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

2　Shinichi Nakagawa[1][*], Malgorzata Lagisz[1], Michael D. Jennions[2], Julia Koricheva[3], Daniel W.A.

3　Noble[2], Timothy H. Parker[4], Alfredo Sánchez-Tójar[5], Yefeng Yang[1], & Rose E. O'Dea[1]

4

5　[1] Evolution & Ecology Research Centre and School of Biological, Earth and Environmental

6　Sciences, University of New South Wales, Sydney, NSW 2052, Australia

7　[2] Division of Ecology and Evolution, Research School of Biology, The Australian National

8　University, Canberra, ACT, Australia

9　[3] Department of Biological Sciences, Royal Holloway University of London, Egham, Surrey, TW20

10　0EX, U.K.

11　[4] Department of Biology, Whitman College, Walla Walla, WA 99362, U.S.A

12　[5] Department of Evolutionary Biology, Bielefeld University, Bielefeld, 33615, Germany

13

14　* Correspondence: S. Nakagawa

15　e-mail: s.nakagawa@unsw.edu.au

16

17　Short title: Publication bias tests for ecology & evolution

18

19　**ORCID**

20　Shinichi Nakagawa: 0000-0002-7765-5182; Malgorzata Lagisz: 0000-0002-3993-6127; Michael D.

21　Jennions: 0000-0001-9221-2788; Julia Koricheva: 0000-0002-9033-0171; Daniel W. A. Noble:

22　0000-0001-9460-8743; Timothy H. Parker: 0000-0003-2995-5284; Alfredo Sánchez-Tójar: 0000-

23　0002-2886-0649; Yefeng Yang: 0000-0002-8610-4016; Rose E. O'Dea: 0000-0001-8177-5075

24

**Abstract**

25   **Abstract**

26   1. Publication bias threatens the validity of quantitative evidence from meta-analyses as it results in

27      some findings being overrepresented in meta-analytic datasets because they are published more

28      frequently or sooner (e.g., 'positive' results). Unfortunately, methods to test for the presence of

29      publication bias, or assess its impact on meta-analytic results, are unsuitable for datasets with

30      high heterogeneity and non-independence, as is common in ecology and evolutionary biology.

31   2. We first review both classic and emerging publication bias tests (e.g., funnel plots, Egger's

32      regression, cumulative meta-analysis, fail-safe $N$, trim-and-fill tests, $p$-curve and selection

33      models), showing that some tests cannot handle heterogeneity, and, more importantly, none of

34      the methods can deal with non-independence. For each method we estimate current usage in

35      ecology and evolutionary biology, based on a representative sample of 102 meta-analyses

36      published in the last ten years.

37   3. Then, we propose a new method using multilevel meta-regression, which can model both

38      heterogeneity and non-independence, by extending existing regression-based methods (i.e.

39      Egger's regression). We describe how our multilevel meta-regression can test not only

40      publication bias, but also time-lag bias, and how it can be supplemented by residual funnel plots.

41   4. Overall, we provide ecologists and evolutionary biologists with practical recommendations on

42      which methods are appropriate to employ given independent and non-independent effect sizes.

43      No method is ideal, and more simulation studies are required to understand how Type 1 and 2

44      error rates are impacted by complex data structures. Still, limitations of these methods do not

45      justify ignoring publication bias in ecological and evolutionary meta-analyses.

46

47   **KEYWORDS:** Outcome reporting bias, $p$-hacking, multilevel meta-analysis, selection bias, radial

48   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):

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

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

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); 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:

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

$$\text{Var}(\ln\text{RR}_i) = \frac{\text{SD}_{1i}^2}{n_{1i}\overline{X}_{1i}^2} + \frac{\text{SD}_{2i}^2}{n_{2i}\overline{X}_{2i}^2}, \quad (4)$$

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):

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

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

where $n_i$ is the $i$th sample size used to obtain the correlation coefficient, $r_i$. Incidentally, the variance of the correlation coefficient is: $\text{Var}(r_i) = (1 - r_i^2)^2 / (n - 1)$, 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 $i$th effect size, normally distributed with a mean of zero and a variance of $\sigma_s^2$ (which is more commonly referred to as $\tau^2$; note when $\sigma_s^2 = 0$, this model reduces to a fixed-effect model), and $m_i$ is the sampling error for the $i$th effect size, distributed with the $i$th 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

6

139 proportion of $\sigma_s^2$ against the total variance is often quantified as $I^2 = \sigma_s^2 / (\sigma_s^2 + \bar{v})$ where $\bar{v}$ is

140 referred to as the 'typical' within-study (sampling) variance, which can be considered as a mean

141 value of $v_i$ (Higgins & Thompson, 2002). In ecological and evolutionary meta-analyses, $I^2$ is

142 around 90%, on average, meaning only ~10% of variation among effect sizes is due to sampling

143 variance (Senior *et al.*, 2016). Therefore, publication bias tests assuming homogeneity ($I^2$ or $\sigma_s^2 = 0$)

144 are unlikely to be useful for ecology and evolution.


145 **3 | PUBLICATION BIAS TESTS**

146 The primary goal of this section is to provide a non-exhaustive but up-to-date overview of

147 publication bias tests, both classic and emerging, especially for ecologists and evolutionary

148 biologists (cf. Moller & Jennions, 2001; Jennions *et al.*, 2013; for thorough technical reviews, see

149 Rothstein, Sutton & Borenstein, 2005; Vevea, Coburn & Sutton, 2019; Marks-Anglin & Chen,

150 2020; Marks-Anglin *et al.*, 2021). Therefore, we summarise different methods of testing for the

151 presence of publication bias and assessing its impact on meta-analytic findings – describing which

152 methods are suitable for datasets with high heterogeneity and non-independence. Our recent survey

153 of publication bias tests used in 102 ecology and evolutionary meta-analyses indicates that many of

154 these methods will be unfamiliar to ecologists and evolutionary biologists; Figure 2 shows the

155 results of the survey (for the details of survey procedure see Supporting Information, Appendix S1,

156 found at https://github.com/itchyshin/publication_bias).

157       Following Sutton (2009) (see also Vevea, Coburn & Sutton, 2019), we categorise

158 publication bias tests into two types: 1) detecting publication bias (e.g., funnel plots, Egger's

159 regression; Section 3.1), and 2) assessing the impact of publication bias (e.g., Fail-safe *N*, trim-and-

160 fill method, and selection models; Section 3.2). Publication bias, including outcome reporting bias,

161 creates patterns of missing data (known as 'funnel asymmetry'; see the next section). Commonly,

162 the magnitude of the overall effect is exaggerated because statistically non-significant effect sizes

163 are less likely to be published, especially when they are based on small sample sizes. For time lag-

164 bias, the magnitude of effect size, and its statistical significance, are related to publication year, so

165 that this bias requires different tests from publication and outcome reporting bias (see Section

166 3.1.3).

**3.1 | Detecting publication bias**

**3.1.1 | Funnel plots**

169 In the absence of publication bias and heterogeneity, plotting effect sizes against a measure of

170 certainty (or uncertainty; see Figure 1) should produce a symmetrical funnel shape around the

171 overall effect, referred to as a funnel plot. These graphs are the most popular method for detecting

172 publication bias in ecological and evolutionary meta-analyses (Figure 2). Funnel plots are also the

173 most preferred graphical tool to detect publication bias in the medical and social sciences (Sterne,

174 Becker & Egger, 2005; Sutton, 2009; Vevea, Coburn & Sutton, 2019; Marks-Anglin & Chen,

175 2020), even though many other graphical methods have been proposed such as weighted histograms

176 and normal quantile plots of effect sizes (as in Figure 2; for other graphical methods, see Rothstein,

177 Sutton & Borenstein, 2005; Marks-Anglin & Chen, 2020).

178       The original funnel plot used sample size as the measure of uncertainty (Light & Pillemer,

179 1984; Figure 3a). Yet, more recent recommendations are to use either SE, precision, variance or the

180 inverse of variance (Figure 1; Sterne, Becker & Egger, 2005; but for why sample size may often be

181 preferred, see Section 4.3). For these four quantities, unlike for sample size, we can draw 95%

182 confidence intervals (based on the *y*-axis; 1.96 x SE) that create a funnel, showing the degree of

183 heterogeneity among effect sizes (if data are homogeneous, most dots will be inside the 95%

184 confidence interval region, e.g., Figure 3b & c). This confidence region also makes it easier to see

185 funnel asymmetry caused by the lack of statistically non-significant effect sizes with high

186 uncertainties (see Figure 3b & c). In a similar vein, a contour-enhanced funnel plot shows different

187 statistical significance regions (around 0) to help detect asymmetry (Peters *et al.*, 2008; Figure 3c).

188 Lastly, Kossmeier, and colleagues (2020) have recently proposed a sunset funnel plot, a type of

189 contour-enhanced plot, which adds visual indicators of statistical power (Figure 3d).

190    One of the limitations of funnel plots is that funnel asymmetry can be caused not just by

191    publication bias (as in Figure 3b, missing large effect sizes of high uncertainties; see also Terrin,

192    Schmid & Lau, 2005). For instance, heterogeneity among effect sizes can create asymmetries of

193    many kinds (Figure 3b); the other potential sources of asymmetry are data irregularities (e.g.,

194    mistakes, frauds, unique observations; cf. Nakagawa & Lagisz, 2016), artefacts (see Section 4.3),

195    and chance (Egger *et al.*, 1997). As mentioned above, high heterogeneity is common in ecological

196    and evolutionary meta-analyses (Senior *et al.*, 2016). Therefore, a standard funnel plot is unlikely to

197    be informative about publication bias. To account for some of the heterogeneity, several researchers

198    recommend plotting residuals from a meta-regression model (Figure 3e; e.g., Roberts & Stanley,

199    2005). In practice, however, no meta-regression model would explain all the heterogeneity. The

200    remaining heterogeneity might still generate asymmetry in a residual funnel plot. The funnel plot

201    should, therefore, be seen as a tool to explore small-study effects where effect sizes based on small

202    sample sizes tend to be larger. Small-study effects may indicate publication bias, but not necessarily

203    (Sterne, Becker & Egger, 2005). Although extensive work exists on funnel plots and heterogeneity,

204    no systematic studies exist asking how funnel plots perform when effect sizes are correlated (but

205    see Section 4.1).

206    Before moving to the next section where we introduce inferential tests of funnel asymmetry

207    (or small-study effects), the radial plot proposed by Galbraith (1988) is worth mentioning, even

208    though our survey found no use of these plots in ecological and evolutionary meta-analyses. The

209    idea of a radial plot is similar to that of a funnel plot. The radial plot shows effect sizes divided by

210    their SEs (essentially, *z* scores) on the *y*-axis and corresponding precisions on the x-axis. The plot,

211    as in Figure 3f, has a slope with a zero intercept (solid line) and its 95% confidence interval based

212    on lines drawn from ±1.96 values (dashed lines) with the steepness of the slope representing the

213    overall mean. The radial plot is useful for visually detecting heterogeneity because data are

214    completely homogeneous when all the data are inside this rectangle (analogous to a funnel shape in

215    funnel plots). These axes of the radial plot (but not those of the funnel plot) help us better

216 understand the original inferential test for observed funnel asymmetry, the so-called Egger's

217 regression (Egger *et al.*, 1997), which is our next topic.

218

219 **3.1.2 | Regression- and correlation-based methods**

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

221
$$z_i = \beta_0 + \beta_1 prec_i + e_i, \qquad (8)$$

222
$$e_i \sim \mathcal{N}(0, \sigma_e^2),$$

223 where $z_i$ is the $i$th $z$ score obtained from dividing an effect size by its SE ($y_i/se_i$), $\beta_0$ is the intercept,

224 $\beta_1$ is the slope for the precision (*prec* or $1/se$) and $e$ is residuals, normally distributed with a

225 variance of $\sigma_e^2$. When $\beta_0$ (not $\beta_1$) is significantly different from zero, then we statistically detected

226 funnel asymmetry (Figure 4a); the more $\beta_0$ deviates from zero, the more severe the asymmetry.

227 Although Egger's regression checks for asymmetry in a funnel plot, Equation 8 does not

228 have effect sizes as a variable, while a funnel plot does (Figure 3). We intuitively like to draw a

229 regression line ($\beta_1$ and $\beta_0$) using Equation 8 in a funnel plot but this could be a confusing task as

230 ones needs to put $\beta_1$ as the intercept and $\beta_0$ as the slope. However, it is possible to reformulate

231 Egger's regression (Equation 8), so that its intercept ($\beta_0$) and its slope ($\beta_1$) can directly be used in a

232 funnel plot, using a weighted regression, as follows (Thompson & Sharp, 1999):

233
$$y_i = \beta_0 + \beta_1 se_i + \epsilon_i, \qquad (9)$$

234
$$\epsilon_i \sim \mathcal{N}(0, v_i\phi),$$

235 where $y_i$ is the $i$th effect size and $\epsilon_i$ is the residuals, normally distributed with a variance of $v_i\phi$,

236 which is sampling variance ($v$) and the multiplicative parameter ($\phi$) estimated in the weighted

237 regression (in a meta-regression, $\phi$ is set to be 1, which assumes that $v_i$ is the exact sampling

238 variance; see the next equation and also cf. Equation 7). Notably, Equation 8's $\beta_0$ is identical to

239 Equation 9's $\beta_1$ and also Equation 8's $\beta_1$ is identical to Equation 9's $\beta_0$ (we demonstrate this in

240 Supplementary Information, Appendix S2). Therefore, we can now look at the statistical

241 significance of the slope of SE ($se_i$ in Equation 9), whose magnitude indicates the severity of

242 asymmetry, and we are also able to put a regression line through a funnel plot (Figure 4b).

243   Given that Equation 9 is very similar to a meta-regression, later versions of Egger's

244 regression variants have taken the same form as a meta-regression (Moreno *et al.*, 2009), for

245 example:

246 $$y_i = \beta_0 + \beta_1 se_i + s_i + m_i, \quad (10)$$

247 $$s_i \sim \mathcal{N}(0, \sigma_s^2), m_i \sim \mathcal{N}(0, v_i),$$

248 which is the same as Equations 7 (the random-effects model) plus the slope of SE ($\beta_1$) (note that

249 different variants have precision, variance of the inverse of variance instead of SE; Moreno *et al.*,

250 2009).

251   According to simulation studies (Macaskill, Walter & Irwig, 2001; Sterne, Egger & Smith,

252 2001; Moreno *et al.*, 2009), Egger's regression and its variants suffer from low power and poor

253 performance when there are fewer than 20 effect sizes, or when the overall effect is large. However,

254 meta-analyses in ecology and evolution often include over 20 effect sizes and our overall effect is

255 usually small (Senior *et al.*, 2016). Therefore, the regression-based method for publication bias is

256 likely to be of use, at least to detect small-study effects. Furthermore, in this meta-regression

257 formulation it is possible to: 1) add moderators to absorb some heterogeneity, and 2) use multilevel

258 meta-regression to account for non-independence among effect sizes. We expand on these

259 possibilities in Section 4.

260   Similar to regression-based publication bias tests, correlation-based methods also

261 statistically test for a relationship between effect sizes and corresponding uncertainties (e.g.

262 sampling variance). All the correlation methods are based on a version of the rank correlation test

263 first proposed by Begg and Mazumdar (1994). This method essentially calculates a Kendall's rank

264 correlation between effect sizes and their sampling variance (or other uncertainty measures,

265 including sample size); a statistically significant correlation can indicate a small-study effect. Thus,

266 it is very simple to implement, but it seems that the rank correlation is less powerful than Egger's

267    regression under many circumstances (Macaskill, Walter & Irwig, 2001). Also, a recent simulation

268    shows that the rank correlation methods, using both sampling variance and sample size, had

269    severely inflated Type I error rates when effect sizes are correlated (Fernandez-Castilla *et al.*,

270    2019). Therefore, we recommend that meta-analysts use regression-based methods instead of

271    correlation-based methods to test for publication bias (in our survey, these methods were roughly

272    equally popular, being reported in around 10% of papers; Figure 2).

273

274    **3.1.3 | Time-lag bias tests**

275    Time-lag bias occurs when larger or statistically significant effects are published more quickly than

276    smaller or non-statistically significant effects, and can manifest as a decline in the magnitude of the

277    overall effect over time (i.e., a decline effect; Koricheva & Kulinskaya, 2019). According to our

278    survey (Figure 2), fewer than 5% of meta-analyses in ecology and evolution tested for this type of

279    publication bias. This is concerning, as time-lag bias is likely to be prevalent in ecology and

280    evolution (Jennions & Moller, 2002; Sanchez-Tojar *et al.*, 2018). To test for time-lag bias, we

281    caution against using correlation-based methods, because this approach does not account for

282    different precisions of effect sizes (e.g., quantifying a rank correlation between effect size and

283    publication year; Barto & Rillig, 2012). Instead, there are two recommended ways to investigate

284    time-lag bias (or a decline effect): 1) using a cumulative meta-analysis, and 2) using a regression-

285    based method (see Trkalinos & Ioannidis, 2005; Koricheva, Jennions & Lau, 2013; Koricheva &

286    Kulinskaya, 2019).

287       Cumulative meta-analysis is where a meta-analytic model (e.g., random-effects model) is

288    applied to a set of effect sizes, which is increased by one effect size at a time iteratively (starting

289    from the oldest effect size). Then, the results are displayed as a forest plot (see Figure 4c). One can

290    easily see when statistical significance or magnitude of the overall effect size changes over time.

291    When multiple effect sizes are obtained from each study, adding one study (one or more effect

292 sizes) rather than one effect size is more practical. For complex data structures (see Section 4.1),

293 limited sample sizes might prevent models from running in the early years of the dataset.

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

295 $$y_i = \beta_0 + \beta_1 year_i + s_i + m_i, \qquad (11)$$

296 where $year_i$ is the publication year for the $i$th study (effect size). As with Equation 8, this method

297 can accommodate other moderators (i.e. potential confounding variables) and also can be

298 extendable to model non-independent effect sizes (see Section 4.2).

299 **3.2 | Assessing the impact of publication bias**

300 **3.2.1 | Fail-safe $N$**

301 We now move to the methods that can assess the impact of publication bias rather than merely

302 detecting it. Fail-safe $N$ (also known as the 'file-drawer number') represents the number of non-

303 significant unpublished results needed to exist to make the overall effect non-significant (e.g.,

304 Rosenthal, 1979; Rosenberg, 2005) or negligible in magnitude (e.g., Owrin, 1983). If the fail-safe $N$

305 is large ($>5N_{study} + 10$), the results of analyses may be considered to be robust with respect to

306 publication bias as such large number of non-significant results is unlikely to exist. The original

307 fail-safe approach by Rosenthal (1979) is the oldest publication bias assessment method and

308 probably the simplest:

309 $$N_{Rosenthal} = \left(\frac{\sum_{i=1}^{N_{study}} z_i}{1.645}\right)^2 - N_{study}, \qquad (12)$$

310 where $z_i$ is the $i$th $z$ value ($y_i/se_i$) as in Equation 7 and 1.645 is the $z$ value for $\alpha = 0.05$ (the one-

311 tailed test). The method by Orwin (1983) relies on the magnitude of the effect size rather than

312 statistical significance; one version of this method can be written as:

313 $$N_{Orwin} = \frac{N_{study}(\bar{y} - y_n)}{y_n}, \qquad (13)$$

314 where $\bar{y}$ is the overall mean (i.e. an estimate from a fixed-effect model) and $y_n$ is the effect size

315 value that is considered to be small or negligible. Although Rosenthal's and Orwin's fail-safe

316    numbers ignore sample sizes (uncertainty) of effect sizes in the dataset, the method proposed by

317    Rosenberg (2005) explicitly includes such information. An equation that assumes a fixed-effect

318    model can be written as:

319
$$N_{Rosenberg} = \frac{N_{study}W}{\sum_{i=1}^{N_{study}} w_i}, \quad (14)$$

320
$$W = \left(\frac{\sum_{i=1}^{N_{study}} w_i y_i}{t_{0.05(N_{stduy})}}\right)^2 - \sum_{i=1}^{N_{study}} w_i,$$

321    where $w_i$ is the inverse of sampling variance ($1/v_i$; note that $w_i$ can be modified for a random-effects

322    model) and $t_{0.05(N_{stduy})}$ denotes the $t$ value with the α level of 0.05 with the number of studies

323    (effect sizes) as the degrees of freedom, DF (for the use of a different DF, see Rosenberg, 2005).

324         Although fail-safe approaches are the most popular method after the funnel plot in our

325    survey (14.1%), Becker (2005) has called for abandoning all the fail-safe approaches, now that

326    other methods for handling publication bias are available. Becker has argued that the fail-safe $N$ is

327    difficult to interpret (e.g., no criterion on what constitutes a small or large $N$), and also that depends

328    on the exact method, a variety of fail-safe numbers can be obtained for the same data set. For

329    example, the $R$ package *metafor* implements the three methods above (Viechtbauer, 2010); its

330    example dataset shows $N_{Rosenthal} = 598$, $N_{Orwin} = 84$, and $N_{Rosenberg} = 370$ (for details, see Supporting

331    Information, Appendix S3). Unfortunately, none of the proposed methods adequately control for

332    heterogeneity (e.g., by incorporating moderators) nor non-independence among effect sizes.

333    Furthermore, none of the methods of fail-safe $N$ are inferential.

334

335    **3.2.2 | Trim-and-fill tests**

336    The trim-and-fill test provides a non-parametric method that can visualize potentially missing data,

337    and statistically both detect and correct for funnel asymmetry (Duval & Tweedie, 2000b; Duval &

338    Tweedie, 2000a). A recent survey showed that the number of studies using the trim-and-fill method

339    is increasing every year (in 2018, over 2000 meta-analyses used this method; Shi & Lin, 2019), and

340     this method is not rare in ecology and evolution (7.5% of the meta-analyses in our survey). In short,

341     this method uses an iterative process to determine how many effect sizes are missing (say, $N_{missing}$)

342     from a funnel, using an initial overall estimate and one of three estimators ($R_0$, $L_0$, & $Q_0$; see an

343     accessible account in Duval, 2005). Then, it 'trims' off $N_{missing}$ effect sizes to suppress funnel

344     asymmetry, and estimates a new overall mean to see whether it can trim more effect sizes until the

345     value $N_{missing}$ stabilizes. Subsequently, $N_{missing}$ effect sizes are 'filled' as mirror images (Figure 4e &

346     f). Finally, an overall effect is re-estimated including the filled values. We note that Duval (2005)

347     has recommended the use of $R_0$ and $L_0$, and that the estimator $R_0$ can provide a significance test for

348     whether the number of missing values is zero or not.

349        The problem with the trim-and-fill test is that the original method assumes homogeneity (i.e.

350     a true mean for all effect sizes). In practice, the trim-and-fill method seems to tolerate some

351     heterogeneity, but performs worse as heterogeneity increases (Peters *et al.*, 2007; Moreno *et al.*,

352     2009). Although trim-and-fill tests have been extended for meta-regressions (Weinhandl & Duval,

353     2012), this implementation of this extension is currently limited to one moderator. Further, recent

354     simulation work by Rogers and Pustejovksy (2020) shows that ignoring non-independence and fitting

355     a trim-and-fill method (using $R_0$) increases Type I error rates, especially when a large overall effect

356     exists.

357

358     **3.2.3 | *P*-value-based methods and selection models**

359     Ecologists and evolutionary biologists have hardly used the available methods based on *p*-values

360     and selection models (*p*-value-based: 1.4%, selection models: 0%, Figure 2), even though both

361     types of methods can provide adjusted overall means. The *p*-curve method was introduced by the

362     same researchers who popularized the terms 'researcher degrees of freedom' (Simmons, Nelson &

363     Simonsohn, 2011) and '*p*-hacking' (Simonsohn, Nelson & Simmons, 2014). The *p*-curve method

364     relies on the distribution of statistically significant *p* values of effect sizes in a dataset (Figure 5a).

365     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.

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

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

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

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

where $\beta_0$ is the overall estimate (or meta-analytic mean); $s_j$ is the between-study effect for the $j$th study, normally distributed with the variance of $\sigma_s^2$; $u_i$ is the between-effect-size effect, or within-study effect, for the $i$th effect size, distributed with a mean of zero and the variance of $\sigma_u^2$; and $m_i$ is as in Equation 7 (but note that $j = 1, 2, \ldots, N_{study}$, the number of studies, and $i = 1, 2, \ldots, N_{effect\text{-}size}$, the number of effect sizes; $N_{effect\text{-}size} > N_{study}$). Equation 15 explicitly models multiple effect sizes per study. Also, in Equation 7, the term $\sigma_s^2$ is the only source of heterogeneity, while in Equation 15, both $\sigma_s^2$ and $\sigma_u^2$ are each contributing to heterogeneity among effect sizes.

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

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

where $\beta_1$ is the slope for $x_1$, a study-level moderator (characteristics of different studies, $j$; e.g., experimental *vs.* observational) and $\beta_2$ 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 (Figure 6a). A complication is that, given Equation 15, we can extract at least 3 different residuals, which are:

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

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

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

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.

416  sample sizes are the same or similar). Contrastingly, conditional residuals no longer show such

417  obvious patterns as we have taken a clustering factor ($s_j$), meaning that these residuals are

418  independent, at least with respect to this factor. Thus, funnel plots with conditional residuals (Figure

419  6c-d) seem like a useful exploratory tool for publication bias when effect sizes are correlated, in

420  addition to using marginal residuals (Figure 6b).

421  As the conditional residuals are supposed to be independent, Nakagawa and Santos (2012)

422  suggested using conditional residuals along with corresponding sampling variance or standard error

423  ($v_i$ or $se_i$) in publication bias tests (e.g., the original Egger's regression and trim-and-fill tests).

424  However, this approach is limited by some assumptions. First, all such residual analyses assume

425  that sampling SE ($se_i$) does not covary with moderators in meta-regression (e.g., $x_1$ and $x_2$ in

426  Equation 16; see Freckleton, 2002). Second, sampling SE is assumed to be the same as the SE of

427  the residuals (which are shown in Figures 6b-d), but they are not the same, although they are often

428  strongly correlated (see Doleman *et al.*, 2020). Finally, in the presence of non-independent data,

429  Equation 15's sampling variances are often correlated; that is, $m_i \sim \mathcal{N}(0, \mathbf{M})$ where $\mathbf{M}$ is a

430  variance-covariance matrix. For example, when $N_{effect\text{-}size} = 3$ and the first two effect sizes' sampling

431  variance are correlated, then we can write $\mathbf{M}$ as:

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

432

433  where $\rho$ is the correlation between the sampling effects of the first two effect sizes ($\rho\sigma_1\sigma_2$ is the

434  covariance). Whenever sampling (error) effects are correlated, neither $resid_{c1}$ nor $resid_{c2}$ are

435  independent. Then, none of publication bias tests reviewed in Section 3 should be used.

436  Incidentally, we note that the robust variance estimator (RVE) originally proposal by Hedges *et al.*

437  (2010) can circumvent modelling the variance-covariance matrix $\mathbf{M}$ even when sampling errors are

438  correlated. This is because covariances are estimated from the data and the associated errors are

439  reflected in standard errors (variance) of point estimates via the RVE (cf. Rodgers & Pustejovsky,

440  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; Fernandez-Castilla *et al.*, 2019; Rodgers & Pustejovsky, 2020):

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

By examining Equation 21, we may realise that $\beta_0$ represents a conditional estimate of an overall

effect when SE is 0, which means, theoretically, there is no uncertainty (Figure 5e). Then, does $\beta_0$

provide an adjusted estimate of an overall effect, when $\beta_1$ is statistically significant (i.e., detecting a

small-study effect)? This question has been examined by Stanley and Doucouliagos (2012; 2014).

They have shown that, with significant $\beta_1$, $\beta_0$ provides an adjusted estimate that is downwardly

biased, when a true positive or a null effect exists (which is illustrated in Figure 5e; note that they

state that with non-statistically significant $\beta_1$, $\beta_0$ provides the best estimate of an adjusted mean). If

the slope of SE ($\beta_1$) is statistically significant then fitting sampling variance instead of SE is

recommended according to the following equation:

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

This is equivalent to fitting $se_i^2$, which is a quadratic term. Stanley and Doucouliagos (2012; 2014)

have shown that $\beta_0$ 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 5f). While this two-step approach may seem simplistic (see also Stanley, 2017;

Stanley, Doucouliagos & Ioannidis, 2017), it provides an easy-to-implement publication bias test

which explicitly models non-independent data.

Further, this regression approach can be used to test time-lag bias (or decline effect) by

modelling the publication year (*year$_j$*):

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

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 $\beta_k$ is the slope for the $k$th moderator ($k = 3, 4,\ldots, N_{mod}$; 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 $\beta_0$ 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

489 (Macaskill, Walter & Irwig, 2001; Deeks, Macaskill & Irwig, 2005), and non-independent effect

490 sizes (Fernandez-Castilla *et al.*, 2019).

491     Instead of the sample size ($n_1 + n_2$), however, for a meta-analysis of SMD or lnRR we

492 propose using the 'effective sample size' ($4\tilde{n}_i$) because it accounts for unbalanced sampling. The

493 effective sample size is given by (Bakbergenuly, Hoaglin & Kulinskaya, 2020b; 2020a; also see;

494 Deeks, Macaskill & Irwig, 2005; Bakbergenuly, Hoaglin & Kulinskaya, 2020c):

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

496 When $n = n_1 = n_2$, the formula reduces to $2n$. Indeed, the inverse of $\tilde{n}_i$ is a part of sampling variance

497 in both SMD and lnRR (Equations 4 & 6):

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

499 where the middle part of the formula corresponds to Equation 2 when setting SMD = 0, while the

500 right-hand side corresponds to Equation 4 when setting CV ($SD/\overline{X}$) = 1. This means that the use of

501 $\tilde{n}_i$ is comparable to that of sampling variance after taking out uncertain elements.

502     Taken together, we can rewrite Equations 21 and 22, respectively, as (Deeks, Macaskill &

503 Irwig, 2005):

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

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

506 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

507 (note that, at the intercept, $\tilde{n}_i$ is infinitely large). We recommend using Equation 27 to check the

508 statistical significance of funnel asymmetry (small-study effects) because it has greater statistical

509 power than Equation 28. Equation 27 can also be used to obtain an adjusted mean when $\beta_1$ is not

510     statistically significant. This is because $\beta_0$ represents an adjusted overall mean when $\sqrt{\frac{1}{\tilde{n}_i}} = 0$. In

511     other words, the predicted overall mean when a study has an infinitely large sample size, $\tilde{n}_i$, and

512     therefore little to no sampling variance. In contrast, when $\beta_1$ is statistically significant in Equation

513     27, we recommend using Equation 28 to obtain an overall estimate adjusted for publication bias

514     because it is less biased. Note that these recommendations are for the effect sizes SMD and lnRR

515     (with *Zr*, we should use Equations 21 and 22). This adjusted estimate should not be taken as a true

516     estimate, however. We should treat it as a possible overall estimate as a part of sensitivity analysis

517     in which we run alternative statistical models to test the robustness of results from the original

518     analysis (Noble *et al.*, 2017).

519        In practice, multilevel meta-analytic models are often more complex. For example,

520     Nakagawa and Santos (2012) proposed a phylogenetic multilevel model with a phylogenetic

521     random factor and a non-phylogenetic random factor as a theoretically sound model when effect

522     sizes are obtained from different species (see also Hadfield & Nakagawa, 2010). The major benefit

523     of our proposed meta-regression approach for publication bias tests is that we can easily extend

524     these models to incorporate other sources of heterogeneity. An example of a meta-regression model

525     testing publication bias and time-lag bias that also includes phylogenetic and non-phylogenetic

526     random effects can be written as:

527
$$y_i = \beta_0 + \beta_1 \sqrt{\frac{1}{\tilde{n}_i}} + \beta_2 c(year_j) + \sum_{k=3}^{N_{mod}} \beta_k x_k + a_h + q_h + s_j + u_i + m_i, \quad (29)$$

528
$$a_h \sim \mathcal{N}(0, \sigma_a^2 \mathbf{A}), q_h \sim \mathcal{N}(0, \sigma_q^2), m_i \sim \mathcal{N}(0, \mathbf{M}),$$

529     where $a_h$ is the phylogenetic effect for the *h*th species, considered multivariate normally distributed

530     with a covariance of $\sigma_a^2 \mathbf{A}$ (**A** is a correlation matrix derived from a phylogeny); $q_h$ is the non-

531     phylogenetic effect for the *h*th species, distributed with the variance of $\sigma_q^2$ ($h = 1, 2, \ldots, N_{species}$, the

532     number of species; $N_{species} \neq N_{study}$); and the other notations are the same as above. Relevantly,

533     when using SMD or lnRR, we may be better off using $\tilde{n}_i$ along with residuals for drawing funnel

534    plots (see Section 4.1; Doleman *et al.*, 2020) rather than SE, precision, or variance. In the

535    Supporting Information we use two datasets and the three effect sizes to illustrate how to practically

536    code these proposed methods (see Appendix S4).

537    **4.4 | Alternative approaches: averaging or sampling**

538    Many of the methods we introduced in Section 3 are still useful, even in the presence of non-

539    independent data, if we aggregate effect sizes per study or sample one effect size per study. When

540    sampling variances are correlated (i.e. **M** as in Equation 29), 'average' sampling variance needs to

541    be calculated by using the following formula (not by simple weighted averaging as for the mean;

542    Borenstein *et al.*, 2009):

$$543 \quad \mathrm{Var}\left(\frac{1}{N_{within}}\sum_{g=1}^{N_{within}} y_g\right) = \left(\frac{1}{N_{within}}\right)^2 \left(\sum_{g=1}^{N_{within}} \sigma_g^2 + \sum_{g\neq l}^{N_{within}} r_{gl}\sqrt{\sigma_g^2 \sigma_l^2}\right), \quad (30)$$

544    where $y_g$ and $y_l$ are the $g$th and $l$th effect size in a study ($g = 1, \ldots, N_{within}$ and $l = 1, \ldots, N_{within}$ where

545    $N_{within}$ is the number of effect sizes within a paper or a species to be combined), $\sigma_i^2$ and $\sigma_g^2$ are the

546    sampling error variances for $y_g$ and $y_l$, and $r_{gl}$ is the correlation between the sampling errors of $y_g$

547    and $y_l$.

548         Overall means will generally not be biased using aggregated or single sample/study effect

549    sizes (Song *et al.*, 2020). Also, Rodgers and Pustejovsky (2020) showed that when averaging effect

550    sizes within studies, all Egger's regression (similar to Equation 10), the trim-and-fill test (using $R_0$

551    estimator) and the three-parameter selection model (as in Vevea & Hedges, 1995) had the

552    appropriate level of Type 1 error, although the three-parameter selection model was noticeably

553    more powerful than the others. However, averaging or sampling is not a general solution when we

554    have a phylogenetic signal ($\sigma_a^2 > 0$; Equation 29). In such a case, averaging or sampling per species

555    will not eliminate non-independence as effect sizes are still correlated via phylogeny (i.e. **A** in

556    Equation 29). Furthermore, even when there is no phylogenetic signal ($\sigma_a^2 = 0$), or we do not have

557    the species-level structure in a dataset, these alternative approaches could be problematic. For

558    example, if we average effect sizes, we will lose all effect-size-level moderators (e.g., one cannot

23

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.

**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).

## 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.

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

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.

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

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

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

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

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.

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

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.

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.

657     Hadfield, J.D. & Nakagawa, S. (2010) General quantitative genetic methods for comparative

658         biology: phylogenies, taxonomies and multi-trait models for continuous and categorical

659         characters. *Journal of Evolutionary Biology,* **23,** 494-508.

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

661         *guide*.

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

663     Hedges, L.V. (1984) Estimation of effect size under nonrandom sampling: the effects of censoring

664         studies yielding statistically insignificant mean differences. *Journal of Educational and*

665         *Behavioral Statistics,* **9,** 61–85.

666     Hedges, L.V., Gurevitch, J. & Curtis, P.S. (1999) The meta-Analysis of response ratios in

667         experimental ecology. *Ecology,* **80,** 1150-1156.

668     Hedges, L.V., Tipton, E. & Johnson, M.C. (2010) Robust variance estimation in meta-regression

669         with dependent effect size estimates. *Research Synthesis Methods,* **1,** 39-65.

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

671         *Medicine,* **21,** 1539-1558.

672     Jennions, M.D., Lorite, C.J., Rosenberg, M.S. & Rothstein, H.R. (2013) Publication and related

673         biases. *The handbook of meta-analysis in ecology and evolution* (eds J. Koricheva, J.

674         Gurevitch & K. Mengersen), pp. 207-236. Princeton University Press, Princeton.

675     Jennions, M.D. & Moller, A.P. (2002) Relationships fade with time: a meta-analysis of temporal

676         trends in publication in ecology and evolution. *Proceedings of the Royal Society B-*

677         *Biological Sciences,* **269,** 43-48.

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

679         *Ecology,* **102,** 828-844.

680     Koricheva, J., Jennions, M.D. & Lau, J. (2013) Temporal trends in effect sizes: causes, detection

681         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.

757 Song, C., Peacor, S.D., Osenberg, C.W. & Bence, J.R. (2020) An assessment of statistical methods

758     for nonindependent data in ecological meta-analyses. *Ecology,* **101,** e03184.

759 Stanley, T.D. (2017) Limitations of PET-PEESE and other meta-analysis methods. *Social*

760     *Psychological and Personality Science,* **8,** 581-591.

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

762     Routledge, New York.

763 Stanley, T.D. & Doucouliagos, H. (2014) Meta-regression approximations to reduce publication

764     selection bias. *Research Synthesis Methods,* **5,** 60-78.

765 Stanley, T.D., Doucouliagos, H. & Ioannidis, J.P.A. (2017) Finding the power to reduce publication

766     bias. *Statistics in Medicine,* **36,** 1580-1598.

767 Sterne, J.A., Egger, M. & Smith, G.D. (2001) Systematic reviews in health care: Investigating and

768     dealing with publication and other biases in meta-analysis. *British Medical Journal,* **323,**

769     101-105.

770 Sterne, J.A.C., Becker, B.J. & Egger, M. (2005) The funnel plot. *Publication bias in meta-analysis :*

771     *prevention, assessment and adjustments* (eds H. Rothstein, A.J. Sutton & M. Borenstein),

772     pp. 75-98. Wiley, Chichester.

773 Sutton, A.J. (2009) Publication bias. *The handbook of research synthesis and meta-analysis* (eds H.

774     Cooper, L.V. Hedges & J.C. Valentine), pp. 435-452. Russell Sage Foundation, New York.

775 Terrin, N., Schmid, C.H. & Lau, J. (2005) In an empirical evaluation of the funnel plot, researchers

776     could not visually identify publication bias. *Journal of Clinical Epidemiology,* **58,** 894-901.

777 Thompson, S.G. & Sharp, S.J. (1999) Explaining heterogeneity in meta-analysis: A comparison of

778     methods. *Statistics in Medicine,* **18,** 2693-2708.

779 Trkalinos, T.A. & Ioannidis, J.P.A. (2005) Assessing the evolution of effect sizes over time.

780     *Publication bias in meta-analysis : prevention, assessment and adjustments* (eds H.

781     Rothstein, A.J. Sutton & M. Borenstein), pp. 241-259. Wiley, Chichester.

782    van Aert, R.C.M., Wicherts, J.M. & van Assen, M.A.L.M. (2016) Conducting Meta-Analyses

783        Based on p Values: Reservations and Recommendations for Applying p-Uniform and p-

784        Curve. *Perspectives on Psychological Science,* **11,** 713-729.

785    van Assen, M.A.L.M., van Aert, R.C.M. & Wicherts, J.M. (2015) Meta-Analysis Using Effect Size

786        Distributions of Only Statistically Significant Studies. *Psychological Methods,* **20,** 293-309.

787    Vevea, J.L., Coburn, K. & Sutton, A.J. (2019) Publication bias. *The handbook of research synthesis*

788        *and meta-analysis* (eds H.M. Cooper, L.V. Hedges & J.C. Valentine), pp. 383-429. Russell

789        Sage Foundation, New York.

790    Vevea, J.L. & Hedges, L.V. (1995) A General Linear-Model for Estimating Effect Size in the

791        Presence of Publication Bias. *Psychometrika,* **60,** 419-435.

792    Viechtbauer, W. (2010) Conducting meta-analyses in R with the metafor package. *Journal of*

793        *Statistical Software,* **36,** 1-48.

794    Weinhandl, E.D. & Duval, S. (2012) Generalization of trim and fill for application in meta-

795        regression. *Research Synthesis Methods,* **3,** 51-67.

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

797

798

799

800

**FIGURE LEGENDS**

**FIGURE 1.** A schematic showing the relationship among 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'.

**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 Supporting Information, 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 as the overall effect estimate as a true effect; e) a residual funnel plot (one moderator removed); and f) a radial plot. We used the *R* packages *metafor* (panels a-c & e; Viechtbauer, 2010), *metaviz* (panel d; Kossmeier, Tran & Voracek, 2020) and *meta* (panel f; Schwarzer, Carpenter & Rücker, 2015) for 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;

851     e.g., Carter *et al.*, 2019; Rodgers & Pustejovsky, 2020). We used the *R* packages *dmetar* (panel a;

852     Harrer *et al.*, 2019) and *metafor* (panel b & c; Viechtbauer, 2010) for visualizations.

853

854     **FIGURE 6.** Examples of funnel plots from a dataset with lnRR ($N_{study}$ = 70; $N_{effect\text{-}size}$ = 271) and a

855     different dataset with *Zr* ($N_{study}$ = 48; $N_{effect\text{-}size}$ = 104): a) a funnel plot of raw data (the same colour

856     indicating effect sizes from the same studies); b) a funnel plot of marginal residuals with the fixed

857     effects removed (as in Equation 17); c) a funnel plot of conditional residuals with fixed effects and

858     the between-study effect removed (as in Equation 18); and d) a funnel plot of conditional residuals

859     with all effects apart from sampling errors removed (as in Equation 19); e) a scatterplot showing a

860     meta-regression on SE (black line; the red line is the same line as in panel f). Note that an overall

861     mean is set to be 0 in this simulated dataset along missing effect sizes imitating publication bias;

862     and f) a scatterplot showing a meta-regression on sampling variance (red line, the same line as in

863     panel 'e'). Both red lines showing to intersect the zero effect size at the intercept. We used the *R*

864     packages *metafor* (panels a-d; Viechtbauer, 2010) and *ggplot2* (panels e-f; Wickham, 2009) for

865     visualizations.

866

867     **FIGURE 7.** A summary of main publication bias tests reviewed in this article, and our

868     recommendations under two different conditions (effect sizes are independent or non-independent).

869     Superscript notes: 1) for funnel plots, residuals from a meta-regression can be plotted instead of raw

870     effect sizes, and using sample sizes instead of standard errors may be a good option for lnRR and

871     SMD; 2) for non-multilevel regression methods, precision and sampling variance (or $\sqrt{1/\tilde{n}_i}$ and

872     $1/\tilde{n}_i$) can be used; 3) technically, fail-safe *N* methods do not provide an adjusted overall mean, but

873     the numbers indicate how many non-significant studies (null effect sizes) would render the overall

874     effect zero (or a particular small effect size value); 4) for trim-and-fill methods, although some

875     heterogeneity can be tolerated the ability to model moderators is limited; alternatively, residuals

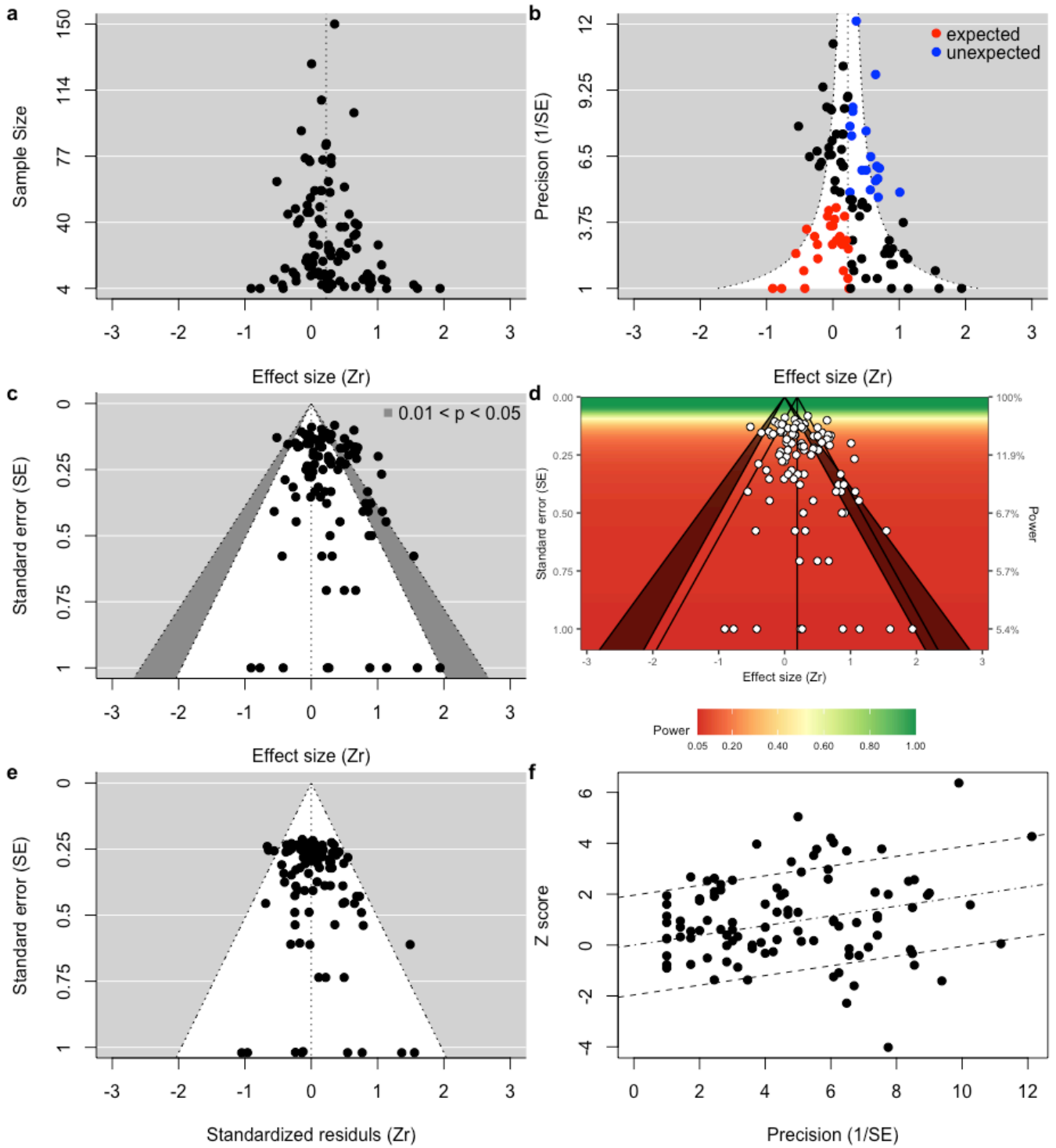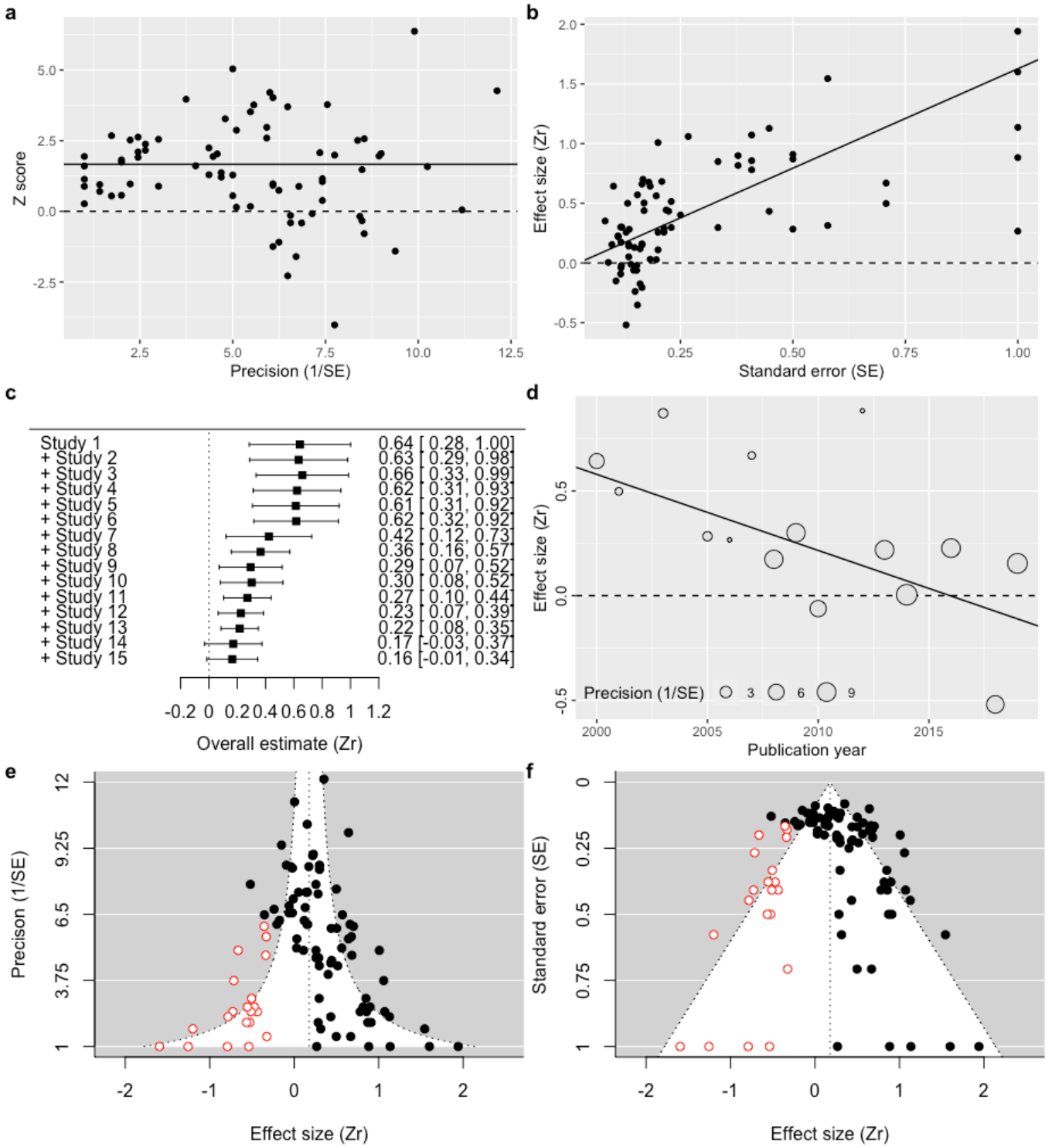876     along with their corresponding variances could be used.

**FIGURE 1**

**FIGURE 2**



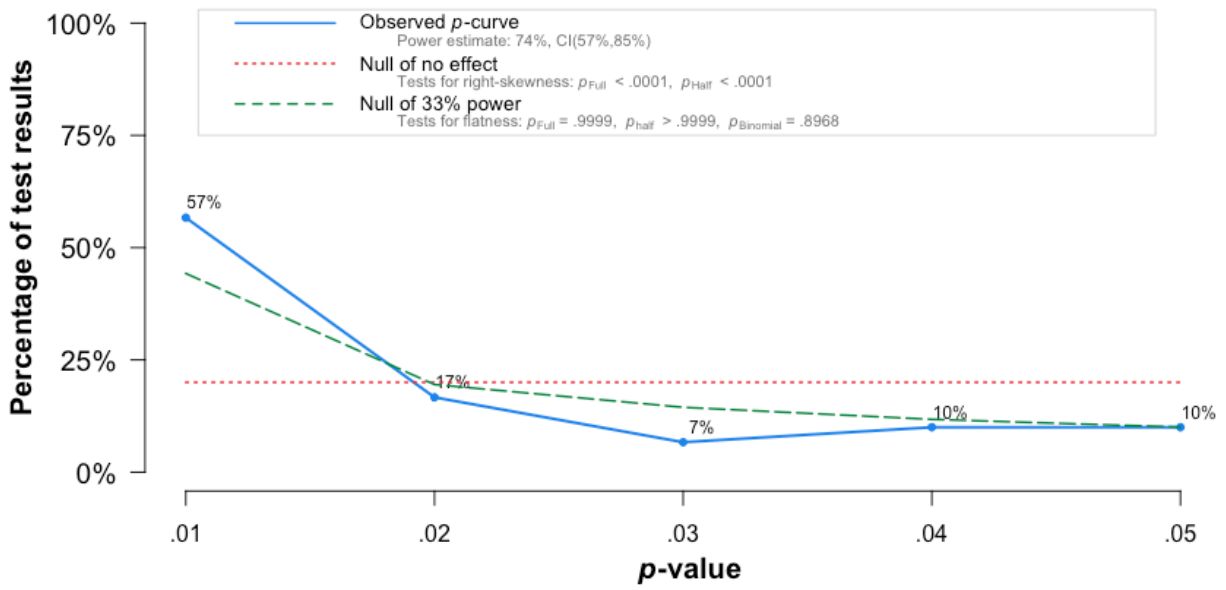number of papers reporting test
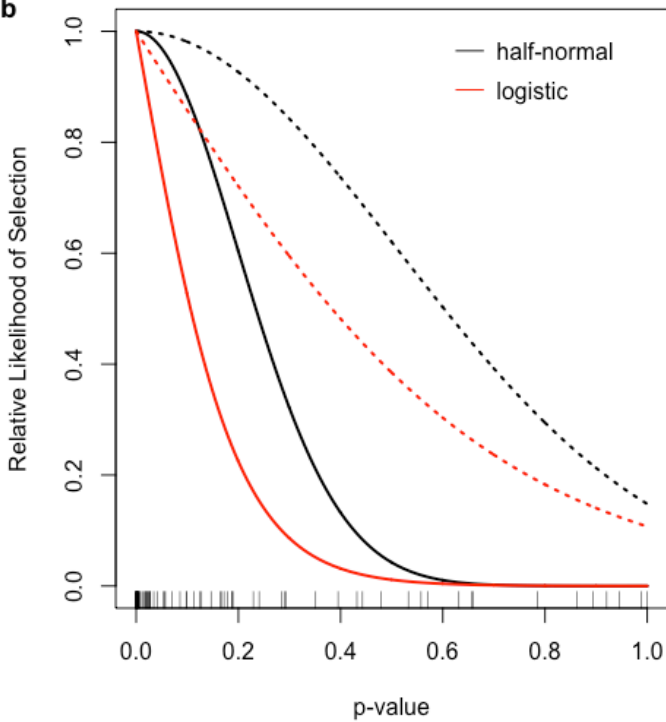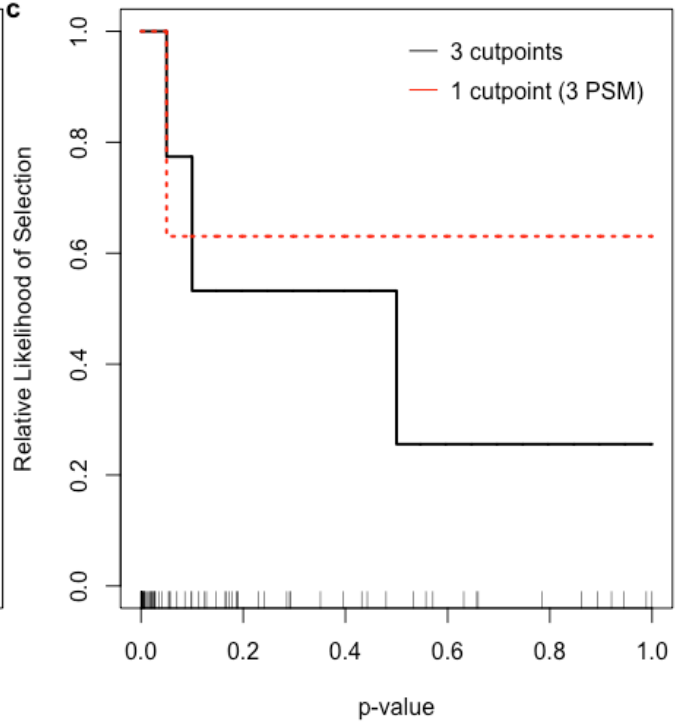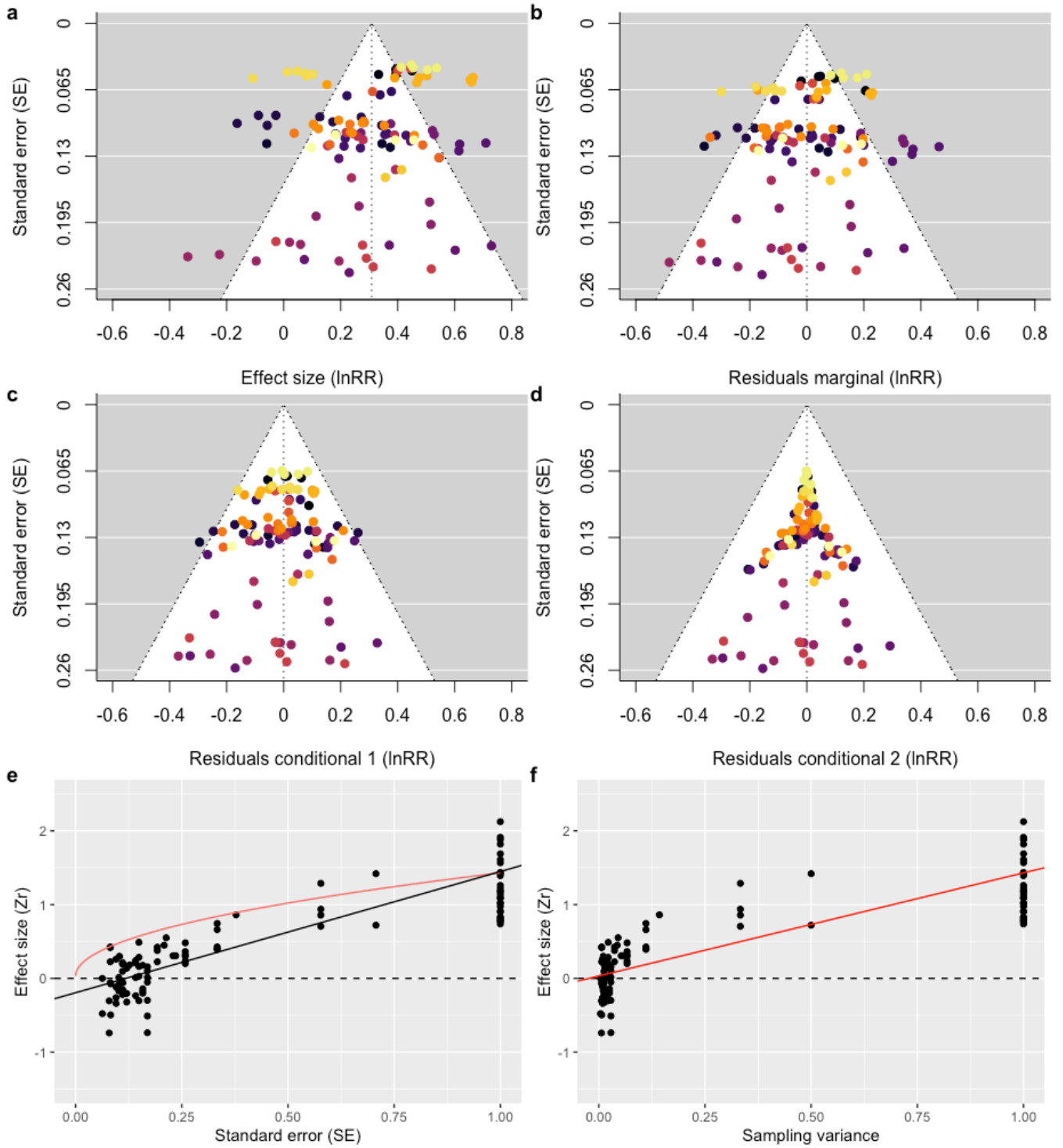
885
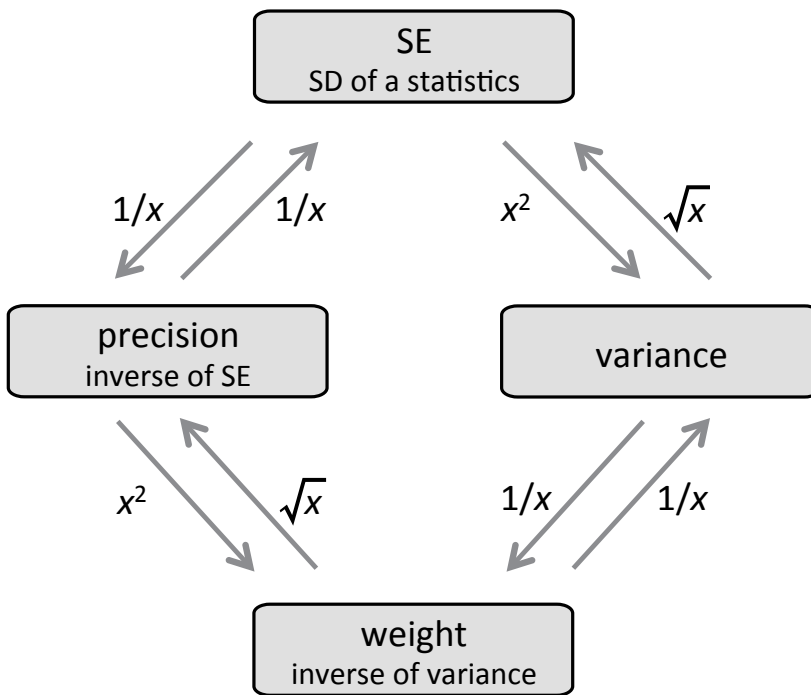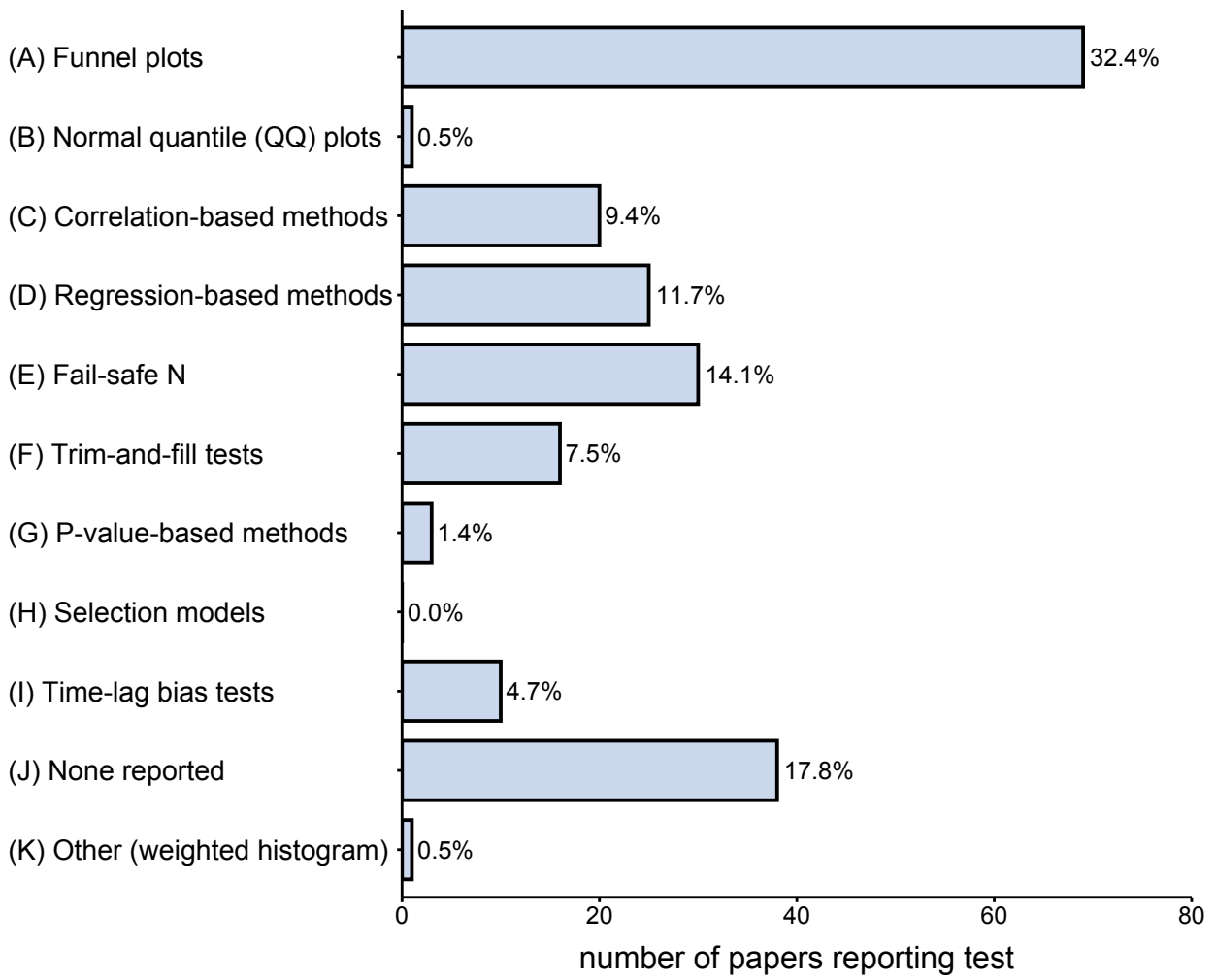
## FIGURE 3



887

888

890

891

**FIGURE 5**

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.

**FIGURE 6**

900

| | 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 |

901

902

903

904

905

906