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 2052, 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, Surrey, TW20 0EX, U.K.

[4] Department of Biology, Whitman College, Walla Walla, WA 99362, U.S.A

[5] Department of Evolutionary Biology, Bielefeld University, Bielefeld, 33615, Germany

* Correspondence: S. Nakagawa

e-mail: 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; Michael D. Jennions: 0000-0001-9221-2788; Julia Koricheva: 0000-0002-9033-0171; Daniel W. A. Noble: 0000-0001-9460-8743; Timothy H. Parker: 0000-0003-2995-5284; Alfredo Sánchez-Tójar: 0000-0002-2886-0649; Yefeng Yang: 0000-0002-8610-4016; Rose E. O'Dea: 0000-0001-8177-5075

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 ten 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 2 error rates are impacted by complex data structures. Still, limitations of these methods do not justify ignoring publication bias in ecological and evolutionary meta-analyses.

KEYWORDS: Outcome reporting bias, *p*-hacking, multilevel meta-analysis, selection bias, radial plot, effective sample size, time-lag bias, decline effect

1 | INTRODUCTION

Evidence from meta-analyses often drives future research, and sometimes leads to changes in policy and practice (Nakagawa *et al.*, 2017; Gurevitch *et al.*, 2018). Therefore, it is essential for meta-analytic evidence to minimise 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 Moller & Jennions, 2001; Jennions *et al.*, 2013; Marks-Anglin *et al.*, 2021). In this article, the following two types are most relevant: 1) outcome reporting bias, where selective reporting occurs within published studies (Marks-Anglin & Chen, 2020; 2021); and 2) time-lag bias, where positive results are published earlier than negative results (Trkalinos & Ioannidis, 2005; Koricheva, Jennions & Lau, 2013; Koricheva & Kulinskaya, 2019). 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 categorised 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, Sutton & Borenstein, 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

74 awareness of the importance of publication bias tests in meta-analysis in ecology and evolution.
75 (Koricheva & Gurevitch, 2014).

76 Two features common to meta-analytic datasets in ecology and evolution pose problems for
77 publication bias tests: high levels of heterogeneity and non-independence. Importantly, many
78 currently available tests for publication bias fail when there are high levels of heterogeneity (e.g.,
79 Macaskill, Walter & Irwig, 2001; Sterne, Egger & Smith, 2001; Moreno *et al.*, 2009). Furthermore,
80 Nakagawa and Santos (2012) noted that, at the time, there were no statistical methods to test for
81 publication bias that could explicitly account for non-independent effect sizes. Highly
82 heterogeneous data are common in ecology and evolutionary biology, as research questions often
83 span many types of ecosystems and species. Non-independence is pervasive because many studies
84 produce multiple effect sizes and, if a meta-analytic dataset includes multiple species, then effect
85 sizes might also be correlated due to phylogenetic relatedness (Noble *et al.*, 2017). Therefore, for a
86 publication bias test to be useful in ecology and evolution, it would need to adequately handle both
87 heterogeneity and non-independence (cf. Fernandez-Castilla *et al.*, 2019; Rodgers & Pustejovsky,
88 2020).

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

2 | KEY STATSTICAL CONCEPTS

2.1 | Sampling variance, standard error, precision and weight

Three types of standardised effect size statistics are most commonly used in meta-analyses in ecology and evolutionary biology (Nakagawa & Santos, 2012; Koricheva & Gurevitch, 2014). The first effect size 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 & Olkin, 1985):

$$\mathrm{SMD}_i = \frac{\overline{\mathrm{X}}_{2i} - \overline{\mathrm{X}}_{1i}}{\sqrt{\frac{(n_{1i}-1)\mathrm{SD}_{1i}^2 + (n_{2i}-1)\mathrm{SD}_{2i}^2}{n_{1i}+n_{2i}-2}}}, \qquad (1)$$

$$\mathrm{Var}(\mathrm{SMD}_i) = \frac{n_{1i}+n_{2i}}{n_{1i}n_{2i}} + \frac{\mathrm{SMD}_i^2}{2(n_{1i}+n_{2i})}, \qquad (2)$$

where the *i*th effect size (SMD) and sampling variance (Var) are a function of the means ($\overline{\mathrm{X}}$), standard deviations (SD of sample) and sample size (*n*) of the two groups (1 and 2); Equation 1 and 2 often include a small sample-size correction factor denoted as *J* (see Borenstein *et al.*, 2009). Second, the logarithm of response ratio (Hedges, Gurevitch & Curtis, 1999; also known as the ratio of means; Friedrich, Adhikari & Beyene, 2008) can be written as:

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

$$\mathrm{Var}(\mathrm{lnRR}_i) = \frac{\mathrm{SD}_{1i}^2}{n_{1i}\overline{\mathrm{X}}_{1i}^2} + \frac{\mathrm{SD}_{2i}^2}{n_{2i}\overline{\mathrm{X}}_{2i}^2}, \qquad (4)$$

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

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

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

where n_i is the ith sample size used to obtain the correlation coefficient, r_i. Incidentally, the variance of the correlation coefficient is: $\mathrm{Var}(r_i) = (1 - r_i^2)^2/(n - 1)$, 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).

Sampling variance is at the heart of meta-analysis as this quantity, which is always a function of sample size, indicates (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; for example, a point estimate with high certainty has low standard error and variance, but high precision and weight (Figure 1).

2.2 | Heterogeneity

Ecologists and evolutionary biologists predominately use a 'random-effects model' of meta-analysis rather than a 'fixed-effect model' (Nakagawa & Santos, 2012; Koricheva & Gurevitch, 2014). 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:

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

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

where 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 (note that $i = 1, 2, \ldots, N_{effect\text{-}size}$, the number of effect sizes; when $N_{effect\text{-}size} = N_{study}$, 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, Sutton & Borenstein, 2005; Vevea, Coburn & Sutton, 2019; Marks-Anglin & Chen, 2020; Marks-Anglin *et al.*, 2021). Therefore, we summarise 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 Supporting Information, Appendix S1, found at https://github.com/itchyshin/publication_bias).

Following Sutton (2009) (see also Vevea, Coburn & Sutton, 2019), we categorise publication bias tests into two types: 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). 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 (Sterne, Becker & Egger, 2005; Sutton, 2009; Vevea, Coburn & Sutton, 2019; Marks-Anglin & Chen, 2020), 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, Sutton & Borenstein, 2005; Marks-Anglin & Chen, 2020).

The original funnel plot used sample size as the measure of uncertainty (Light & Pillemer, 1984; Figure 3a). Yet, more recent recommendations are to use either SE, precision, variance or the inverse of variance (Figure 1; Sterne, Becker & Egger, 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 x 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, and colleagues (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; see also Terrin, Schmid & Lau, 2005). For instance, heterogeneity among effect sizes can create asymmetries of many kinds (Figure 3b); the other potential sources of asymmetry are data irregularities (e.g., mistakes, frauds, unique observations; cf. Nakagawa & Lagisz, 2016), artefacts (see Section 4.3), and chance (Egger *et al.*, 1997). 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, Becker & Egger, 2005). 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 SEs (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:

$$z_i = \beta_0 + \beta_1 prec_i + e_i, \qquad (8)$$

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

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 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):

$$y_i = \beta_0 + \beta_1 se_i + \epsilon_i, \qquad (9)$$

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

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 Supplementary Information, 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:

$$y_i = \beta_0 + \beta_1 se_i + s_i + m_i, \qquad (10)$$

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

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

According to simulation studies (Macaskill, Walter & Irwig, 2001; Sterne, Egger & Smith, 2001; Moreno *et al.*, 2009), 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: 1) add moderators to absorb some heterogeneity, and 2) 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). 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, Walter & Irwig, 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.*, 2019). 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; Sanchez-Tojar *et al.*, 2018). To test for time-lag bias, we caution against using correlation-based methods, because this approach does not account for different precisions of effect sizes (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): 1) using a cumulative meta-analysis, and 2) using a regression-based method (see Trkalinos & Ioannidis, 2005; Koricheva, Jennions & Lau, 2013; Koricheva & Kulinskaya, 2019).

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):

$$y_i = \beta_0 + \beta_1 year_i + s_i + m_i, \qquad (11)$$

where $year_i$ is the publication year for the ith 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 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 non-significant unpublished results needed to exist to make the overall effect non-significant (e.g., Rosenthal, 1979; Rosenberg, 2005) or negligible in magnitude (e.g., Owrin, 1983). If the fail-safe *N* is large ($>5N_{study} + 10$), the results of analyses may be considered to be robust with respect to publication bias as such large number of 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:

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

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 Orwin (1983) relies on the magnitude of the effect size rather than statistical significance; one version of this method can be written as:

$$N_{Orwin} = \frac{N_{study}(\overline{y} - y_n)}{y_n}, \qquad (13)$$

where $\overline{y}$ 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:

$$N_{Rosenberg} = \frac{N_{study} W}{\sum_{i=1}^{N_{study}} w_i}, \qquad (14)$$

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

where w_i is the inverse of sampling variance ($1/v_i$; note that w_i can be modified for a random-effects model) 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 the 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 depends on the exact method, 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); its example dataset shows $N_{Rosenthal}$ = 598, N_{Orwin} = 84, and $N_{Rosenberg}$ = 370 (for details, 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, none of the methods of fail-safe *N* are inferential.

3.2.2 | Trim-and-fill tests

The trim-and-fill test provides a non-parametric method that can visualize potentially missing data, and statistically both detect and correct for funnel asymmetry (Duval & Tweedie, 2000b; Duval & Tweedie, 2000a). 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 not rare 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, $N_{missing}$) 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 (Peters *et al.*, 2007; Moreno *et al.*, 2009). Although trim-and-fill tests have been extended for meta-regressions (Weinhandl & Duval, 2012), this implementation of this extension is currently limited to one moderator. Further, recent simulation work by Rogers and Pustejovksy (2020) 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, Nelson & Simonsohn, 2011) and '*p*-hacking' (Simonsohn, Nelson & Simmons, 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, van Aert & Wicherts, 2015). Interestingly, McShane et al. (2016) has 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, Wicherts & van Assen, 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 Rothstein, Sutton & Borenstein, 2005; Vevea, Coburn & Sutton, 2019; Marks-Anglin & Chen, 2020). There are probably as many selection models as all other methods combined (Marks-Anglin & Chen, 2020), but 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 (e.g., Preston, Ashby & Smyth, 2004; Carter *et al.*, 2019; Rodgers & Pustejovsky, 2020;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.

391 **4.1 | A multilevel meta-analysis and funnel plots**

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

393 $$y_i = \beta_0 + s_j + u_i + m_i, \qquad (15)$$

394 $$s_j \sim \mathcal{N}(0, \sigma_s^2), u_i \sim \mathcal{N}(0, \sigma_u^2), m_i \sim \mathcal{N}(0, v_i),$$

395 where β_0 is the overall estimate (or meta-analytic mean); s_j is the between-study effect for the *j*th
396 study, normally distributed with the variance of σ_s^2; u_i is the between-effect-size effect, or within-
397 study effect, for the *i*th effect size, distributed with a mean of zero and the variance of σ_u^2; and m_i is
398 as in Equation 7 (but note that j = 1, 2, …, N_{study}, the number of studies, and i = 1, 2, …, $N_{effect\text{-}size}$,
399 the number of effect sizes; $N_{effect\text{-}size} > N_{study}$). Equation 15 explicitly models multiple effect sizes per
400 study. Also, in Equation 7, the term σ_s^2 is the only source of heterogeneity, while in Equation 15,
401 both σ_s^2 and σ_u^2 are each contributing to heterogeneity among effect sizes.

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

404 $$y_i = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i} + s_j + u_i + m_i, \qquad (16)$$

405 where β_1 is the slope for x_1, a study-level moderator (characteristics of different studies, *j*; e.g.,
406 experimental *vs.* observational) and β_2 is the slope for x_2, an effect-size-level moderator
407 (characteristics of effect sizes, *i*; different measurements or sexes). We have mentioned that we can
408 draw a funnel plot with residuals rather than the observed effect sizes (Figure 6a). A complication is
409 that, given Equation 15, we can extract at least 3 different residuals, which are:

410 $$resid_{mi} = y_i - (\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i}), \qquad (17)$$

411 $$resid_{c1i} = y_i - (\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i} + s_j), \qquad (18)$$

412 $$resid_{c2i} = y_i - (\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i} + s_j + u_i), \qquad (19)$$

413 where $resid_m$ represents marginal residuals (subtracting only fixed effects from the observations;
414 Figure 6b), whereas $resid_{c1}$ and $resid_{c2}$ are conditional residuals (Figure 6c & d; Nobre & Singer,
415 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 Figures 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 $N_{effect\text{-}size} = 3$ and the first two effect sizes' sampling variance are correlated, then we can write $\mathbf{M}$ as:

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

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, 2020).

441 **4.2 | Multilevel meta-regression and Egger's regression**

442 As an alternative to using residual analysis, we can directly model sampling SE in Equation 15 (cf.
443 Equation 10; Fernandez-Castilla *et al.*, 2019; Rodgers & Pustejovsky, 2020):

444 $$y_i = \beta_0 + \beta_1 se_i + s_j + u_i + m_i. \quad (21)$$

445 By examining Equation 21, we may realise that β_0 represents a conditional estimate of an overall
446 effect when SE is 0, which means, theoretically, there is no uncertainty (Figure 5e). Then, does β_0
447 provide an adjusted estimate of an overall effect, when β_1 is statistically significant (i.e., detecting a
448 small-study effect)? This question has been examined by Stanley and Doucouliagos (2012; 2014).
449 They have shown that, with significant β_1, β_0 provides an adjusted estimate that is downwardly
450 biased, when a true positive or a null effect exists (which is illustrated in Figure 5e; note that they
451 state that with non-statistically significant β_1, β_0 provides the best estimate of an adjusted mean). If
452 the slope of SE (β_1) is statistically significant then fitting sampling variance instead of SE is
453 recommended according to the following equation:

454 $$y_i = \beta_0 + \beta_1 v_i + s_j + u_i + m_i. \quad (22)$$

455 This is equivalent to fitting se_i^2, which is a quadratic term. Stanley and Doucouliagos (2012; 2014)
456 have shown that β_0 in Equation 22 is still downwardly biased, but much less so, although Equation
457 21 is more powerful (i.e. an adjustment tends to underestimate) when there is a positive (or no)
458 effect (cf. Figure 5f). While this two-step approach may seem simplistic (see also Stanley, 2017;
459 Stanley, Doucouliagos & Ioannidis, 2017), it provides an easy-to-implement publication bias test
460 which explicitly models non-independent data.

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

463 $$y_i = \beta_0 + \beta_1 year_j + s_j + u_i + m_i. \quad (23)$$

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

$$y_i = \beta_0 + \beta_1 se_i + \beta_2 c(year_j) + \sum_{k=3}^{N_{mod}} \beta_k x_k + s_j + u_i + m_i, \qquad (24)$$

where β_k is the slope for the *k*th moderator (*k* = 3, 4,…, N_{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 Supporting Information, Appnedix S4). 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, Walter & Irwig, 2001; Deeks, Macaskill & Irwig, 2005). This is especially true under two scenarios: 1) when there is a (mathematical) relationship between effect size and sampling SE not due to publication bias, and 2) when SE is not estimated accurately.

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 inaccurately, we now go back to comparing sampling variance among the three commonly used effect sizes (Equations 2, 4 and 6). The SMD's variance has the square of the point estimate (i.e. SMD; Equation 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, $\overline{X}$ (sample mean) and especially SD (sample standard deviation) will be poorly estimated, resulting 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). 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); this approach is known as the funnel plot test (Macaskill, Walter & Irwig, 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

(Macaskill, Walter & Irwig, 2001; Deeks, Macaskill & Irwig, 2005), and non-independent effect sizes (Fernandez-Castilla *et al.*, 2019).

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' ($4\tilde{n}_i$) because it accounts for unbalanced sampling. The effective sample size is given by (Bakbergenuly, Hoaglin & Kulinskaya, 2020b; 2020a; also see; Deeks, Macaskill & Irwig, 2005; Bakbergenuly, Hoaglin & Kulinskaya, 2020c):

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

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 & 6):

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

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, Macaskill & Irwig, 2005):

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

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

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 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 *et al.*, 2017).

In practice, multilevel meta-analytic models are often more complex. 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:

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

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

where a_h is the phylogenetic effect for the *h*th 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 *h*th 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 (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. **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):

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

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 (2020) showed that when averaging effect sizes within studies, 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 & Hedges, 1995) had the appropriate level of Type 1 error, although the three-parameter selection model was noticeably more powerful 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. **A** in Equation 29). 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

559 average categorical moderators such as measurement types, evaluation methods or sex). Although
560 iteratively sampling one effect size per study could capture moderating effects, this approach also
561 reduces the information content of the dataset. Despite these limitations, under some circumstances,
562 averaging and sampling could be useful (examples and implementations for the trim-and-fill test
563 and a selection model in Supporting Information, Appendix S5).

564 **5 | CONCLUSIONS**

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

573 Few simulation studies exist explicitly investigating the performance of publication bias
574 tests with non-independent data. Two studies that we are aware of supported similar models to the
575 multilevel-regression method we proposed here (Fernandez-Castilla *et al.*, 2019; Rodgers &
576 Pustejovsky, 2020). In addition, a general point to take from these two simulation studies is that
577 most methods are prone to Type 2 error, with a possible exception of some selection models, even
578 when the methods have nominal Type 1 error rates. Therefore, not detecting publication bias in a
579 publication bias test should not be taken as a proof of no publication bias, including for multilevel
580 regression. Clearly, we need more methodological and simulation-based work in the future.

581 Finally, we repeat that the results of publication bias tests should always be cautiously
582 interpreted because no methods will ever be able to verify the actual number of missing effect sizes.
583 By way of example, a recent study compared the results of 15 meta-analyses and pre-registered
584 replication projects on the same topics (Kvarven, Stromland & Johannesson, 2020). 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 two-step regression model (the method by Stanley & Doucouliagos, 2012; 2014). 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 try to identify related biases.

ACKNOWLEDGEMENTS

We are grateful for Wolfgang Viechtbauer who help SN arrive at the multilevel-regression method described in this article. 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.

AUTHORS' CONTRIBUTIONS

Conceptualization: SN & REO; Data curation REO, AST, & YY; Formal Analysis, REO, AST, YY & SN; Validation: REO, DWAN & SN; Investigation: SN, ML, MDJ, JK, DWAN, THP, & REO; Visualization: SN, ML & REO; Methodology: SN; Writing – original draft: SN; Project administration: SN & REO; Writing - review & editing: all authors. We note that the supplementary information (Appendices S1-S5) was put together by REO, AST, YY & SN.

DATA AVAILABILITY

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

607 **REFERENCES**

608 Bakbergenuly, I., Hoaglin, D.C. & Kulinskaya, E. (2020a) Estimation in meta-analyses of mean
609 difference and standardized mean difference. *Statistics in Medicine,* **39,** 171-191.

610 Bakbergenuly, I., Hoaglin, D.C. & Kulinskaya, E. (2020b) Estimation in meta-analyses of response
611 ratios. *Bmc Medical Research Methodology,* **20**.

612 Bakbergenuly, I., Hoaglin, D.C. & Kulinskaya, E. (2020c) Methods for estimating between-study
613 variance and overall effect in meta-analysis of odds ratios. *Research Synthesis Methods,* **11,**
614 426-442.

615 Barto, E.K. & Rillig, M.C. (2012) Dissemination biases in ecology: effect sizes matter more than
616 quality. *Oikos,* **121,** 228-235.

617 Becker, B.J. (2005) Fail safe *N* or file-drawer number. *Publication bias in meta-analysis :*
618 *prevention, assessment and adjustments* (eds H. Rothstein, A.J. Sutton & M. Borenstein),
619 pp. 111-125. John Wiley, Chichester.

620 Begg, C.B. & Mazumdar, M. (1994) Operating characteristics of a rank correlation test for
621 publication bias. *Biometrics,* **50,** 1088-1101.

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

624 Carter, E.C., Schönbrodt, F.D., Gervais, W.M. & Hilgard, J. (2019) Correcting for Bias in
625 Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices*
626 *in Psychological Science,* **2,** 115-144.

627 Citkowicz, M. & Vevea, J.L. (2017) A Parsimonious Weight Function for Modeling Publication
628 Bias. *Psychological Methods,* **22,** 28-41.

629 Deeks, J.J., Macaskill, P. & Irwig, L. (2005) The performance of tests of publication bias and other
630 sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal*
631 *of Clinical Epidemiology,* **58,** 882-893.

632 Doleman, B., Freeman, S.C., Lund, J.N., Williams, J.P. & Sutton, A.J. (2020) Funnel plots may
633 show asymmetry in the absence of publication bias with continuous outcomes dependent on
634 baseline risk: presentation of a new publication bias test. *Research Synthesis Methods,* **11,**
635 522-534.
636 Duval, S. (2005) The trim and fill method. *Publication bias in meta-analysis: prevention,*
637 *assessment and adjustments* (eds H. Rothstein, A.J. Sutton & M. Borenstein), pp. 127-144.
638 John Wiley, Chichester.
639 Duval, S. & Tweedie, R. (2000a) A nonparametric "trim and fill" method of accounting for
640 publication bias in meta-analysis. *Journal of the American Statistical Association,* **95,** 89-98.
641 Duval, S. & Tweedie, R. (2000b) Trim and fill: A simple funnel-plot-based method of testing and
642 adjusting for publication bias in meta-analysis. *Biometrics,* **56,** 455-463.
643 Egger, M., Smith, G.D., Schneider, M. & Minder, C. (1997) Bias in meta-analysis detected by a
644 simple, graphical test. *British Medical Journal,* **315,** 629-634.
645 Fernandez-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S.N., Onghena, P. & Van den
646 Noortgate, W. (2019) Detecting selection bias in meta-analyses with multiple Outcomes: a
647 simulation study. *Journal of Experimental Education*.
648 Freckleton, R.P. (2002) On the misuse of residuals in ecology: regression of residuals vs. multiple
649 regression. *Journal of Animal Ecology,* **71,** 542-545.
650 Friedrich, J.O., Adhikari, N.K.J. & Beyene, J. (2008) The ratio of means method as an alternative to
651 mean differences for analyzing continuous outcome variables in meta-analysis: A simulation
652 study. *Bmc Medical Research Methodology,* **8**.
653 Galbraith, R.F. (1988) A note on graphical presentation of estimated odds ratios from several
654 clinical-trials. *Statistics in Medicine,* **7,** 889-894.
655 Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. (2018) Meta-analysis and the science of
656 research synthesis. *Nature,* **555,** 175-182.

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.

Harrer, M., Cuijpers, P., Furukawa, T. & Ebert, D.D. (2019) *Doing meta-analysis in R: a hands-on guide*.

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

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.

Hedges, L.V., Gurevitch, J. & Curtis, P.S. (1999) The meta-Analysis of response ratios in experimental ecology. *Ecology,* **80,** 1150-1156.

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.

Higgins, J.P.T. & Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine,* **21,** 1539-1558.

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

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.

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

Koricheva, J., Jennions, M.D. & Lau, J. (2013) Temporal trends in effect sizes: causes, detection and implications. *The handbook of meta-analysis in ecology and evolution* (eds J.

682 Koricheva, J. Gurevitch & K. Mengersen), pp. 237-254. Princeton University Press,
683 Princeton.
684 Koricheva, J. & Kulinskaya, E. (2019) Temporal instability of evidence base: a threat to policy
685 making? *Trends in Ecology & Evolution,* **34,** 895-902.
686 Kossmeier, M., Tran, U.S. & Voracek, M. (2020) Power-Enhanced Funnel Plots for Meta-Analysis
687 The Sunset Funnel Plot. *Zeitschrift Fur Psychologie-Journal of Psychology,* **228,** 43-49.
688 Kvarven, A., Stromland, E. & Johannesson, M. (2020) Comparing meta-analyses and preregistered
689 multiple-laboratory replication projects. *Nature Human Behaviour,* **4,** 423-+.
690 Lajeunesse, M.J. (2015) Bias and correction for the log response ratio in ecological meta-analysis.
691 *Ecology,* **96,** 2056-2063.
692 Light, R.J. & Pillemer, D.B. (1984) *Summing up : the science of reviewing research*. Harvard
693 University Press, Cambridge, Mass. ; London.
694 Macaskill, P., Walter, S.D. & Irwig, L. (2001) A comparison of methods to detect publication bias
695 in meta-analysis. *Statistics in Medicine,* **20,** 641-654.
696 Marks-Anglin, A. & Chen, Y. (2020) A historical review of publication bias. *Research Synthesis*
697 *Methods,* **11,** 725-742.
698 Marks-Anglin, A., Duan, R., Chen, Y., Panagiotou, O. & Schmid, C.H. (2021) Publication and
699 outcome reporting bias. *Handbook of meta-analysis* (eds C.H. Schmid, T. Stijnen & R.W.
700 White), pp. 283-312. CRC, Boca Raton.
701 McShane, B.B., Bockenholt, U. & Hansen, K.T. (2016) Adjusting for Publication Bias in Meta-
702 Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on*
703 *Psychological Science,* **11,** 730-749.
704 Moller, A.P. & Jennions, M.D. (2001) Testing and adjusting for publication bias. *Trends in Ecology*
705 *& Evolution,* **16,** 580-586.

706 Moreno, S.G., Sutton, A.J., Ades, A.E., Stanley, T.D., Abrams, K.R., Peters, J.L. & Cooper, N.J.
707 (2009) Assessment of regression-based methods to adjust for publication bias through a
708 comprehensive simulation study. *Bmc Medical Research Methodology,* **9**.
709 Nakagawa, S. & Lagisz, M. (2016) Visualizing unbiased and biased unweighted meta-analyses.
710 *Journal of Evolutionary Biology,* **29,** 1914-1916.
711 Nakagawa, S., Noble, D.W., Senior, A.M. & Lagisz, M. (2017) Meta-evaluation of meta-analysis:
712 ten appraisal questions for biologists. *BMC Biology,* **15,** 18.
713 Nakagawa, S. & Santos, E.S.A. (2012) Methodological issues and advances in biological meta-
714 analysis. *Evolutionary Ecology,* **26,** 1253-1274.
715 Noble, D.W.A., Lagisz, M., O'Dea R, E. & Nakagawa, S. (2017) Nonindependence and sensitivity
716 analyses in ecological and evolutionary meta-analyses. *Molecular Ecology,* **26,** 2410-2425.
717 Nobre, J.S. & Singer, J.D. (2007) Residual analysis for linear mixed models. *Biometrical Journal,*
718 **49,** 863-875.
719 Owrin, R.G. (1983) A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics,*
720 **8,** 157-159.
721 Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R. & Rushton, L. (2007) Performance of the trim
722 and fill method in the presence of publication bias and between-study heterogeneity.
723 *Statistics in Medicine,* **26,** 4544-4562.
724 Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R. & Rushton, L. (2008) Contour-enhanced meta-
725 analysis funnel plots help distinguish publication bias from other causes of asymmetry.
726 *Journal of Clinical Epidemiology,* **61,** 991-996.
727 Preston, C., Ashby, D. & Smyth, R. (2004) Adjusting for publication bias: modelling the selection
728 process. *Journal of Evaluation in Clinical Practice,* **10,** 313-322.
729 Roberts, C.J. & Stanley, T.D. (2005) *Meta-regression analysis: issues of publicaiton bias in*
730 *economics*. Blackwell, Malden, MA.

731 Rodgers, M.A. & Pustejovsky, J.E. (2020) Evaluating meta-analytic methods to detect selective
732 reporting in the presence of dependent effect sizes. *Psychological Methods*.
733 Rosenberg, M.S. (2005) The file-drawer problem revisited: A general weighted method for
734 calculating fail-safe numbers in meta-analysis. *Evolution,* **59,** 464-468.
735 Rosenthal, R. (1979) The "file drawer problem" and tolerance for null results. *Psychological*
736 *Bulletin,* **86,** 638-641.
737 Rothstein, H., Sutton, A.J. & Borenstein, M. (2005) Publication bias in meta-analysis: prevention,
738 assessment and adjustments. John Wiley, Chichester.
739 Sanchez-Tojar, A., Nakagawa, S., Sanchez-Fortun, M., Martin, D.A., Ramani, S., Girndt, A.,
740 Bokony, V., Kempenaers, B., Liker, A., Westneat, D.F., Burke, T. & Schroeder, J. (2018)
741 Meta-analysis challenges a textbook example of status signalling and demonstrates
742 publication bias. *Elife,* **7**.
743 Schwarzer, G., Carpenter, J.R. & Rücker, G. (2015) *Meta-analysis with R*.
744 Senior, A.M., Grueber, C.E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E.S.A. & Nakagawa, S.
745 (2016) Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and
746 implications. *Ecology,* **97,** 3293-3299.
747 Senior, A.M., Viechtbauer, W. & Nakagawa, S. (2020) Revisiting and expanding the meta-analysis
748 of variation: The log coefficient of variation ratio. *Research Synthesis Methods,* **11,** 553-
749 567.
750 Shi, L.Y. & Lin, L.F. (2019) The trim-and-fill method for publication bias: practical guidelines and
751 recommendations based on a large database of meta-analyses. *Medicine,* **98**.
752 Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011) False-positive psychology: undisclosed
753 flexibility in data collection and analysis allows presenting anything as significant.
754 *Psychological Science,* **22,** 1359-1366.
755 Simonsohn, U., Nelson, L.D. & Simmons, J.P. (2014) P-curve: a key to the file-drawer. *Journal of*
756 *Experimental Psychology-General,* **143,** 534-547.

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.

Stanley, T.D. (2017) Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science,* **8,** 581-591.

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

Stanley, T.D. & Doucouliagos, H. (2014) Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods,* **5,** 60-78.

Stanley, T.D., Doucouliagos, H. & Ioannidis, J.P.A. (2017) Finding the power to reduce publication bias. *Statistics in Medicine,* **36,** 1580-1598.

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.

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

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

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.

Thompson, S.G. & Sharp, S.J. (1999) Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine,* **18,** 2693-2708.

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

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.

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

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.

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.

Wickham, H. (2009) *ggplot2 : elegant graphics for data analysis*. Springer, New York ; London.

800

801 **FIGURE LEGENDS**

802 **FIGURE 1.** A schematic showing the relationship among standard error (SE), sampling variance,
803 precision (the inverse of SE) and weight (the inverse of variance). Note that the inverse of variance
804 is the weight for a fixed-effect model (the weight for a random-effect model is the inverse of the
805 sum of sampling variance and between-study variance). In the statistical literature, the inverse of
806 variance is also referred to as precision. Importantly, 'standard error' (SE) can be referred to as
807 'standard deviation' (SD), which is not incorrect because standard error is 'standard deviation of a
808 statistic' – not to be confounded with 'standard deviation of a sample'.

809

810 **FIGURE 2.** Frequencies of the usages of different publication bias tests in our survey of 102 meta-
811 analyses in ecology and evolution. Note that only one paper employed a method (a weighted
812 histogram) belonging to a category that was not pre-specified (including 'None reported'; the labels
813 for items A-K match the labels used in our survey). For the details of the survey, see Supporting
814 Information, Appendix S1.

815

816 **FIGURE 3.** Examples of funnel plots and a radial plot using the same dataset ($N_{effect\text{-}size} = N_{study} =$
817 100): a) a funnel plot with sample size as a measure of uncertainty; b) a funnel plot with precision
818 (1/SE) as a measure of uncertainty, red dots representing 'expected' missing data under publication
819 bias, and blue dots representing 'unexpected' missing data; c) a counter enhanced funnel plot with
820 SE as a measure of uncertainty; d) a sunset plot showing statistical power of data as the overall
821 effect estimate as a true effect; e) a residual funnel plot (one moderator removed); and f) a radial
822 plot. We used the *R* packages *metafor* (panels a-c & e; Viechtbauer, 2010), *metaviz* (panel d;
823 Kossmeier, Tran & Voracek, 2020) and *meta* (panel f; Schwarzer, Carpenter & Rücker, 2015) for
824 visualizations.

825

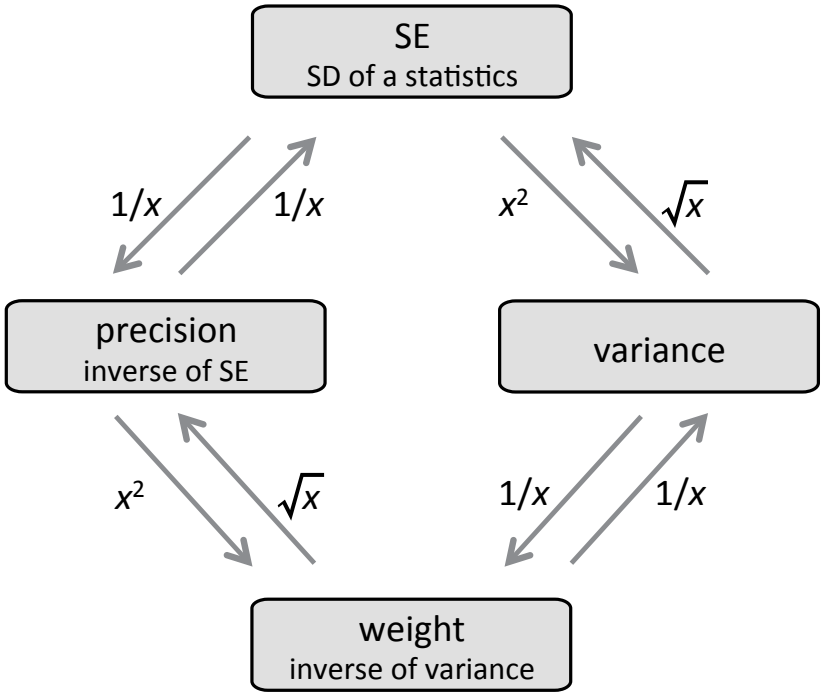
826 **FIGURE 4.** Examples of various plots (using the same dataset as Figure 3b minus 25 red
827 datapoints, therefore $N_{effect\text{-}size}$ = 75): a) a scatter plot with the height of the solid line representing
828 the degree of funnel asymmetry (cf. the radial plot at Figure 3 f); b) a scatter plot with the steepness
829 of the slope representing the degree of funnel asymmetry; c) a forest plot showing results of
830 cumulate meta-analyses, where only a portion of the dataset ($N_{effect\text{-}size}$ = 15) was used; d) a bubble
831 plot showing a 'decline effect' over time, where only a portion of the dataset ($N_{effect\text{-}size}$ = 15) was
832 used; e) a funnel plot with precision (1/SE) and with a trim-and-fill method filling missing data
833 (red circles; using the R_0 estimator); and f) the same as panel e but with SE as a measure of
834 uncertainty. We used the *R* packages *gglot2* (panels a, b & d; Wickham, 2009) and *metafor* (panel ;
835 Viechtbauer, 2010) for visualizations.
836
837 **FIGURE 5.** Example plots for *p*-curves and selection models (using the same dataset as in Figure
838 4; $N_{effect\text{-}size}$ = 75): a) a line plot showing the distribution of statistically significant *p* values under 3
839 scenarios: 1) with the observed *p* values (blue solid line), 2) when there is no effect (red dotted
840 line), and 3) when there is an effect (i.e. an observed overall effect as a true effect) with 33%
841 statistical power (note that if a blue line increases at the α level of 0.05, this is a sign of *p*-hacking;
842 for more details of this plot, see www.p-curve.com); b) a plot showing 4 different weight functions
843 that model, based on the data, the likelihood of effect sizes being selected for publication: 1) a half-
844 normal function based on *p* values (black solid line), 2) the same function but based both on *p*
845 values and precisions (black dotted line), 3) a logistic function based on *p* values (red solid line),
846 and 4) the same function but based both on *p* values and precisions (red dotted line; these functions
847 are based on Preston, Ashby & Smyth, 2004); and c) a plot showing two different 'step' weight
848 function based on: 1) three cut-points (α = 0.05, 0.1, 0.5) and 2) one cut-point (α = 0.05; this model
849 is sometimes referred to as a 3 parameter selection model, PSM with the 3 parameters being an
850 overall mean, the between-study variance, and an index determining the likelihood of selection;

e.g., Carter *et al.*, 2019; Rodgers & Pustejovsky, 2020). We used the *R* packages *dmetar* (panel a; Harrer *et al.*, 2019) and *metafor* (panel b & c; Viechtbauer, 2010) for visualizations.

FIGURE 6. Examples of funnel plots from a dataset with lnRR (N_{study} = 70; $N_{effect\text{-}size}$ = 271) and a different dataset with *Zr* (N_{study} = 48; $N_{effect\text{-}size}$ = 104): 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). 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, the same line as in panel 'e'). Both red lines showing to intersect the zero effect size at the intercept. We used the *R* packages *metafor* (panels a-d; Viechtbauer, 2010) and *ggplot2* (panels e-f; Wickham, 2009) for visualizations.

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 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.

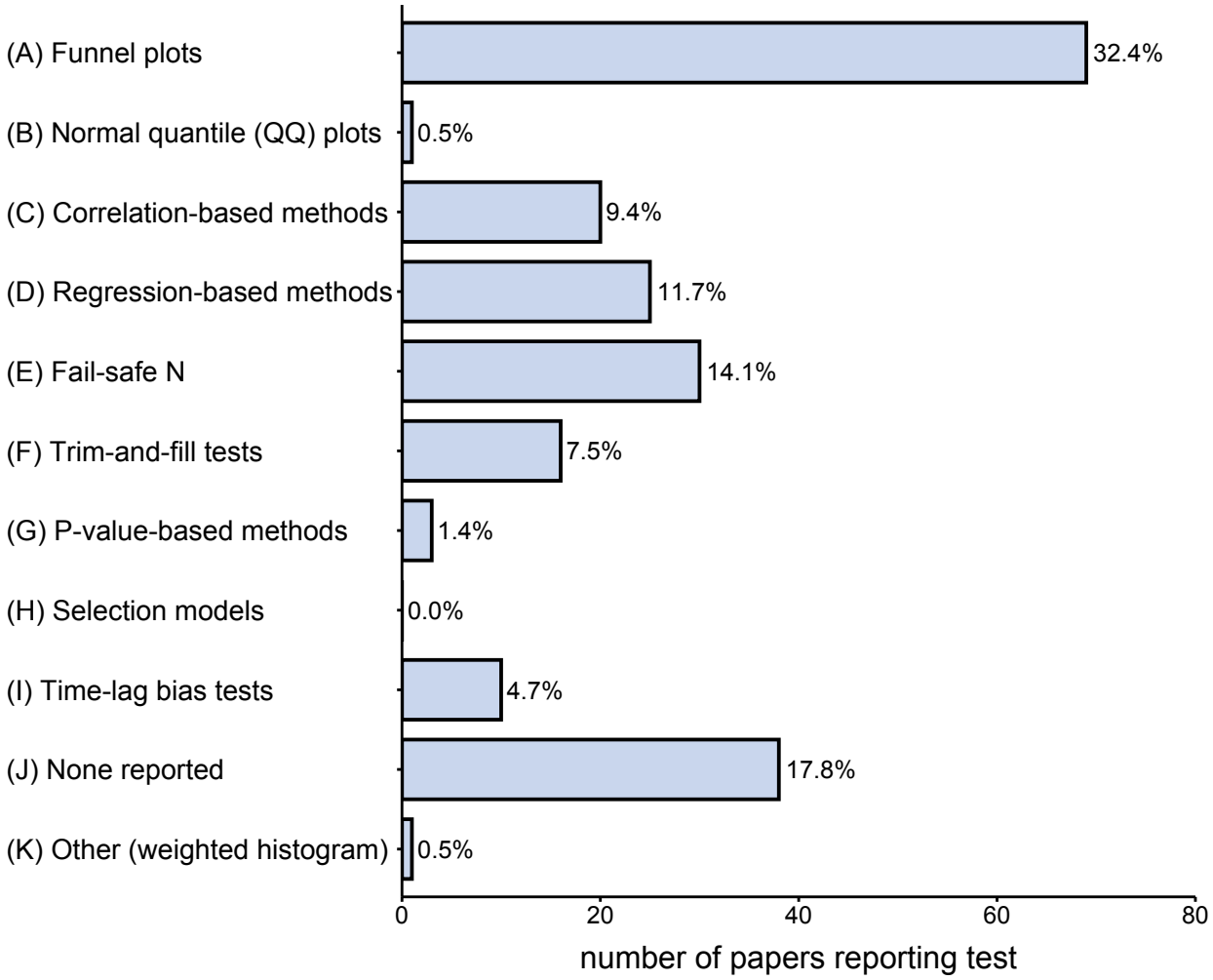
877 **FIGURE 1**



878

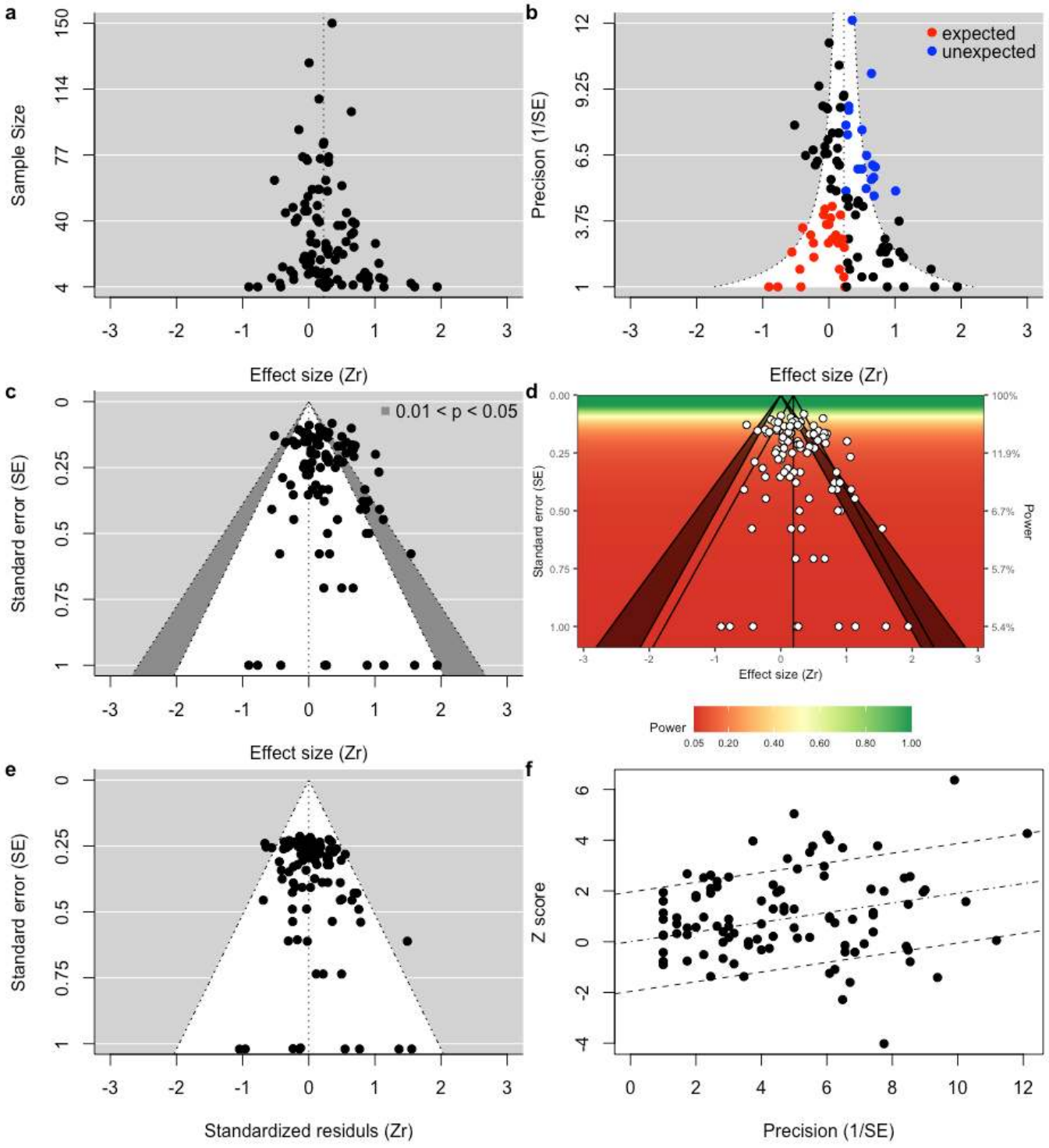
879

FIGURE 2


(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%
0
20
40
60
80
number of papers reporting test

885

886 **FIGURE 3**



887

888

FIGURE 4

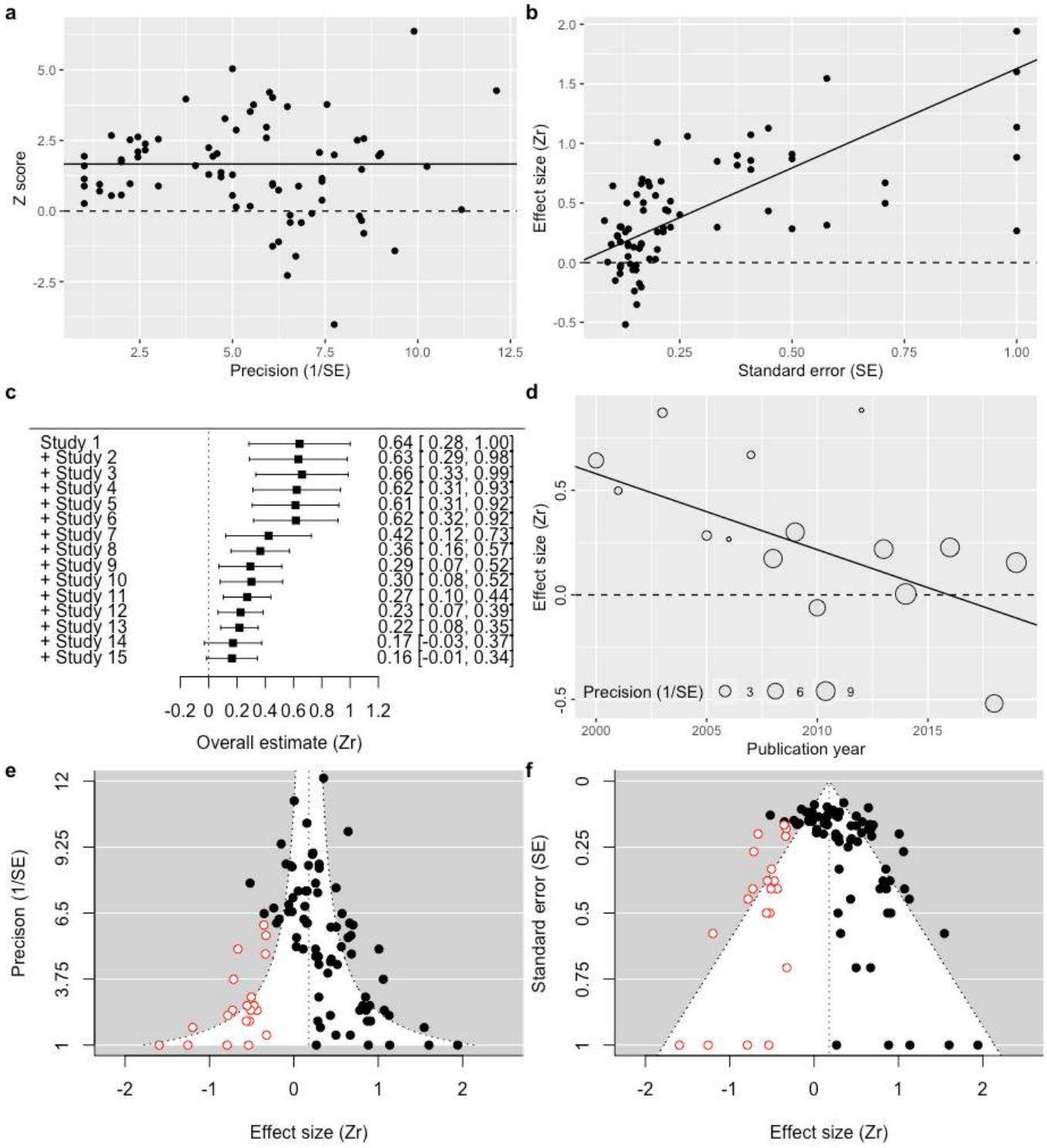

a
Z score
Precision (1/SE)
b
Effect size (Zr)
Standard error (SE)
c
Study 1 0.64 [0.28, 1.00]
+ Study 2 0.63 [0.29, 0.98]
+ Study 3 0.66 [0.33, 0.99]
+ Study 4 0.62 [0.31, 0.93]
+ Study 5 0.61 [0.31, 0.92]
+ Study 6 0.62 [0.32, 0.92]
+ Study 7 0.42 [0.12, 0.73]
+ Study 8 0.36 [0.16, 0.57]
+ Study 9 0.29 [0.07, 0.52]
+ Study 10 0.30 [0.08, 0.52]
+ Study 11 0.27 [0.10, 0.44]
+ Study 12 0.23 [0.07, 0.39]
+ Study 13 0.22 [0.08, 0.35]
+ Study 14 0.17 [-0.03, 0.37]
+ Study 15 0.16 [-0.01, 0.34]
Overall estimate (Zr)
d
Effect size (Zr)
Precision (1/SE)
Publication year
e
Precison (1/SE)
Effect size (Zr)
f
Standard error (SE)
Effect size (Zr)

FIGURE 5

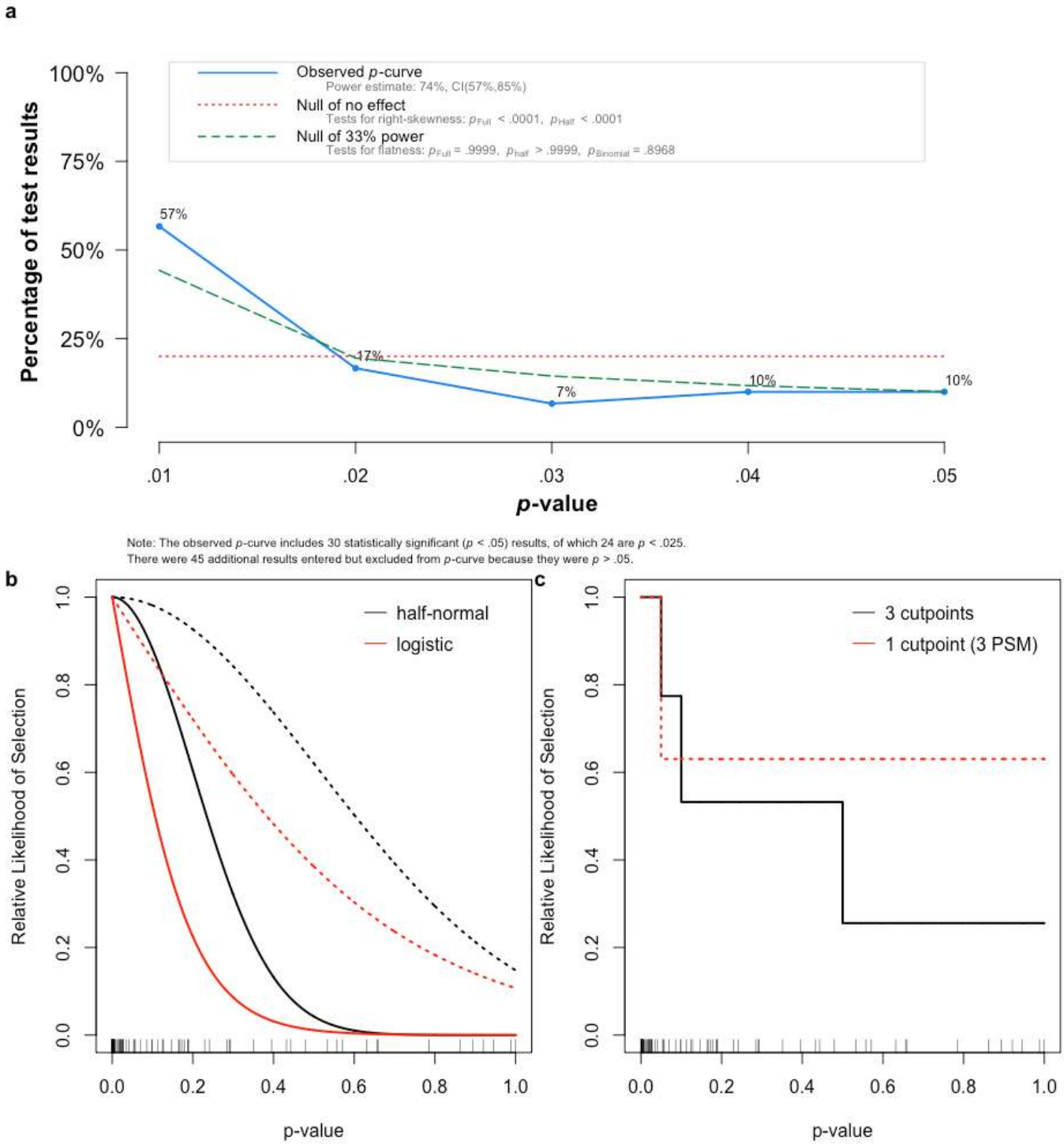

a
Observed p-curve
Power estimate: 74%, CI(57%,85%)
Null of no effect
Tests for right-skewness: p_{Full} < .0001, p_{Half} < .0001
Null of 33% power
Tests for flatness: p_{Full} = .9999, p_{half} > .9999, $p_{Binomial}$ = .8968
Percentage of test results
100%
75%
50%
25%
0%
57%
17%
7%
10%
10%
.01
.02
.03
.04
.05
p-value
Note: The observed p-curve includes 30 statistically significant (p < .05) results, of which 24 are p < .025.
There were 45 additional results entered but excluded from p-curve because they were p > .05.
b
half-normal
logistic
Relative Likelihood of Selection
0.0
0.2
0.4
0.6
0.8
1.0
p-value
c
3 cutpoints
1 cutpoint (3 PSM)
Relative Likelihood of Selection
p-value

FIGURE 6

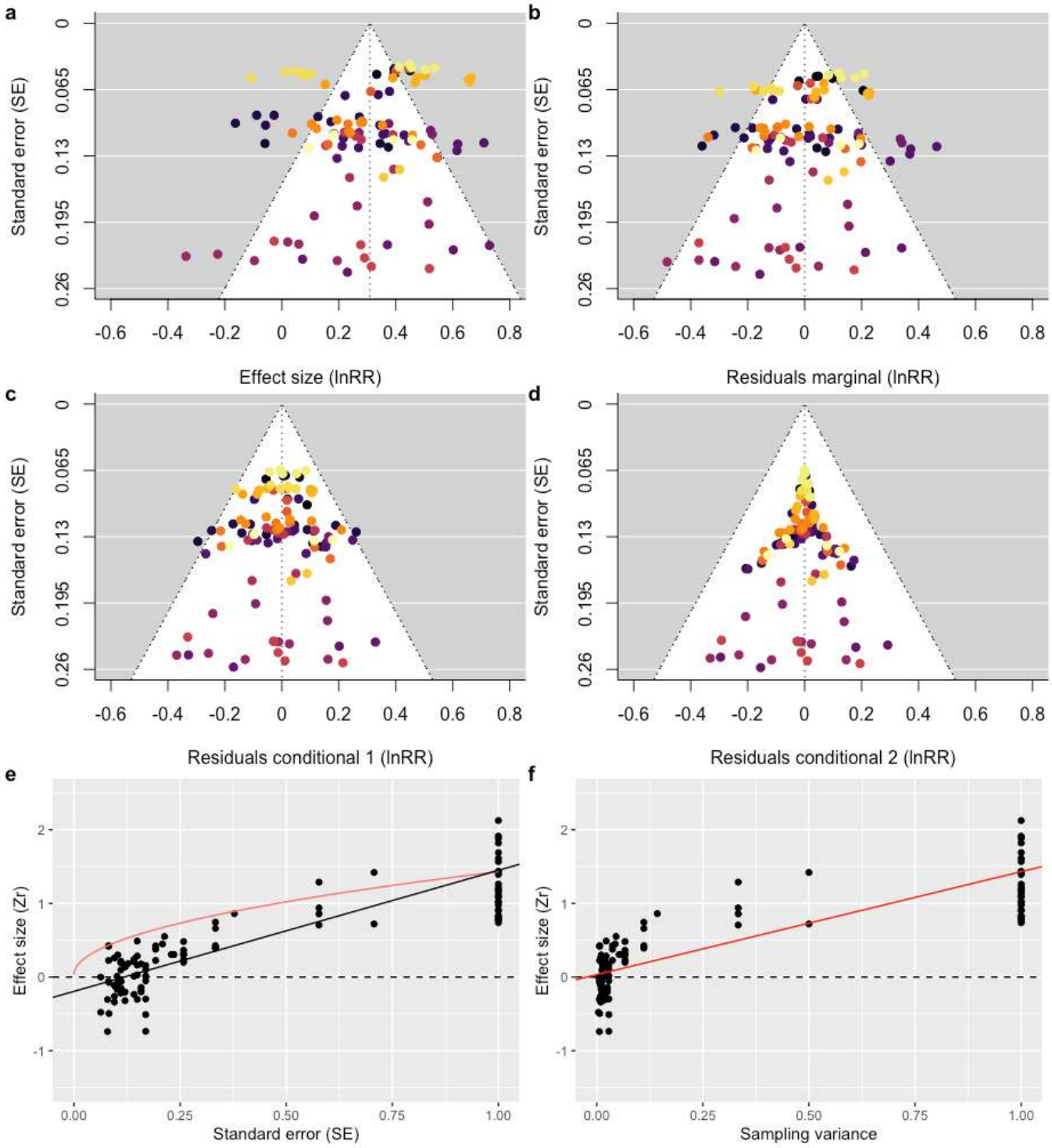

a
b
c
d
e
f
Standard error (SE)
0
0.065
0.13
0.195
0.26
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
Effect size (lnRR)
Residuals marginal (lnRR)
Residuals conditional 1 (lnRR)
Residuals conditional 2 (lnRR)
Effect size (Zr)
2
1
0
-1
0.00
0.25
0.50
0.75
1.00
Standard error (SE)
Sampling variance

FIGURE 7

	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	not applicable	no	no
Selection models	no	yes	yes	not applicable	yes	no
Multilevel meta-regression	no	yes	yes	yes	not applicable	yes