the relationship among common terminology in the meta-analytic literature: standard error (*SE*), sampling variance, precision (the inverse of *SE*) and weight (the inverse of variance). Note that the inverse of variance is the weight for a fixed-effect model (the weight for a random-effect model is the inverse of the sum of sampling variance and between-study variance). In the statistical literature, the inverse of variance is also referred to as precision. Importantly, 'standard error' (*SE*) can be referred to as 'standard deviation' (*SD*), which is not incorrect because standard error is 'standard deviation of a statistic'—not to be confounded with 'standard deviation of a sample'

$_{\text{size}}$, the number of effect sizes; when $N_{\text{effect size}} = N_{\text{study}}$, the number of studies, effect sizes are usually independent). The proportion of σ_s^2 against the total variance is often quantified as $I^2 = \sigma_s^2/(\sigma_s^2 + \overline{v})$, where $\overline{v}$ is referred to as the 'typical' within-study (sampling) variance, which can be considered as a mean value of v_i (Higgins & Thompson, 2002). 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 et al., 2016). Therefore, publication bias tests assuming homogeneity (I^2 or $\sigma_s^2 = 0$) 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, especially for ecologists and evolutionary biologists (cf. Moller & Jennions, 2001; Jennions et al., 2013; for thorough technical reviews, see Rothstein et al., 2005; Vevea et al., 2019; Marks-Anglin & Chen, 2020a; Marks-Anglin et al., 2021). Therefore, we summarize different methods of testing for the presence of publication bias and assessing its impact on meta-analytic findings—describing which methods are suitable for datasets with high heterogeneity and non-independence. Our recent survey of publication bias tests used in 102 ecology and evolutionary meta-analyses indicates that many of these methods will be unfamiliar to ecologists and evolutionary

biologists; Figure 2 shows the results of the survey (for the details of survey procedure, see Appendix S1).

Following Sutton (2009; see also Vevea et al., 2019), we categorize publication bias tests into two types: (a) detecting publication bias (e.g. funnel plots, Egger's regression; Section 3.1) and (b) 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). Commonly, the magnitude of the overall effect is exaggerated because statistically non-significant effect sizes are less likely to be published, especially when they are based on small sample sizes. For time-lag bias, the magnitude of effect size and its statistical significance are related to publication year so that this bias requires different tests from publication and outcome reporting bias (see Section 3.1.3).

3.1 | Detecting publication bias

3.1.1 | Funnel 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 a funnel plot. These graphs are the most popular method for detecting publication bias in ecological and evolutionary meta-analyses (Figure 2). Funnel plots are also the most preferred graphical tool to detect publication bias in the medical and social sciences (Marks-Anglin & Chen, 2020a; Sterne et al., 2005; Sutton, 2009; Vevea et al., 2019), even though many other graphical methods have been proposed such as weighted histograms and normal quantile plots of effect sizes (as in Figure 2; for other graphical methods, see Rothstein et al., 2005; Marks-Anglin & Chen, 2020a).

The original funnel plot used sample size as the measure of uncertainty (Light & Pillemer, 1984; Figure 3a). Yet, more recent recommendations are to use *SE*, precision, variance or the inverse of variance (Figure 1; Sterne et al., 2005; but for why sample size may often be preferred, see Section 4.3). For these four quantities, unlike for sample size, we can draw 95% confidence intervals (based on the *y*-axis; $1.96 \times SE$) that create a funnel, showing the degree of heterogeneity among effect sizes (if data are homogeneous, most dots will be inside the 95% confidence interval region, e.g. Figure 3b,c). This confidence region also makes it easier to see funnel asymmetry caused by the lack of statistically non-significant effect sizes with high uncertainties (see Figure 3b,c). In a similar vein, a contour-enhanced funnel plot shows different statistical significance regions (around 0) to help detect asymmetry (Peters et al., 2008; Figure 3c). Lastly, Kossmeier et al. (2020) have recently proposed a sunset funnel plot, a type of contour-enhanced plot, which adds visual indicators of statistical power (Figure 3d).

One of the limitations of funnel plots is that funnel asymmetry can be caused not just by publication bias (as in Figure 3b, missing large effect sizes of high uncertainties or unexpected missing data points can create such asymmetry; see also Terrin et al., 2005). For instance, heterogeneity among effect sizes can create asymmetries

(A) Funnel plots — 32.4%
(B) Normal quantile (QQ) plots — 0.5%
(C) Correlation-based methods — 9.4%
(D) Regression-based methods — 11.7%
(E) Fail-safe *N* — 14.1%
(F) Trim-and-fill tests — 7.5%
(G) *p*-value-based methods — 1.4%
(H) Selection models — 0.0%
(I) Time-lag bias tests — 4.7%
(J) None reported — 17.8%
(K) Other (weighted histogram) — 0.5%

number of papers reporting test
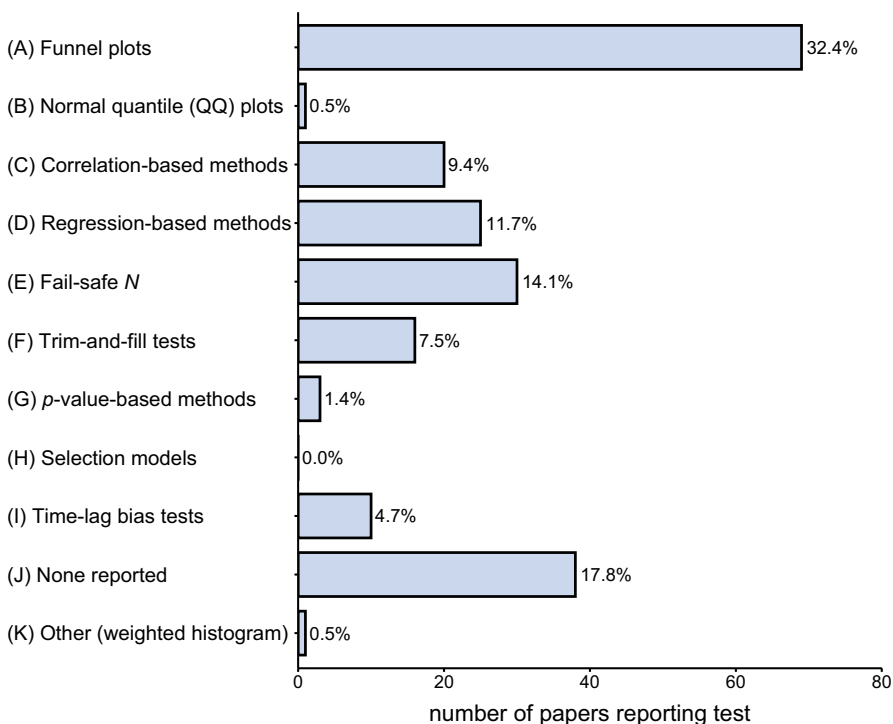
FIGURE 2 Frequencies of the usages of different publication bias tests in our survey of 102 meta-analyses in ecology and evolution. Note that only one paper employed a method (a weighted histogram) belonging to a category that was not pre-specified (including 'None reported'; the labels for items A–K match the labels used in our survey). For the details of the survey, see Appendix S1
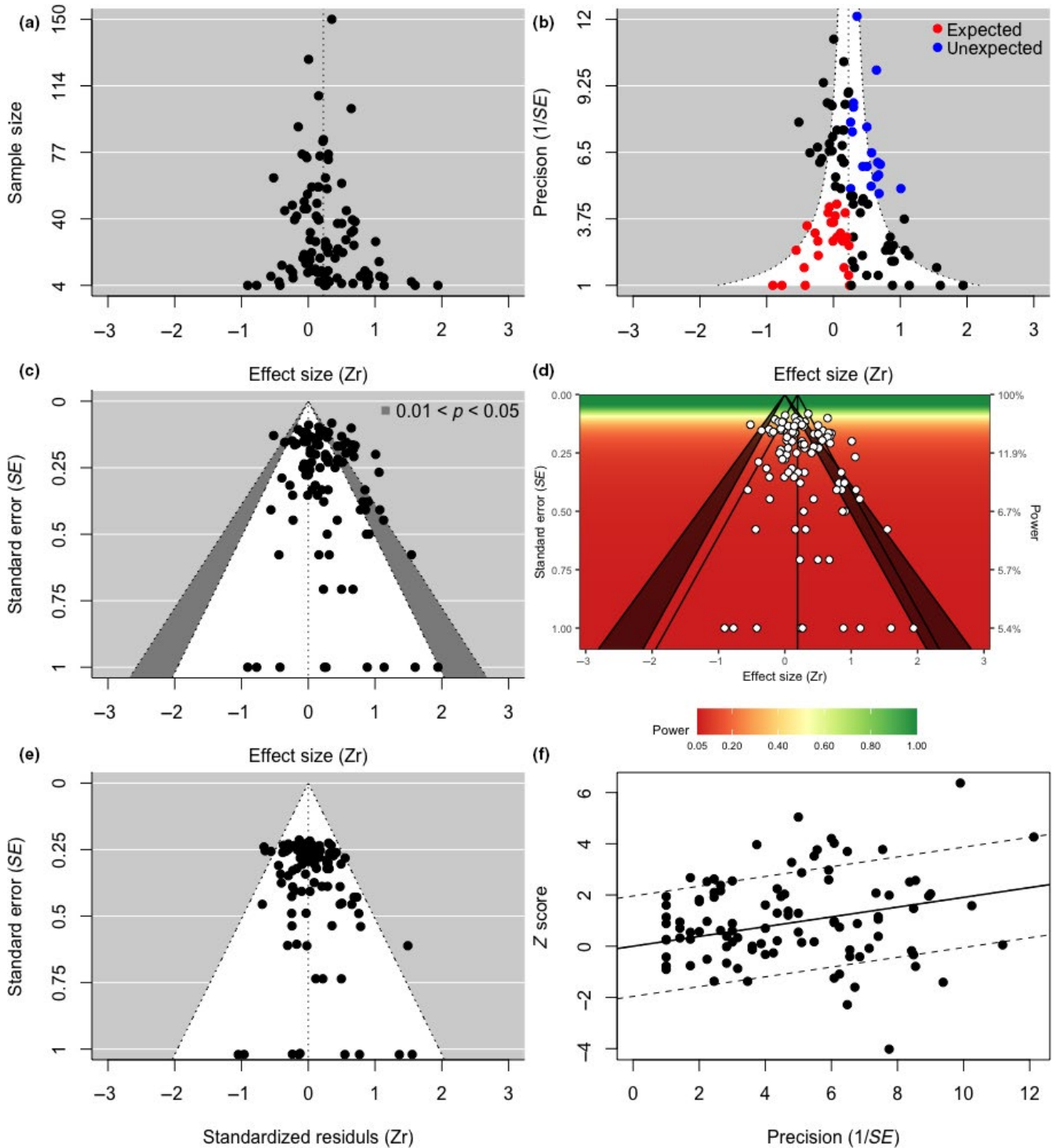
FIGURE 3 Examples of funnel plots and a radial plot using the same dataset ($N_{effect\ size} = N_{study} = 100$): (a) a funnel plot with sample size as a measure of uncertainty; (b) a funnel plot with precision ($1/SE$) as a measure of uncertainty, red dots representing 'expected' missing data under publication bias, and blue dots representing 'unexpected' missing data; (c) a counter enhanced funnel plot with SE as a measure of uncertainty; (d) a sunset plot showing statistical power of data using the overall effect estimate as a true effect (the black line indicates the overall effect); (e) a residual funnel plot from a meta-regression with one moderator; and (f) a radial plot showing the overall effect by its slope's steepness and heterogeneity with the degree of scattering of the data points (for more details, see the main text). We used the R packages METAFOR (panels a–c and e; Viechtbauer, 2010), METAVIZ (panel d; Kossmeier et al., 2020) and META (panel f; Schwarzer et al., 2015) for visualizations

of many kinds (Figure 3b). Incidentally, the other potential sources of asymmetry are data irregularities (e.g. mistakes, frauds, unique observations; cf. Nakagawa & Lagisz, 2016), artefacts and chance (Egger et al., 1997). Among these other sources, it is artefacts due to the intrinsic associations between many standardized effect size statistics and sampling variance (or *SE*) that are probably most important. Therefore, we expand on this later (see Section 4.3).

As mentioned above, high heterogeneity is common in ecological and evolutionary meta-analyses (Senior et al., 2016). Therefore, a standard funnel plot is unlikely to be informative about publication bias. To account for some of the heterogeneity, several researchers recommend plotting residuals from a meta-regression model (Figure 3e; e.g. Roberts & Stanley, 2005). In practice, however, no meta-regression model would explain all the heterogeneity. The remaining heterogeneity might 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 necessarily (Sterne et al., 2005, 2011). Although extensive work exists on funnel plots and heterogeneity, no systematic studies exist asking how funnel plots perform when effect sizes are correlated (but see Section 4.1).

Before moving to the next section where we introduce inferential tests of funnel asymmetry (or small-study effects), the radial plot proposed by Galbraith (1988) is worth mentioning, even though our survey found no use of these plots in ecological and evolutionary meta-analyses. The idea of a radial plot is similar to that of a funnel plot. The radial plot shows effect sizes divided by their *SE*s (essentially, z scores) on the y-axis and corresponding precisions on the x-axis. The plot, as in Figure 3f, has a slope with a zero intercept (solid line) and its 95% confidence interval based on lines drawn from ± 1.96 values (dashed lines) with the steepness of the slope representing the overall mean. 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 (but not those of the funnel plot) help us better understand the original inferential test for observed funnel asymmetry, the so-called Egger's regression (Egger et al., 1997), 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:

$$z_i = \beta_0 + \beta_1 \text{prec}_i + e_i, \tag{8}$$

$$e_i \sim \mathcal{N}\left(0, \sigma_e^2\right),$$

where z_i is the ith z score obtained from dividing an effect size by its *SE* (y_i/se_i), β_0 is the intercept, β_1 is the slope for the precision (prec or $1/se$) and e is residuals, normally distributed with a variance of σ_e^2. When β_0 (not β_1) is statistically significantly different from zero, then

we statistically detected funnel asymmetry (Figure 4a); the more β_0 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 line (β_1 and β_0) using Equation 8 in a funnel plot but this could be a confusing task as ones needs to put β_1 as the intercept and β_0 as the slope. However, it is possible to reformulate Egger's regression (Equation 8) so that its intercept (β_0) and its slope (β_1) can directly be used in a funnel plot, using a weighted regression, as follows (Thompson & Sharp, 1999):

$$y_i = \beta_0 + \beta_1 se_i + \epsilon_i, \tag{9}$$

$$\epsilon_i \sim \mathcal{N}\left(0, v_i \phi\right),$$

where y_i is the ith effect size and ϵ_i is the residuals, normally distributed with a variance of $v_i\phi$, which is sampling variance (v) and the multiplicative parameter (ϕ) estimated in the weighted regression (in a meta-regression, ϕ is set to be 1, which assumes that v_i is the exact sampling variance; see the next equation and also cf. Equation 7). Notably, Equation 8's β_0 is identical to Equation 9's β_1 and also Equation 8's β_1 is identical to Equation 9's β_0 (we demonstrate this in Appendix S2). Therefore, we can now look at the statistical significance of the slope of *SE* (se_i in Equation 9), whose magnitude indicates the severity of asymmetry, and we are also 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 et al., 2009), for example:

$$y_i = \beta_0 + \beta_1 se_i + s_i + m_i, \tag{10}$$

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

which is the same as Equation 7 (the random-effects model) plus the slope of *SE* (β_1) (note that different variants have precision, variance or the inverse of variance instead of *SE*; Moreno et al., 2009).

According to simulation studies (Macaskill et al., 2001; Moreno et al., 2009; Sterne et al., 2001), Egger's regression and its variants suffer from low power and poor performance when there are fewer than 20 effect sizes, or when the overall effect is large. However, meta-analyses in ecology and evolution often include over 20 effect sizes and our overall effect is usually small (Senior et al., 2016). 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 it is possible to: (a) add moderators to absorb some heterogeneity and (b) use multilevel meta-regression to account for non-independence among effect sizes. We expand on these possibilities in Section 4.

Similar to regression-based publication bias tests, correlation-based methods also statistically test for a relationship between effect sizes and corresponding uncertainties (e.g. sampling variance).
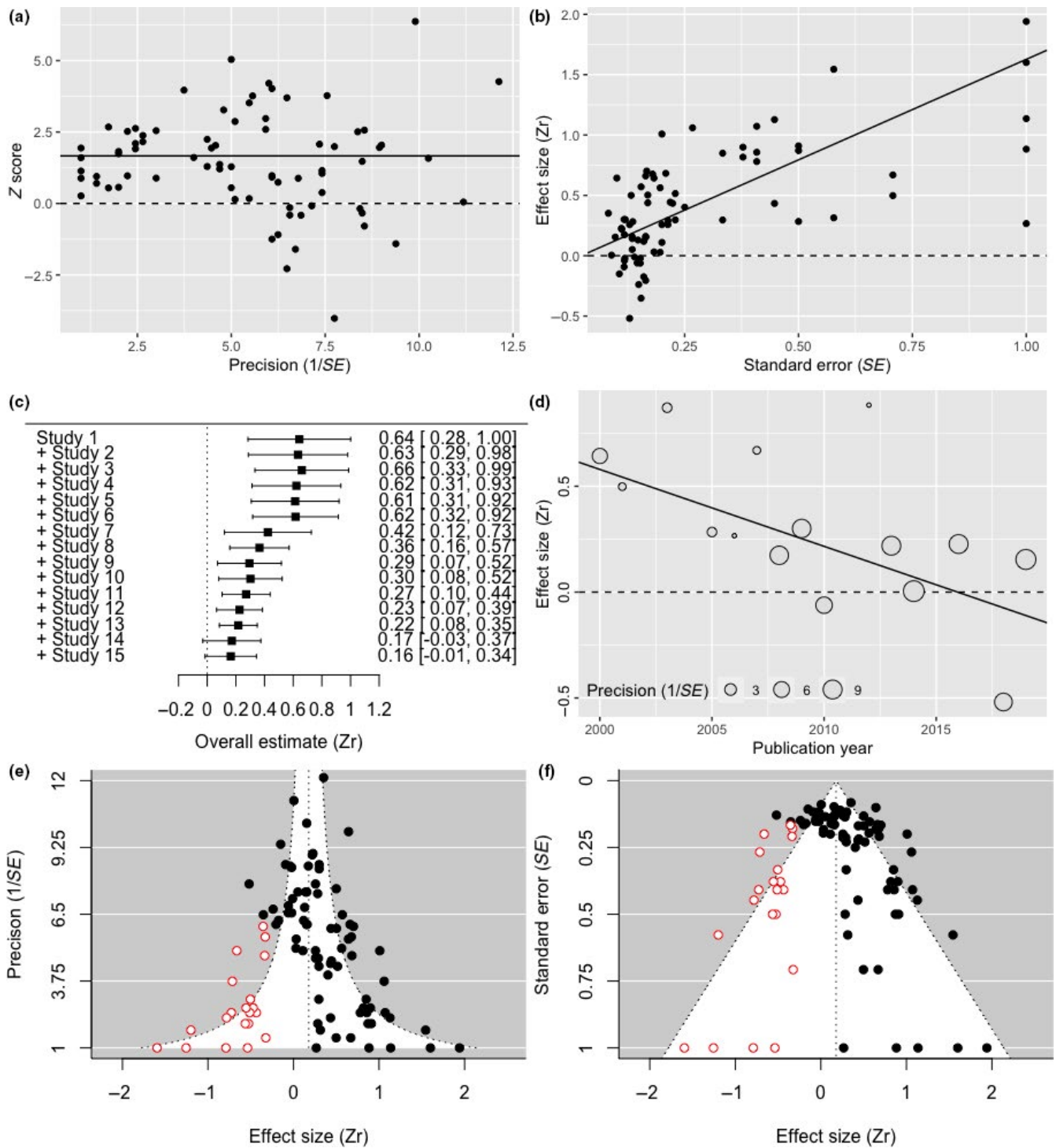
FIGURE 4 Examples of various plots (using the same dataset as Figure 3b minus 25 red datapoints; therefore, $N_{\text{effect size}} = 75$): (a) a scatter plot with the height of the solid line representing the degree of funnel asymmetry (cf. the radial plot at Figure 3f); (b) a scatter plot with the steepness of the slope representing the degree of funnel asymmetry; (c) a forest plot showing results of cumulate meta-analyses, where only a portion of the dataset ($N_{\text{effect size}} = 15$) was used; (d) a bubble plot showing a 'decline effect' over time, where only a portion of the dataset ($N_{\text{effect size}} = 15$) was used; (e) a funnel plot with precision (1/SE) and with a trim-and-fill method filling missing data (red circles; using the R_0 estimator); and (f) the same as panel (e) but with SE as a measure of uncertainty. We used the R packages GGPLOT2 (panels a, b and d; Wickham, 2009) and METAFOR (panels e and f; Viechtbauer, 2010) for visualizations

All the correlation methods are based on a version of the rank correlation test first proposed by Begg and Mazumdar (1994). This method essentially calculates a Kendall's rank correlation between effect sizes and their sampling variance (or other uncertainty measures, including sample size); a statistically significant correlation can indicate a small-study effect. Thus, it is very simple to implement,

but it seems that the rank correlation is less powerful than Egger's regression under many circumstances (Macaskill et al., 2001). Also, a recent simulation shows that the rank correlation methods, using both sampling variance and sample size, had severely inflated Type I error rates when effect sizes are correlated (Fernandez-Castilla et al., 2021). Therefore, we recommend that meta-analysts use regression-based methods instead of correlation-based methods to test for publication bias (in our survey, these methods were roughly equally popular, being reported in around 10% of papers; Figure 2).

3.1.3 | Time-lag bias tests

Time-lag bias occurs when larger or statistically significant effects are published more quickly than smaller or non-statistically significant effects, and can manifest as a decline in the magnitude of the overall effect over time (i.e. a decline effect; Koricheva & Kulinskaya, 2019). 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, as time-lag bias is likely to be prevalent in ecology and evolution (Jennions & Moller, 2002; Sánchez-Tójar et al., 2018). To test for time-lag bias, we caution against using correlation-based methods because this approach does not account for effect size precision (e.g. quantifying a rank correlation between effect size and publication year; Barto & Rillig, 2012). Instead, there are two recommended ways to investigate time-lag bias (or a decline effect): (a) using a cumulative meta-analysis and (b) using a regression-based method (see Koricheva et al., 2013; Koricheva & Kulinskaya, 2019; Trkalinos & Ioannidis, 2005). Regardless of the method, the key feature of time-lag bias tests is that, as more studies accumulate, the mean effect size is expected to converge on its true value. As such, we expect to see a change in the mean effect size as studies accumulate across the time.

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 statistical significance or magnitude of the overall effect size changes over time. 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. For complex data structures (see Section 4.1), limited sample sizes might prevent models from running in the early years of the dataset.

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

$$y_i = \beta_0 + \beta_1 \text{year}_i + s_i + m_i, \tag{11}$$

where year_i is the publication year for the ith study (effect size). It is noted that, in Equation 11, we are assuming a simple linear change in effect size over time which is to be expected with a 'decline effect' (as described above). This assumption may be unrealistic depending on how time-lag bias manifests. One way of dealing with this is to

model the logarithm of (publication) year, the inverse of year or the quadratic effect of year along with the linear effect (for an example of log 'publication year', see Stanley & Jarrell, 1998; see also Jarrell & Stanley, 2004). As with Equation 8, this method can accommodate other moderators (i.e. potential confounding variables) and also can be extended 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 the methods that can assess the impact of publication bias rather than merely detecting it. Fail-safe N (also known as the 'file-drawer number') represents the number of statistically non-significant unpublished results needed to exist to make the overall effect non-significant (e.g. Rosenberg, 2005; Rosenthal, 1979) or negligible in magnitude (e.g. Owrin, 1983). If the fail-safe N is large ($>5N_{\text{study}} + 10$), the results of the analyses may be considered to be robust with respect to publication bias as such large number of statistically non-significant results is unlikely to exist. The original fail-safe approach by Rosenthal (1979) is the oldest publication bias assessment method and probably the simplest:

$$N_{\text{Rosenthal}} = \left(\frac{\sum_{i=1}^{N_{\text{study}}} z_i}{1.645} \right)^2 - N_{\text{study}}, \tag{12}$$

where z_i is the ith z value (y_i/se_i) as in Equation 7 and 1.645 is the z value for $\alpha = 0.05$ (the one-tailed test). The method by Owrin (1983) relies on the magnitude of the effect size rather than statistical significance; one version of this method can be written as:

$$N_{\text{Orwin}} = \frac{N_{\text{study}} \left(\bar{y} - y_n \right)}{y_n}, \tag{13}$$

where $\bar{y}$ is the overall mean (i.e. an estimate from a fixed-effect model) and y_n 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) explicitly includes such information. An equation that assumes a fixed-effect model can be written as:

$$N_{\text{Rosenberg}} = \frac{N_{\text{study}} W}{\sum_{i=1}^{N_{\text{study}}} w_i}, \tag{14}$$

$$W = \left(\frac{\sum_{i=1}^{N_{\text{study}}} w_i y_i}{t_{0.05(N_{\text{stduy}})}} \right)^2 - \sum_{i=1}^{N_{\text{study}}} w_i,$$

where w_i is the inverse of sampling variance or weight ($1/v_i$; note that w_i can be modified for a random-effects model), W is the amount

of additional weight required to reach statistical significance and $t_{0.05(N_{stduy})}$ 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).

Although fail-safe approaches are the most popular method after the funnel plot in our survey (14.1%), Becker (2005) has called for abandoning all fail-safe approaches, now that other methods for handling publication bias are available. Becker 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 a variety of fail-safe numbers can be obtained for the same dataset depending on the exact methods. For example, the R package METAFOR implements the three methods above (Viechtbauer, 2010); its example dataset shows $N_{Rosenthal} = 598$, $N_{Orwin} = 84$, and $N_{Rosenberg} = 370$ (for details, see Appendix S3). Unfortunately, none of the proposed methods adequately control for heterogeneity (e.g. by incorporating moderators) or non-independence among effect sizes. Furthermore, none of the methods of fail-safe N are inferential.

3.2.2 | Trim-and-fill tests

The trim-and-fill test provides a nonparametric method that can visualize potentially missing data, and statistically both detect and correct for funnel asymmetry (Duval & Tweedie, 2000a, 2000b). 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 & Lin, 2019), and this method is used in 7.5% of the ecology and evolution meta-analyses in our survey. In short, this method uses an iterative process to determine how many effect sizes are missing (say, $N_{missing}$) from a funnel, using an initial overall estimate and one of three estimators (R_0, L_0 & Q_0; see an accessible account in Duval, 2005). Then, it 'trims' off $N_{missing}$ effect sizes to suppress funnel asymmetry, and estimates a new overall mean to see whether it can trim more effect sizes until the value $N_{missing}$ stabilizes. Subsequently, $N_{missing}$ effect sizes are 'filled' as mirror images (Figure 4e,f). Finally, an overall effect is re-estimated including the filled values. We note that Duval (2005) 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 with 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 et al., 2009; Peters et al., 2007). Although trim-and-fill tests have been extended to meta-regressions (Weinhandl & Duval, 2012), the implementation of this extension is currently limited to one moderator. Furthermore, recent simulation work by Rodgers and Pustejovsky (2021) shows that ignoring non-independence and fitting a trim-and-fill method (using R_0) increases Type I error rates, especially when a large overall effect exists.

3.2.3 | *p*-value-based methods and selection models

Ecologists and evolutionary biologists have hardly used the available methods based on p-values and selection models (p-value-based: 1.4%, selection models: 0%, Figure 2), even though both types of methods can provide adjusted overall means. The p-curve method was introduced by the same researchers who popularized the terms 'researcher degrees of freedom' (Simmons et al., 2011) and 'p-hacking' (Simonsohn et al., 2014). The p-curve method relies on the distribution of statistically significant p values of effect sizes in a dataset (Figure 5a). The p-uniform method is a similar method, which also exploits the distribution of p values (van Assen et al., 2015). Interestingly, McShane et al. (2016) have pointed out that both p-curve and p-uniform tests are versions of a selection model first suggested by Hedges (1984); all of these methods, unfortunately, do not perform well with heterogeneity as they assume one true effect (see also, van Aert & van Assen, 2021; van Aert et al., 2016). 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, complex class of publication bias methods (reviewed in Marks-Anglin & Chen, 2020a; Rothstein et al., 2005; Vevea et al., 2019). There are probably as many selection models as all other methods combined (Marks-Anglin & Chen, 2020a), but a property common to all of them is that they model how effect sizes are missing (or selected to be published), based on one or more statistical parameters, for example, p values, effect sizes or sampling variance (e.g. Carter et al., 2019; Preston et al., 2004; Rodgers & Pustejovsky, 2021; Figure 5b,c). Importantly, selection models can tolerate and model heterogeneity. Indeed, the recent model by Citkowicz and Vevea (2017) 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 non-independent effect sizes, and as far as we are aware, such implementation is extremely challenging.

4 | METHODS FOR DEPENDENT EFFECT SIZES

In this section, we first define a multilevel model that explicitly incorporates non-independence among effect sizes. Next, 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 testing method. This new method can both detect and correct for funnel asymmetry or small-study effects, while modelling heterogeneity and complex non-independence involving both correlation and variance–covariance matrices.
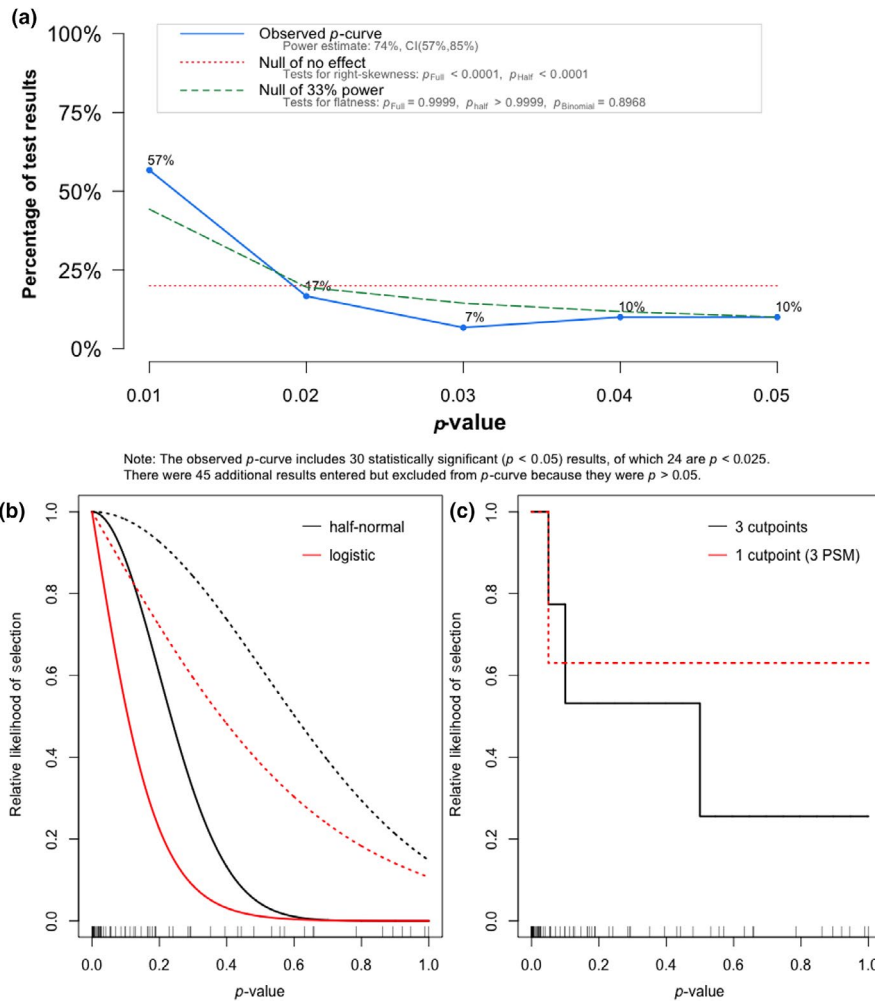
FIGURE 5 Example plots for p-curves and selection models (using the same dataset as in Figure 4; $N_{\text{effect size}} = 75$): (a) a line plot showing the distribution of statistically significant p values under three scenarios: (1) with the observed p values (blue solid line), (2) when there is no effect (red dotted line) and (3) when there is an effect (i.e. an observed overall effect as a true effect) with 33% statistical power (note that if a blue line increases at the α level of 0.05, this is a sign of p-hacking; for more details of this plot, see www.p-curve.com); (b) a plot showing four different weight functions that model, based on the data, the likelihood of effect sizes being selected for publication: (1) a half-normal function based on p values (black solid line), (2) the same function but based both on p values and precisions (black dotted line), (3) a logistic function based on p values (red solid line) and (4) the same function but based both on p values and precisions (red dotted line; these functions are based on Preston et al., 2004); and (c) a plot showing two different 'step' weight functions based on: (1) three cutpoints ($\alpha = 0.05, 0.1, 0.5$) and (2) one cut-point ($\alpha = 0.05$; this model is sometimes referred to as a three-parameter selection model, PSM with the three parameters being an overall mean, the between-study variance and an index determining the likelihood of selection; e.g. Carter et al., 2019; Rodgers & Pustejovsky, 2021). We used the R packages DMETAR (panel a; Harrer et al., 2021) and METAFOR (panels b and c; Viechtbauer, 2010) for visualizations

4.1 | A multilevel meta-analysis and funnel plots

The simplest multilevel meta-analytic model can be written as (Nakagawa & Santos, 2012):

$$y_i = \beta_0 + s_j + u_i + m_i, \tag{15}$$

$$s_j \sim \mathcal{N}\left(0, \sigma_s^2\right), \quad u_i \sim \mathcal{N}\left(0, \sigma_u^2\right), \quad m_i \sim \mathcal{N}\left(0, v_i\right),$$

where β_0 is the overall estimate (or meta-analytic mean); s_j is the between-study effect for the jth study, normally distributed with the variance of σ_s^2; u_i is the between-effect-size effect, or

within-study effect, for the ith effect size, distributed with a mean of zero and the variance of σ_u^2; and m_i is the sampling error (as in Equation 7; note that $j = 1, 2, ..., N_{\text{study}}$, the number of studies, and $i = 1, 2, ..., N_{\text{effect size}}$, the number of effect sizes; $N_{\text{effect size}} > N_{\text{study}}$). Equation 15 explicitly models multiple effect sizes per study. Also, in Equation 7, the term σ_s^2 is the only source of heterogeneity, while in Equation 15, both σ_s^2 and σ_u^2 are each contributing to heterogeneity among effect sizes.

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

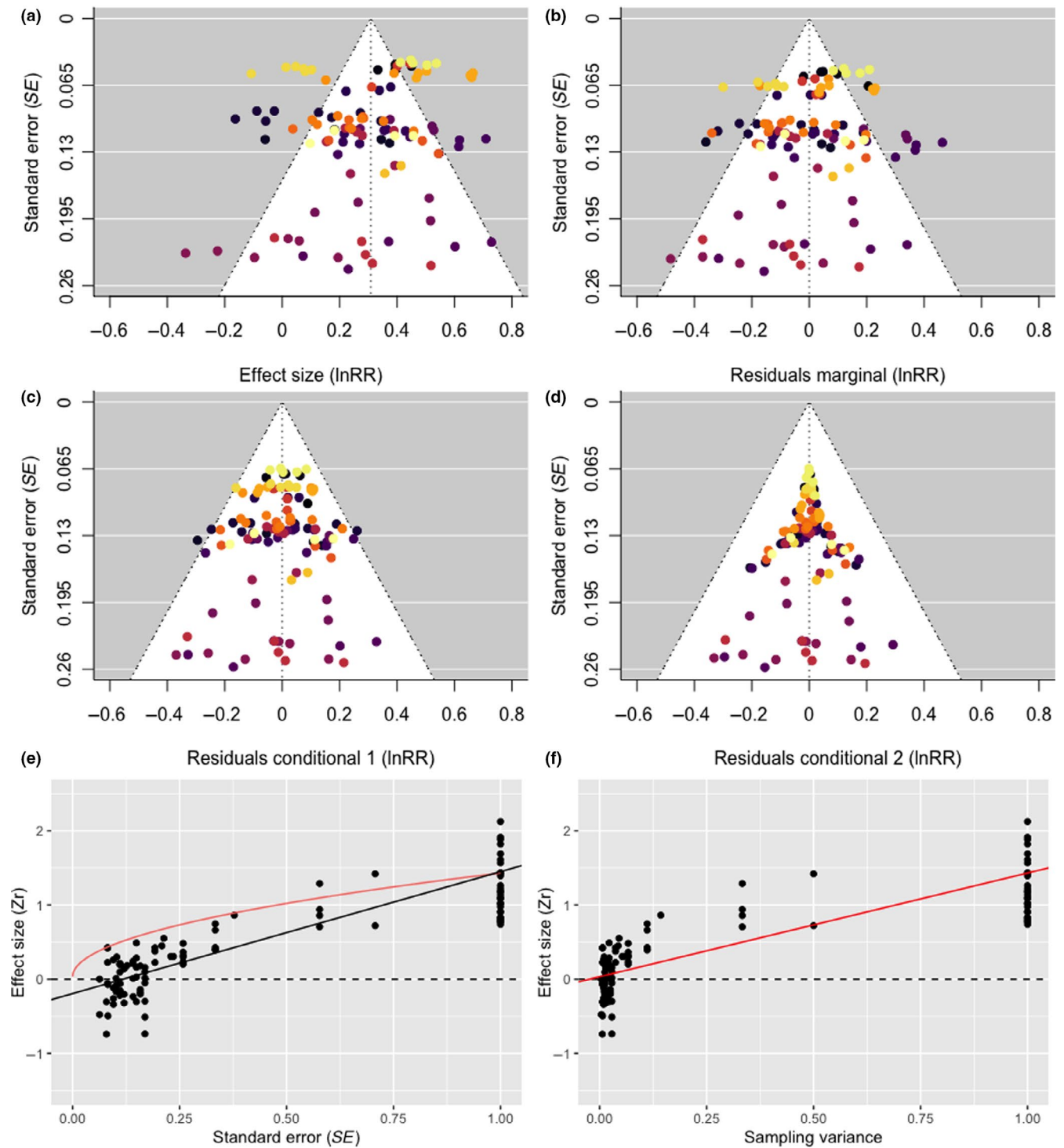$$y_i = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i} + s_j + u_i + m_i, \tag{16}$$

FIGURE 6 Examples of funnel plots from a dataset with lnRR (N_{study} = 70; $N_{effect\ size}$ = 271; panels a–d) and a different dataset with Zr (N_{study} = 48; $N_{effect\ size}$ = 104; panels e, f): (a) a funnel plot of raw data (the same colour indicating effect sizes from the same studies); (b) a funnel plot of marginal residuals with the fixed effects removed (as in Equation 17); (c) a funnel plot of conditional residuals with fixed effects and the between-study effect removed (as in Equation 18); and (d) a funnel plot of conditional residuals with all effects apart from sampling errors removed (as in Equation 19); (e) a scatterplot showing a meta-regression on SE (black line; the red line is the same line as in panel (f), scaled to the standard error shown on the x-axis of panel e). Note that an overall mean is set to be 0 in this simulated dataset along missing effect sizes imitating publication bias; and (f) a scatterplot showing a meta-regression on sampling variance (red line, an equivalent line as the one in panel e). The red lines, for both panel (e) and panel (f), intersect the effect size at the intercept because they are the same regression lines plotted on different x-axes. We used the R packages METAFOR (panels a–d; Viechtbauer, 2010) and GGPLOT2 (panels e, f; Wickham, 2009) for visualizations

where β_1 is the slope for x_1, a study-level moderator (characteristics of different studies, j; e.g. experimental vs. observational) and β_2 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 (for an example of a funnel plot with non-independent effect sizes, see Figure 6a). A complication is that, given Equation 15, we can extract at least three different residuals, which are:

$$\text{resid}_{mi} = y_i - \left(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i}\right), \tag{17}$$

$$\text{resid}_{c1i} = y_i - \left(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i} + s_j\right), \tag{18}$$

$$\text{resid}_{c2i} = y_i - \left(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i} + s_j + u_i\right), \tag{19}$$

where resid_m represents marginal residuals (subtracting only fixed effects from the observations; Figure 6b), whereas resid_{c1} and resid_{c2} are conditional residuals (Figure 6c,d; Nobre & Singer, 2007). As shown in Figure 6a–d, marginal residuals still show the patterns due to study origin (i.e. sample sizes are the same or similar). Contrastingly, conditional residuals no longer show such obvious patterns as we have taken a clustering factor (s_j), meaning that these residuals are independent, at least with respect to this factor. Thus, funnel plots with conditional residuals (Figure 6c,d) seem like a useful exploratory tool for publication bias when effect sizes are correlated, in addition to using marginal residuals (Figure 6b).

As the conditional residuals are supposed to be independent, Nakagawa and Santos (2012) suggested using conditional residuals along with corresponding sampling variance or standard error (v_i or se_i) in publication bias tests (e.g. the original Egger's regression and trim-and-fill tests). However, this approach is limited by some assumptions. First, all such residual analyses assume that sampling SE (se_i) does not covary with moderators in meta-regression (e.g. x_1 and x_2 in Equation 16; see Freckleton, 2002). Second, sampling SE is assumed to be the same as the SE of the residuals (which are shown in Figure 6b–d), but they are not the same, although they are often strongly correlated (see Doleman et al., 2020). Finally, in the presence of non-independent data, Equation 15's sampling variances are often correlated; that is, $m_i \sim \mathcal{N}(0, \mathbf{M})$ where $\mathbf{M}$ is a variance–covariance matrix. For example, when $N_{\text{effect size}} = 3$ and the first two effect sizes' sampling variance are correlated, then we can write $\mathbf{M}$ as:

$$\mathbf{M} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}, \tag{20}$$

where ρ is the correlation between the sampling effects of the first two effect sizes ($\rho\sigma_1\sigma_2$ is the covariance). Whenever sampling (error) effects are correlated, neither resid_{c1} nor resid_{c2} are independent. Then, none of publication bias tests reviewed in Section 3 should be used. Incidentally, we note that the robust variance estimator (RVE) originally proposal by Hedges et al. (2010) can circumvent modelling

the variance–covariance matrix $\mathbf{M}$ even when sampling errors are correlated. This is because covariances are estimated from the data and the associated errors are reflected in standard errors (variance) of point estimates via the RVE (cf. Rodgers & Pustejovsky, 2021; also see Bom & Rachinger, 2020).

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; Doucouliagos & Stanley, 2009; Fernandez-Castilla et al., 2021; Havranek & Irsova, 2011; Rodgers & Pustejovsky, 2021):

$$y_i = \beta_0 + \beta_1 se_i + s_j + u_i + m_i. \tag{21}$$

By examining Equation 21, we may realize that β_0 represents a conditional estimate of an overall effect when SE is 0, which means, theoretically, there is no uncertainty (Figure 6e). Then, we can ask 'does β_0 provide an adjusted estimate of an overall effect, when β_1 is statistically significant (i.e. detecting a small-study effect)?' Stanley and Doucouliagos (2012, 2014) have shown that, with statistically significant β_1, β_0 provides an adjusted estimate that is downwardly biased, when a true positive or a null effect exists (Figure 6e). They also state that with non-statistically significant β_1, β_0 provides the best estimate of an adjusted mean. If the slope of SE (β_1) is statistically significant then fitting sampling variance instead of SE is recommended according to the following equation:

$$y_i = \beta_0 + \beta_1 v_i + s_j + u_i + m_i. \tag{22}$$

This is equivalent to fitting se_i^2, which is a quadratic term. Stanley and Doucouliagos (2012, 2014) have shown that β_0 in Equation 22 is still downwardly biased, but much less so, although Equation 21 is more powerful (i.e. an adjustment tends to underestimate) when there is a positive (or no) effect (cf. Figure 6f). While this two-step approach, using Equations 21 and 22, may seem simplistic (see also Stanley, 2017; Stanley et al., 2017), it provides an easy-to-implement publication bias test that explicitly models non-independent data.

Furthermore, this regression approach can be used to test time-lag bias (or decline effect) by modelling the publication year (year_j):

$$y_i = \beta_0 + \beta_1 \text{year}_j + s_j + u_i + m_i. \tag{23}$$

When heterogeneity exists, it is best to combine Equations 21 and 23 with moderators. Such a model can be written as:

$$y_i = \beta_0 + \beta_1 se_i + \beta_2 c\left(\text{year}_j\right) + \sum_{k=3}^{N_{mod}} \beta_k x_k + s_j + u_i + m_i, \tag{24}$$

where β_k is the slope for the kth moderator ($k = 3, 4, ..., N_{\text{mod}}$; the number of moderators), the other parameters are as above, but one will need

to centre the moderator, $year_j$ (i.e. set the mean value of $year_j$ as 0) or other continuous variables to keep β_0 meaningful to be interpreted as an adjusted overall effect (see more details in Appendix S4). Similar models to Equation 24, including the publication year as a covariate, can be found in meta-analyses in the social sciences, especially the field of economics (e.g. Costa-Font et al., 2013; Havranek & Sokolova, 2020; Jarrell & Stanley, 2004; Matousek et al., 2021). However, simulation studies have shown Egger's regression variants with sampling standard error as a moderator (e.g. Equations 10 and 21) perform poorly, even when adequately powered (Deeks et al., 2005; Macaskill et al., 2001). This is especially true when there is a (mathematical) relationship between effect size and sampling SE not due to publication bias. Furthermore, sampling SE can often be poorly estimated.

4.3 | Multilevel meta-regression using sample size: A proposed approach

To understand how a correlation between effect size and SE can come about, and when SE can be estimated inaccurately, we now go back to comparing sampling variance among the three commonly used effect sizes (Equations 2, 4 and 6). The SMD's sampling variance contains the square of the point estimate (Equation 2), whereas lnRR's sampling variance contains both the treatment and control means that are also contained in the point estimate (Equation 4; Costa-Font et al., 2013; Zwetsloot et al., 2017; Doncaster & Spake, 2018; Pustejovskyei & Rodgers, 2019). This can lead to a correlation between point estimates (i.e. lnRR and SMD) and their sampling SE, resulting in 'artefactual' funnel asymmetry (Section 3.2; note that this issue is widespread, and also found in other standardized effect sizes, such as odds ratio and risk difference; Peters et al., 2006). Furthermore, we also notice that in Equation 4 (i.e. lnRR's variance), when sample sizes (n_1 and n_2) are small, the sample mean ($\overline{X}$) and particularly, the sample standard deviation (SD) will be poorly estimated. This will result in an unreliable estimate of sampling variance (this is also the case for Equation 2). These issues do not affect the sampling variance of Zr, which is a function only of sample size (n; Equation 6; cf. Rucker et al., 2008). Therefore, the sample size ($n_1 + n_2$) has been suggested as a moderator instead of SE (e.g. Equation 21) when we use effect size statistics such as SMD and lnRR (also correlation, r; see Section 2.1; Macaskill et al., 2001). Simulations suggest using the sample size as a moderator outperforms SE with close to nominal Type 1 error rates in the cases of both independent (Deeks et al., 2005; Macaskill et al., 2001), and non-independent effect sizes (Fernandez-Castilla et al., 2021).

Instead of the sample size ($n_1 + n_2$), however, for a meta-analysis of SMD or lnRR, we propose using the 'effective sample size' ($4\tilde{n}_i$ or just $\tilde{n}_i$) because it accounts for unbalanced sampling (cf. Stanley, 2005). The effective sample size is given by (Bakbergenuly et al., 2020a, 2020b; also see; Deeks et al., 2005; Bakbergenuly et al., 2020c):

$$4\tilde{n}_i = \frac{4n_{1i}n_{2i}}{n_{1i} + n_{2i}}. \tag{25}$$

When $n = n_1 = n_2$, the formula reduces to $2n$. Indeed, the inverse of $\tilde{n}_i$ is a part of sampling variance in both SMD and lnRR (Equations 4 and 6):

$$\frac{1}{\tilde{n}_i} = \frac{n_{1i} + n_{2i}}{n_{1i}n_{2i}} = \frac{1}{n_{1i}} + \frac{1}{n_{2i}}, \tag{26}$$

where the middle part of the formula corresponds to Equation 2 when setting SMD = 0, while the right-hand side corresponds to Equation 4 when setting CV ($SD/\overline{X}$) = 1. This means that the use of $\tilde{n}_i$ is comparable to that of sampling variance after taking out uncertain elements.

Taken together, we can rewrite Equations 21 and 22, respectively, as (Deeks et al., 2005):

$$y_i = \beta_0 + \beta_1 \sqrt{\frac{1}{\tilde{n}_i}} + s_j + u_i + m_i, \tag{27}$$

$$y_i = \beta_0 + \beta_1 \left(\frac{1}{\tilde{n}_i}\right) + s_j + u_i + m_i, \tag{28}$$

where $\sqrt{1/\tilde{n}_i}$ is a replacement of se_i in Equation 21, and $1/\tilde{n}_i$ is a replacement of v_i in Equation 22 (note that, at the intercept, $\tilde{n}_i$ is infinitely large). We recommend using Equation 27 to check the statistical significance of funnel asymmetry (small-study effects) because it has greater statistical power than Equation 28. Equation 27 can also be used to obtain an adjusted mean when β_1 is not statistically significant. This is because β_0 represents an adjusted overall mean when $\sqrt{\frac{1}{\tilde{n}_i}} = 0$. In other words, the predicted overall mean when a study has an infinitely large sample size, $\tilde{n}_i$, and therefore little to no sampling variance. In contrast, when β_1 is statistically significant in Equation 27, we recommend using Equation 28 to obtain an overall estimate adjusted for publication bias because it is less biased. Note that these recommendations are for the effect sizes SMD and 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 this adjusted estimate 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 et al., 2017).

In practice, multilevel meta-analytic models are often more complex than what is shown above. For example, Nakagawa and Santos (2012) 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 & Nakagawa, 2010). The major benefit of our proposed meta-regression approach for publication bias tests 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 that also includes phylogenetic and non-phylogenetic random effects can be written as:

$$y_i = \beta_0 + \beta_1 \sqrt{\frac{1}{\tilde{n}_i}} + \beta_2 c\left(year_j\right) + \sum_{k=3}^{N_{mod}} \beta_k x_k + a_h + q_h + s_j + u_i + m_i, \tag{29}$$

$$a_h \sim \mathcal{N}\left(0, \sigma_a^2 A\right), \ q_h \sim \mathcal{N}\left(0, \sigma_q^2\right), \ m_i \sim \mathcal{N}\left(0, \mathbf{M}\right),$$

where a_h is the phylogenetic effect for the hth species, considered multivariate normally distributed with a covariance of $\sigma_a^2 \mathbf{A}$ ($\mathbf{A}$ is a correlation matrix derived from a phylogeny); q_h is the non-phylogenetic effect for the hth species, distributed with the variance of σ_q^2 ($h = 1$, 2, ..., $N_{species}$, the number of species; $N_{species} \neq N_{study}$); and the other notations are the same as above. Relevantly, when using SMD or lnRR, we may be better off using $\tilde{n}_i$ along with residuals for drawing funnel plots (see Section 4.1; Doleman et al., 2020) rather than SE, precision, or variance. In the Supporting Information, we use two datasets and the three effect sizes to illustrate how to practically code these proposed methods; there, we have redrawn Figure 6a–d using $\tilde{n}_i$ instead of SE (see 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. When sampling variances are correlated (i.e. $\mathbf{M}$ 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 et al., 2009):

$$\text{Var}\left(\frac{1}{N_{within}} \sum_{g=1}^{N_{within}} y_g \right) = \left(\frac{1}{N_{within}} \right)^2 \left(\sum_{g=1}^{N_{within}} \sigma_g^2 + \sum_{g \neq l}^{N_{within}} r_{gl} \sqrt{\sigma_g^2 \sigma_l^2} \right), \quad (30)$$

where y_g and y_l are the gth and lth effect size in a study ($g = 1, ..., N_{within}$ and $l = 1, ..., N_{within}$ where N_{within} is the number of effect sizes within a paper or a species to be combined), σ_i^2 and σ_g^2 are the sampling error variances for y_g and y_l, and r_{gl} is the correlation between the sampling errors of y_g and y_l.

Overall means will generally not be biased using aggregated or single sample/study effect sizes (Song et al., 2020). Also, Rodgers and Pustejovsky (2021) showed that when averaging effect sizes within studies, 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 & Hedges, 1995) all had the appropriate level of Type 1 error, although the three-parameter selection model was noticeably more powerful in detecting publication bias than the others. However, averaging or sampling is not a general solution when we have a phylogenetic signal ($\sigma_a^2 > 0$; 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. $\mathbf{A}$ in Equation 29; Nakagawa, Senior, et al., 2021). Furthermore, even when there is no phylogenetic signal ($\sigma_a^2 = 0$), 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 measurement types, evaluation methods or sex). Although iteratively sampling one effect size per study could capture moderating effects, this approach also reduces the information content of the dataset. 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 Appendix S5).

5 | CONCLUSIONS

Given the high levels of heterogeneity and prevalence of non-independence in ecological and evolutionary meta-analytic datasets, our choice of suitable tests for publication bias is limited. We have described the main methods for testing publication bias alongside our recommendations, as summarized in Figure 7. Our proposed multilevel regression method appears to be the only practical method fulfilling statistical assumptions under most circumstances. Although using averaging or sampling are not a universal solution, they may be useful in supplementing our multilevel meta-regression method. This is because all publication bias tests should be seen as a part of sensitivity analysis (Noble et al., 2017), meaning that we should run more than one publication bias test.

Few simulation studies have explicitly investigated the performance of publication bias tests with non-independent data. Two simulation studies that we are aware of supported similar models to the multilevel-regression method we proposed here (Fernandez-Castilla et al., 2021; Rodgers & Pustejovsky, 2021). In addition, a general point to take from these two simulation studies is that most methods are prone to Type 2 error, with a possible exception of some selection models, even when the methods have nominal Type 1 error rates. Therefore, not detecting publication bias in a publication bias test should not be taken as a proof of no publication bias, including for multilevel regression. Clearly, we need more methodological and simulation-based work in the future.

Finally, we repeat that the results of publication bias tests should always be cautiously interpreted because no methods will ever be able to verify the actual number of missing effect sizes. By way of example, a recent study compared the results of 15 meta-analyses with pre-registered replication projects on the same topics (Kvarven et al., 2020). The overall effects from the replication projects were smaller than those of the meta-analyses indicating the meta-analysis results were likely susceptible to publication bias. Interestingly, the replication projects' estimates were also smaller than the adjusted effects from the trim-and-fill method and the three-parameter selection model. In contrast, the two-step regression model (the method by Stanley & Doucouliagos, 2012, 2014) provided similar estimates to the replication projects. This is good news as our main recommendation is a version of this two-step approach. Nonetheless, caution needs to be exercised to acknowledge the limitations and assumptions of any publication bias test. Overall, we suggest that all future meta-analyses in ecology and evolution should test for publication bias, and try to identify related biases. For meta-analysts to achieve this goal, all empiricists need to report their statistical results, including their sample sizes and estimates of uncertainty (SE and SD),

	Test by visual inspection	Adjusts for the overall mean	Deals with heterogeneity	Time-lag (decline) effect	Independence: recommend?	Non-independence: recommend?
Funnel plot	yes	no	yes	not applicable	yes	yes[1]
Regression method (non-multilevel)	no	yes[2]	yes	yes	yes	no
Correlation method	no	no	no	maybe	no	no
Cumulative meta-analysis (forest plot)	yes	not applicable	not applicable	yes	yes	no
Fail-safe *N* method	no	yes[3]	no	not applicable	no	no
Trim & fill method	no	yes	maybe[4]	not applicable	maybe	no
_p_-value based methods	no	yes	no[5]	not applicable	no	no
Selection models	no	yes	yes	not applicable	yes	no
Multilevel meta-regression	no	yes	yes	yes	not applicable	yes

FIGURE 7 A summary of main publication bias tests reviewed in this article, and our recommendations under two different conditions (effect sizes are independent or non-independent). Superscript notes: (1) for funnel plots, residuals from a meta-regression can be plotted instead of raw effect sizes, and using sample sizes instead of standard errors may be a good option for lnRR and SMD; (2) for non-multilevel regression methods, precision and sampling variance (or $\sqrt{1/\tilde{n}_i}$ and $1/\tilde{n}_i$) can be used; (3) technically, fail-safe *N* methods do not provide an adjusted overall mean, but the numbers indicate how many statistically non-significant studies (null effect sizes) would render the overall effect zero (or a particular small effect size value); (4) for trim-and-fill methods, although some heterogeneity can be tolerated, the ability to model moderators is limited; alternatively, residuals along with their corresponding variances could be used; and (5) as an exception, a new method, named *p*-uniform*, could potentially tolerate heterogeneity (van Aert & van Assen, 2021)

transparently and compressively (Gerstner et al., 2017; Hennessy et al., 2021).

CONFLICT OF INTEREST
The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS
Conceptualization: S.N. and R.E.O.; Data curation: R.E.O., A.S.-T. and Y.Y.; Formal Analysis, R.E.O., A.S.-T., Y.Y. and S.N.; Validation: R.E.O., D.W.A.N. and S.N.; Investigation: S.N., M.L., M.D.J., J.K., D.W.A.N., T.H.P. and R.E.O.; Visualization: S.N., M.L. and R.E.O.; Methodology: S.N.; Writing—original draft: S.N.; Project administration: S.N. and R.E.O.; Writing—review and editing: all authors. We note that the supplementary information (Appendices S1–S5) was put together by R.E.O., A.S.-T., Y.Y., and S.N.

PEER REVIEW
The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13724.

DATA AVAILABILITY STATEMENT
We have relevant data and code available at Zenodo https://zenodo.org/record/5504537#.YXIMxXlxWss (Nakagawa, O'Dea, et al., 2021).

ORCID
Shinichi Nakagawa https://orcid.org/0000-0002-7765-5182
Malgorzata Lagisz https://orcid.org/0000-0002-3993-6127
Michael D. Jennions https://orcid.org/0000-0001-9221-2788
Julia Koricheva https://orcid.org/0000-0002-9033-0171
Daniel W. A. Noble https://orcid.org/0000-0001-9460-8743
Timothy H. Parker https://orcid.org/0000-0003-2995-5284
Alfredo Sánchez-Tójar https://orcid.org/0000-0002-2886-0649
Yefeng Yang https://orcid.org/0000-0002-8610-4016
Rose E. O'Dea https://orcid.org/0000-0001-8177-5075

REFERENCES

Bakbergenuly, I., Hoaglin, D. C., & Kulinskaya, E. (2020a). Estimation in meta-analyses of mean difference and standardized mean difference. *Statistics in Medicine*, *39*, 171–191. https://doi.org/10.1002/sim.8422

Bakbergenuly, I., Hoaglin, D. C., & Kulinskaya, E. (2020b). Estimation in meta-analyses of response ratios. *BMC Medical Research Methodology*, *20*, 1–24. https://doi.org/10.1186/s12874-020-01137-1

Bakbergenuly, I., Hoaglin, D. C., & Kulinskaya, E. (2020c). Methods for estimating between-study variance and overall effect in meta-analysis of odds ratios. *Research Synthesis Methods*, *11*, 426–442. https://doi.org/10.1002/jrsm.1404

Barto, E. K., & Rillig, M. C. (2012). Dissemination biases in ecology: Effect sizes matter more than quality. *Oikos*, *121*, 228–235. https://doi.org/10.1111/j.1600-0706.2011.19401.x

Becker, B. J. (2005). Fail safe *N* or file-drawer number. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111–125). John Wiley.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088–1101. https://doi.org/10.2307/2533446

Bom, P. R. D., & Rachinger, H. (2020). A generalized-weights solution to sample overlap in meta-analysis. *Research Synthesis Methods*, *11*, 812–832.

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

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*, 115–144. https://doi.org/10.1177/2515245919847196

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods*, *22*, 28–41. https://doi.org/10.1037/met0000119

Cohen, J. (1988). *Statistical power analysis for the beahvioral sciences* (2nd ed.). Lawrence Erlbaum.

Costa-Font, J., McGuire, A., & Stanley, T. (2013). Publication selection in health policy research: The winner's curse hypothesis. *Health Policy*, *109*, 78–87. https://doi.org/10.1016/j.healthpol.2012.10.015

Deeks, J. J., Macaskill, P., & Irwig, L. (2005). The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology*, *58*, 882–893. https://doi.org/10.1016/j.jclinepi.2005.01.016

Doleman, B., Freeman, S. C., Lund, J. N., Williams, J. P., & Sutton, A. J. (2020). Funnel plots may show asymmetry in the absence of publication bias with continuous outcomes dependent on baseline risk: Presentation of a new publication bias test. *Research Synthesis Methods*, *11*, 522–534. https://doi.org/10.1002/jrsm.1414

Doncaster, C. P., & Spake, R. (2018). Correction for bias in meta-analysis of little-replicated studies. *Methods in Ecology and Evolution*, *9*, 634–644. https://doi.org/10.1111/2041-210X.12927

Doucouliagos, H., & Stanley, T. D. (2009). Publication selection bias in minimum-wage research? A meta-regression analysis. *British Journal of Industrial Relations*, *47*, 406–428. https://doi.org/10.1111/j.1467-8543.2009.00723.x

Duval, S. (2005). The trim and fill method. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 127–144). John Wiley.

Duval, S., & Tweedie, R. (2000a). A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98. https://doi.org/10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fernandez-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *Journal of Experimental Education*, *89*, 125–144. https://doi.org/10.1080/00220973.2019.1582470

Freckleton, R. P. (2002). On the misuse of residuals in ecology: Regression of residuals vs. multiple regression. *Journal of Animal Ecology*, *71*, 542–545. https://doi.org/10.1046/j.1365-2656.2002.00618.x

Friedrich, J. O., Adhikari, N. K. J., & Beyene, J. (2008). The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study. *BMC Medical Research Methodology*, *8*, 32. https://doi.org/10.1186/1471-2288-8-32

Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical-trials. *Statistics in Medicine*, *7*, 889–894. https://doi.org/10.1002/sim.4780070807

Gerstner, K., Moreno-Mateos, D., Gurevitch, J., Beckmann, M., Kambach, S., Jones, H. P., & Seppelt, R. (2017). Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. *Methods in Ecology and Evolution*, *8*, 777–784. https://doi.org/10.1111/2041-210X.12758

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, *555*, 175–182. https://doi.org/10.1038/nature25753

Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, *23*, 494–508. https://doi.org/10.1111/j.1420-9101.2009.01915.x

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2021). *Doing meta-analysis in R: A hands-on guide*. Chapman and Hall/CRC.

Havranek, T., & Irsova, Z. (2011). Estimating vertical spillovers from FDI: Why results vary and what the true effect is. *Journal of International Economics*, *85*, 234–244. https://doi.org/10.1016/j.jinteco.2011.07.004

Havranek, T., & Sokolova, A. (2020). Do consumers really follow a rule of thumb? Three thousand estimates from 144 studies say 'probably not'. *Review of Economic Dynamics*, *35*, 97–122. https://doi.org/10.1016/j.red.2019.05.004

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, *9*, 61–85. https://doi.org/10.3102/10769986009001061

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*, 1150–1156. https://doi.org/10.1890/0012-9658(1999)080%5B1150:TMAORR%5D2.0.CO;2

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. https://doi.org/10.1002/jrsm.5

Hennessy, E. A., Acabchuk, R. L., Arnold, P. A., Dunn, A. G., Foo, Y. Z., Johnson, B. T., Geange, S. R., Haddaway, N. R., Nakagawa, S., Mapanga, W., Mengersen, K., Page, M. J., Sánchez-Tójar, A., Welch, V., & McGuinness, L. A. (2021). Ensuring prevention science research is synthesis-ready for immediate and lasting scientific impact. *Prevention Science*. https://doi.org/10.1007/s11121-021-01279-8

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558. https://doi.org/10.1002/sim.1186

Jarrell, S. B., & Stanley, T. D. (2004). Declining bias and gender wage discrimination? A meta-regression analysis. *Journal of Human Resources*, *39*, 828–838. https://doi.org/10.2307/3558999

Jennions, M. D., Lorite, C. J., Rosenberg, M. S., & Rothstein, H. R. (2013). Publication and related biases. In J. Koricheva, J. Gurevitch, & K. Mengersen (Eds.), *The handbook of meta-analysis in ecology and evolution* (pp. 207–236). Princeton University Press.

Jennions, M. D., & Moller, A. P. (2002). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, *269*, 43–48. https://doi.org/10.1098/rspb.2001.1832

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, *102*, 828–844. https://doi.org/10.1111/1365-2745.12224

Koricheva, J., Jennions, M. D., & Lau, J. (2013). Temporal trends in effect sizes: Causes, detection and implications. In J. Koricheva, J. Gurevitch, & K. Mengersen (Eds.), *The handbook of meta-analysis in ecology and evolution* (pp. 237–254). Princeton University Press.

Koricheva, J., & Kulinskaya, E. (2019). Temporal instability of evidence base: A threat to policy making? *Trends in Ecology & Evolution*, *34*, 895–902. https://doi.org/10.1016/j.tree.2019.05.006

Kossmeier, M., Tran, U. S., & Voracek, M. (2020). Power-enhanced funnel plots for meta-analysis the sunset funnel plot. *Zeitschrift Fur Psychologie – Journal of Psychology*, *228*, 43–49. https://doi.org/10.1027/2151-2604/a000392

Kvarven, A., Stromland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*, 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lajeunesse, M. J. (2015). Bias and correction for the log response ratio in ecological meta-analysis. *Ecology*, *96*, 2056–2063. https://doi.org/10.1890/14-2402.1

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Harvard University Press.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*, 641–654. https://doi.org/10.1002/sim.698

Marks-Anglin, A., & Chen, Y. (2020a). A historical review of publication bias. *Research Synthesis Methods*, *11*, 725–742. https://doi.org/10.1002/jrsm.1452

Marks-Anglin, A., & Chen, Y. (2020b). Small-study effects: Current practice and challenges for future research. *Statistics and its Interface*, *13*(4), 475–484.

Marks-Anglin, A., Duan, R., Chen, Y., Panagiotou, O., & Schmid, C. H. (2021). Publication and outcome reporting bias. In C. H. Schmid, T. Stijnen, & R. W. White (Eds.), *Handbook of meta-analysis* (pp. 283–312). CRC.

Matousek, J., Havranek, T., & Irsova, Z. (2021). Individual discount rates: A meta-analysis of experimental evidence. *Experimental Economics*, https://doi.org/10.1007/s10683-021-09716-9

McShane, B. B., Bockenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749. https://doi.org/10.1177/1745691616662243

Møller, A. P., & Jennions, M. D. (2001). Testing and adjusting for publication bias. *Trends in Ecology & Evolution*, *16*, 580–586. https://doi.org/10.1016/S0169-5347(01)02235-2

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, *9*, 1–17. https://doi.org/10.1186/1471-2288-9-2

Nakagawa, S., & Lagisz, M. (2016). Visualizing unbiased and biased unweighted meta-analyses. *Journal of Evolutionary Biology*, *29*, 1914–1916. https://doi.org/10.1111/jeb.12945

Nakagawa, S., Noble, D. W., Senior, A. M., & Lagisz, M. (2017). Meta-evaluation of meta-analysis: Ten appraisal questions for biologists. *BMC Biology*, *15*, 18. https://doi.org/10.1186/s12915-017-0357-7

Nakagawa, S., O'Dea, R. E., Yang, Y., Sánchez-Tójar, A., & Noble, D. W. A. (2021). itchyshin/publication_bias: First release (v1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.5504537

Nakagawa, S., & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, *26*, 1253–1274. https://doi.org/10.1007/s10682-012-9555-5

Nakagawa, S., Senior, A. M., Viechtbauer, W., & Noble, D. W. A. (2021). An assessment of statistical methods for non-independent data in ecological meta-analyses: Comment. *Ecology*. https://doi.org/10.1002/ecy.3490

Noble, D. W. A., Lagisz, M., O'Dea, R. E., & Nakagawa, S. (2017). Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. *Molecular Ecology*, *26*, 2410–2425. https://doi.org/10.1111/mec.14031

Nobre, J. S., & Singer, J. D. (2007). Residual analysis for linear mixed models. *Biometrical Journal*, *49*, 863–875. https://doi.org/10.1002/bimj.200610341

Owrin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, *8*, 157–159.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA – Journal of the American Medical Association*, *295*, 676–680. https://doi.org/10.1001/jama.295.6.676

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*, 4544–4562. https://doi.org/10.1002/sim.2889

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*, 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

Preston, C., Ashby, D., & Smyth, R. (2004). Adjusting for publication bias: Modelling the selection process. *Journal of Evaluation in Clinical Practice*, *10*, 313–322. https://doi.org/10.1111/j.1365-2753.2003.00457.x

Pustejovskyei, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*, 57–71. https://doi.org/10.1002/jrsm.1332

Roberts, C. J., & Stanley, T. D. (2005). *Meta-regression analysis: Issues of publicaiton bias in economics*. Blackwell.

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*, 141–160.

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*, 464–468. https://doi.org/10.1111/j.0014-3820.2005.tb01004.x

Rosenthal, R. (1979). The 'file drawer problem' and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley.

Rucker, G., Schwarzer, G., & Carpenter, J. (2008). Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine*, *27*, 746–763. https://doi.org/10.1002/sim.2971

Sánchez-Tójar, A., Nakagawa, S., Sanchez-Fortun, M., Martin, D. A., Ramani, S., Girndt, A., Bokony, V., Kempenaers, B., Liker, A., Westneat, D. F., Burke, T., & Schroeder, J. (2018). Meta-analysis challenges a textbook

example of status signalling and demonstrates publication bias. *eLife*, 7, e37385. https://doi.org/10.7554/eLife.37385

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R*.

Senior, A. M., Grueber, C. E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E. S. A., & Nakagawa, S. (2016). Heterogeneity in ecological and evolutionary meta-analyses: Its magnitude and implications. *Ecology*, 97, 3293–3299. https://doi.org/10.1002/ecy.1591

Senior, A. M., Viechtbauer, W., & Nakagawa, S. (2020). Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Research Synthesis Methods*, 11, 553–567. https://doi.org/10.1002/jrsm.1423

Shi, L. Y., & Lin, L. F. (2019). The trim-and-fill method for publication bias: Practical guidelines and recommendations based on a large database of meta-analyses. *Medicine*, 98, e15987. https://doi.org/10.1097/MD.0000000000015987

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology-General*, 143, 534–547. https://doi.org/10.1037/a0033242

Song, C., Peacor, S. D., Osenberg, C. W., & Bence, J. R. (2020). An assessment of statistical methods for nonindependent data in ecological meta-analyses. *Ecology*, 101, e03184. https://doi.org/10.1002/ecy.3184

Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys*, 19, 309–345. https://doi.org/10.1111/j.0950-0804.2005.00250.x

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8, 581–591. https://doi.org/10.1177/1948550617693062

Stanley, T. D., & Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*. Routledge.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5, 60–78. https://doi.org/10.1002/jrsm.1095

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, 36, 1580–1598. https://doi.org/10.1002/sim.7228

Stanley, T. D., & Jarrell, S. B. (1998). Gender wage discrimination bias? A meta-regression analysis. *Journal of Human Resources*, 33, 947–973. https://doi.org/10.2307/146404

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75–98). Wiley.

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal*, 323, 101–105. https://doi.org/10.1136/bmj.323.7304.101

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ – British Medical Journal*, 343, d4002. https://doi.org/10.1136/bmj.d4002

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 435–452). Russell Sage Foundation.

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology*, 58, 894–901. https://doi.org/10.1016/j.jclinepi.2005.01.006

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, 18, 2693–2708. https://doi.org/10.1002/(SICI)1097-0258(19991030)18:20<2693:AID-SIM235>3.0.CO;2-V

Trkalinos, T. A., & Ioannidis, J. P. A. (2005). Assessing the evolution of effect sizes over time. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 241–259). Wiley.

van Aert, R. C. M., & van Assen, M. A. L. M. (2021). Correcting for publication bias in a meta-analysis with the p-uniform* Method. *OSF preprint*, https://osf.io/ebq6m/

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, 11, 713–729.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20, 293–309. https://doi.org/10.1037/met0000025

Vevea, J. L., Coburn, K., & Sutton, A. J. (2019). Publication bias. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 383–429). Russell Sage Foundation.

Vevea, J. L., & Hedges, L. V. (1995). A general linear-model for estimating effect size in the presence of publication bias. *Psychometrika*, 60, 419–435. https://doi.org/10.1007/BF02294384

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

Weinhandl, E. D., & Duval, S. (2012). Generalization of trim and fill for application in meta-regression. *Research Synthesis Methods*, 3, 51–67. https://doi.org/10.1002/jrsm.1042

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer.

Zwetsloot, P. P., Van der Naald, M., Sena, E. S., Howells, D. W., IntHout, J., De Groot, J. A. H., Chamuleau, S. A. J., MacLeod, M. R., & Wevers, K. E. (2017). Standardized mean differences cause funnel plot distortion in publication bias assessments. *eLife*, 6, e24260. https://doi.org/10.7554/eLife.24260

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.