Please cite as:

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021). Detecting Selection Bias in Meta-Analyses with Multiple Outcomes: A Simulation Study, *The Journal of Experimental Education*, 89:1, 125-144.
DOI: 10.1080/00220973.2019.1582470

Detecting Selection Bias in Meta-Analyses with Multiple Outcomes: A Simulation Study

Belén Fernández-Castilla[1,2], Lies Declercq[1,2], Laleh Jamshidi[1,2], S. Natasha Beretvas[4], Patrick Onghena[1] & Wim Van den Noortgate[1,2]

[1] Faculty of Psychology and Educational Sciences, KU Leuven, University of Leuven, Belgium

[2] Imec-ITEC, KU Leuven, University of Leuven, Belgium

[4] University of Texas at Austin, Texas, USA

Corresponding Author:

Belén Fernández-Castilla (Department of Psychology and Educational Science, KU Leuven – IICK Building – Box 1.33, Etienne Sabelaan 51, 8500, Kortrijk, Belgium)

Email address: belen.fernandezcastilla@kuleuven.be

Abstract

This study explores the performance of classical methods for detecting publication bias, namely Egger's Regression test, Funnel Plot test, Begg's Rank Correlation and Trim and Fill method, in meta-analysis of studies that report multiple effects. Publication bias, outcome reporting bias, and a combination of both were generated. Egger's Regression and Funnel Plot test were extended to three-level models, and possible cutoffs for the L_0^+ estimator of the Trim and Fill method were explored. Furthermore, we checked whether the combination of results of several methods yielded a better control of Type I error rates. Results show that no method works well across all conditions, and that their performance depends mainly on the population effect size value and on the total variance.

Keywords: meta-analysis, multiple effect sizes, publication bias, selective outcome reporting bias, simulation study.

Detecting Selection Bias in Meta-Analyses with Multiple Outcomes: A Simulation Study

Publication bias constitutes one of the biggest threats to the validity of meta-analytic results. It is defined as the propensity for publishing positive, significant results to the detriment of studies that report negative or non-significant results (Rosenthal, 1979). Because studies with no novel or positive results take more time to publish or are not even published at all (Decullier, Lheritier, & Chapuis, 2005; Ioannidis, 1998), it is less likely that they will find their way into meta-analyses. Unfortunately, a large body of research provides evidence of the presence of publication bias in many disciplines, including, for example, medicine (Dwan, Gamble, Williamson, & Kirkham, 2013; Ioannidis, 1998) and psychology (e.g., Fanelli, 2012; Franco, Malhotra, & Simonovits, 2014). A related kind of selection bias occurs when only a non-random selection of effect sizes are reported (e.g. the most impressive or significant ones) among multiple outcomes obtained within a study (Fusar-Poli, Nosek, & David, 2014; Hutton & Williamson, 2000; Tannock, 1996).

Several approaches exist to deal with publication bias. The most commonly used methods include the Egger Regression test (Egger, Davey-Smith, Schneider, & Minder, 1997), the Funnel Plot test (Macaskill, Walter, & Irwig, 2001), the Begg Rank Correlation test (Begg & Mazumdar, 1994) and the Trim and Fill method (Duval & Tweedie, 2000a, 2000b). They are all based on the relationship between effect size and sample size that is expected when selection bias exists. This effect is known as the ‘small-study effect’, whereby studies with large effect sizes and small sample sizes are more likely to be published than studies with the same sample size but reporting unfavorable or non-significant results. This effect size – sample size relationship leads to an asymmetrical funnel plot shape, with some suppressed studies in the lower left part of the figure (if the real or expected overall effect size is positive). As said before, these methods are the most commonly applied (Ferguson & Brannick, 2012). Possible explanations for their popularity are

that they are easy to understand and that they are available in most of the common software for performing meta-analysis (e.g., package *metafor* in R, SAS, Stata or Comprehensive Meta-Analysis). However, it should be kept in mind that the asymmetry of the funnel plot does not necessarily indicate the presence of publication bias. Asymmetric funnel plots can be the result of other underlying phenomena, such as high heterogeneity in the meta-analytic data, study – quality effects or a relation between the size of the study and the types of intervention studied (Egger et al., 1997; Ioannidis, Cappelleri, & Lau, 1998). This is why serious doubts have been raised about their performance in previous studies (e.g. Terrin, Schmid, Lau, & Olkin, 2003).

Several simulation studies have explored the performance of these methods (e.g., Kromrey & Rendina-Gobioff, 2006; Macaskill et al., 2001; Moreno et al., 2009; Peters, Sutton, Jones, Abrams, & Rushton, 2006; Sterne, Gavaghan, & Egger, 2000). However, no study has ever explored the performance of the classical methods for detecting publication bias in the common situation where multiple (and dependent) effects sizes are reported within primary studies. In fact, studies that introduce and explain techniques for combining studies with multiple effect sizes (e.g., Cheung, 2014; Hedges, Tipton, & Johnson, 2010; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013, 2015) do not recommend ways in which selection bias can be assessed. This lack of information is acknowledged in some applied meta-analyses, like the one authored by Platt et al., (2016) who claim that publication bias was not appropriately addressed in their paper because ‘tests for publication bias do not yet exist for multilevel meta-analyses’ (p. 17). Similarly, Assink and Wibbelink (2016) have encouraged an evaluation of the performance of methods for detecting publication bias when effects for multiple outcomes are reported within primary studies. On top of that, previous simulations have never explored the performance of these methods under other types of selection biases, including selective outcome reporting bias. Given this need for such methods, the aim of this study is to fill

this gap by exploring the performance of (the extension of) the classical methods for detecting publication and selective outcome reporting bias when multiple effect sizes are reported in primary studies.

Besides the classical methods, other more sophisticated procedures exist to deal with publication bias that may perform better. For instance, Stanley and Doucouliagos (2014) have proposed a modification of Egger's regression test, and have shown how this meta-regression, called precision-effect estimate with standard error (PEESE), corrects better for publication bias than other regression methods. Other alternative methods exist, such as selection methods (e.g.; Copas, 1999; Hedges & Vevea, 1996; Vevea & Hedges, 1995) or approaches such as *p*-curve (Simonsohn, Nelson, & Simmons, 2014) and *p*-uniform (Van Assen, Van Aert, & Wicherts, 2015). However, researchers do not often use the PEESE approach, presumably due to its recent development. Also, selection methods are much less frequently used due to their computational complexity and the difficulty of their application (Jin, Zhou, & He, 2015; Peters, Sutton, Jones, Abrams, & Rushton, 2007; Rothstein, Sutton, & Borenstein, 2005), and one of the conditions for the application of the *p*-curve and *p*-uniform approaches is the use of one *p*-value per study. Furthermore, all the aforementioned methods are primarily meant to correct for selection bias, and in this paper we only focus on detection. Therefore, these methods are not studied further in this paper.

In the following section we give more details about the rationale and application of Egger's Regression test, the Funnel Plot test, Begg's Rank Correlation, and the Trim and Fill method. We also propose a multilevel extension of Egger's Regression test and the Funnel Plot test so that they account for dependency among effect sizes within the same study.

Egger's Regression test

In the original version of Egger's Regression test, the standardized effect sizes are regressed on the inverse of their standard errors. If the intercept is statistically significant, this is considered as an indication of publication bias. It has been shown that this equation equals a weighted least squared regression model, weighted by the inverse of the sampling variance, where the standard error of the effect size is introduced as a covariate (Sterne et al., 2000):

$$d_k = \gamma_0 + \gamma_1\sqrt{\sigma_k^2} + r_k \quad (1)$$

d_k represents the effect size reported in study *k*, $\sqrt{\sigma_k^2}$ refers to the (estimated) standard error of d_k, and r_k is a random residual normally distributed with mean 0 and variance σ_k^2. In meta-analysis, the sampling variance σ_k^2 is typically estimated beforehand, and therefore in the meta-analysis itself is considered as known. The term γ_1 is the regression coefficient representing the influence of the standard error on the effect sizes. If this regression coefficient is statistically significant (i.e., *p*-value < .05), then this is considered evidence for publication bias. The intercept, γ_0, can be interpreted as the overall effect if there is no publication bias. This model can be extended in order to account for between-studies heterogeneity by adding a random component, u_k (Sterne & Egger, 2005):

$$d_k = \gamma_0 + \gamma_1\sqrt{\sigma_k^2} + r_k + u_k \quad (2)$$

where u_k is a normal error term with mean 0 and variance σ_u^2 (also written in some studies as τ^2), that represents the residual between-studies variability, that is, the variability that is not due to sampling variance, $\sigma_{r_k}^2$, but rather refers to systematic variance between population effect sizes across studies.

In the scenario in which multiple effect sizes are reported within studies, an additional random residual can be added in order to account for within-study variability, that is, the variability in population effect sizes for outcomes that belong to the same study (Van den Noortgate et al., 2013, 2015). Effect sizes reported in the same study are dependent, and if this dependency is ignored (by fitting the two-level model in Equation 2), then the information contained in the data could be overestimated, which in turn would lead to the underestimation of the standard error of the pooled effect and of the second level predictors (Moerbeek, 2004). In this case, the second level predictor is the standard error of the observed effect sizes. Therefore, the Type I error rate when detecting selection bias would be inflated. Another reason why it is important to extend the model is that if a traditional random effects model (Equation 2) was applied, then a study reporting multiple effect sizes would get a much larger weight on the estimation of the pooled effect, whereas if a three-level model is used, each study contributes to the pooled effect with only a specific study-mean, and the assigned weight depends on how dependent the effect sizes reported in that study are. For all these reasons, it is important to extend this method to a three-level model if primary studies include multiple effect sizes. The model including the random component accounting for the within-study variance is as follows:

$$d_{jk} = \gamma_{00} + \gamma_1\sqrt{\sigma_{jk}^2} + r_{jk} + v_{jk} + u_{0k} \quad (3)$$

where d_{jk} represents the j^{th} observed effect size in study k and v_{jk} is a normal error term with mean zero and variance σ_v^2, (within-study variance). As mentioned before, if γ_1 is statistically significant, selection bias could exist. Previous simulation studies (Macaskill et al., 2001; Peters et al., 2006; Sterne et al., 2000) have shown that Egger's Regression test (based on Equation 1) led to inflated Type I error rates when the population effect size (in their case, an odds ratio) increased, when the number of studies included in the meta-analysis was large or when effect

sizes were heterogeneous. Another simulation study focused on standardized mean differences (Kromrey & Rendina-Gobioff, 2006) reported similar results and furthermore showed that the power was in general low. Based on these results, we also do not expect good performance for this three-level version of Egger's Regression test, although we expect it will perform better in terms if Type I error than the two-level version if multiple effects are reported within studies.

Begg's Rank Correlation test

Begg and Mazumdar (1994) proposed to test the relationship between the standardized effect sizes and their corresponding sampling variances using Kendall's Tau correlation. Kendall's Tau correlation is a non-parametric analysis that is not based on assumptions including the independence of effect sizes, and therefore there is no need to adapt the approach for multilevel meta-analyses. Yet, we use an adapted notation to account for the scenario where multiple outcomes are reported within studies. Although no extensions are proposed for this method, the performance of Begg's Rank Correlation test is explored in this study because Kromrey and Rendina-Gobioff (2006) found that this method had a good control on Type I error rates, so we aim to study to what degree this also holds in this context of multiple within-study effect sizes. The standardized effect sizes (d_{jk_*}) are calculated according to the following formula:

$$d_{jk_*} = \frac{d_{jk} - d_.}{\sqrt{\sigma^2_{jk_*}}} \tag{4}$$

where the d_{jk} represents the j^{th} effect size of study *k*, $d_.$ stands for the estimated pooled effect size using a fixed effect model, and $\sigma^2_{jk_*}$ is the sampling variance of the standardized effect size *j* in study *k*, that is calculated by:

$$\sigma^2_{jk_*} = \sigma^2_{jk} - \left(\sum \frac{1}{\sigma^2_{jk}}\right)^{-1} \tag{5}$$

where σ_{jk}^2 is the variance of the non-standardized effect sizes. If the Kendall's Tau correlation between d_{jk_*} and $\sigma_{jk_*}^2$ is statistically significant, this is an indication of publication bias. To know whether a given Kendall's Tau correlation is significant or not, a normalized test statistic can be calculated (see, e.g., Begg & Mazumdar, 1994). Following the notation used by Kromrey and Rendina-Gobioff (2006), we will refer to this method as Begg's Rank Correlation (V).

Begg and Mazumdar (1994) also proposed to calculate and test the Kendall's Tau correlation between the standardized effect size and the sample size on which it is based (N_k). We will refer to this approach as Begg's Rank Correlation (N). Sterne et al. (2000), found that the power of Begg's Rank Correlation (V) was rather low, especially if there were fewer than 10 studies in the meta-analysis. Macaskill et al., (2001) showed that Begg's Rank Correlation (V) became liberal as the population effect deviated from zero, while in the same condition Begg's Rank Correlation (N) became more conservative. Conversely, Begg's Rank Correlation (V) led to higher power rates compared to Begg's Rank Correlation (N), especially if the number of studies was large and the population effect was low. Kromrey and Rendina-Gobioff (2006) found that the Type I error rate of Begg's Rank Correlation (N) was close to the nominal value, except for large population effect sizes, whereas Begg's Rank Correlation (V) yielded too many Type I errors.

Funnel plot test

Macaskill et al. (2001) proposed to use sample sizes as a predictor of the effect sizes:

$$d_k = \gamma_0 + \gamma_3 N_k + r_k \quad (6)$$

In this study, this method has been extended in the same way as Egger's Regression test, that is, two random effects have been added to account for the within-study and between-studies variances in order to prevent for inflated Type I error rates:

$$d_{jk} = \gamma_{00} + \gamma_3 N_{jk} + r_{jk} + u_{0k} + v_{jk} \quad (7)$$

The variances of $v_{jk,}$ and u_{0k} are the within-study variance σ_v^2, and the between-studies variance σ_u^2, respectively. Macaskill et al., (2001) found in a simulation study that the standard version of this method (Equation 6) outperformed Begg's Rank Correlation and Egger's Regression test in terms of Type I error rate control, although its power was limited especially when the number of studies included in the meta-analysis was small. Similar results were obtained by Kromrey and Rendina-Gobioff (2006): the Funnel Plot test had a conservative Type I error rate in most of the conditions, and the power was low but increased as the population effect size increased.

Trim and Fill method

The Trim and Fill method not only allows for detecting publication bias but also offers the possibility to estimate a new overall effect size corrected for publication bias. This re-estimated effect size then can be used to evaluate again the publication bias, and to obtain a new estimated effect size, and so on. However, only the first step of this iterative method will be implemented in this study, because the focus in this paper is on bias assessment, and not on correction for bias. This same procedure has been followed in previous simulation studies (Ferguson & Brannick, 2012; Kromrey & Rendina-Gobioff, 2006), and, moreover, the adjusted estimate (corrected for publication bias) obtained through the iterative procedure is not always precise (Peters et al., 2007), especially under conditions of high heterogeneity (Terrin et al., 2003), and should be used only to test how sensitive the results are to the possible presence of publication bias.

The first step of the Trim and Fill method consists of estimating the number of studies suppressed due to publication bias. Duval and Tweedie (2000a, 2000b) talk about 'suppressed studies' to refer to relevant studies that are not available due to publication bias and estimate this number, k_0, by assuming that these suppressed studies are the ones that report more negative

results, that is, the ones that found non-significant differences or differences in the unexpected direction. In order to estimate the suppressed studies, Duval and Tweedie (2000a, 2000b) proposed two estimators of k_0, namely R_0 and L_0. For estimating R_0, the following steps have to be followed: first a rank (r_k) has to be assigned to the absolute difference between a given effect size d_k and the estimated pooled effect size, $d_.$:

$$r_k = |d_k - d_.| \tag{8}$$

A rank of 1 indicates the lowest absolute difference between d_k and $d_.$. Second, a negative sign is assigned to the ranks that refer to effect sizes smaller than the estimated pooled effect ($d_k < d_.$). Third, we should look for the absolute value of the lowest rank, r_{lowest}. The estimator R_0 is then calculated with the following formula:

$$R_0 = \theta^* - 1 \tag{9}$$

where $\theta^* = m - r_{lowest}$ and m is the total number of observed effect sizes within the dataset. For estimating estimator L_0, the following equations have to be applied:

$$L_0 = \frac{[(4T_m) - m(m+1)]}{2m-1} \tag{10}$$

$$T_m = \sum_{(d_k - d_.)>0} r_k \tag{11}$$

where T_m is the Wilcoxon rank test statistics for the observed m values. Following Duval and Tweedie (2000a, 2000b) procedure, we will refer to R_0^+ and L_0^+ as the R_0 and L_0 estimators in which negative values are truncated to 0. In this context where there are multiple effect sizes reported in each study, the unit of this analysis shifts from studies to effect sizes, meaning that R_0^+ and L_0^+ actually represent the number of suppressed effect sizes and not the number of suppressed studies. These suppressed effect sizes can belong to studies that were never available for the meta-analyst, or to published studies where authors selectively reported the most salient effect sizes while hiding effect sizes that were non-significant. Kromrey and Rendina-Gobioff

(2006) considered the existence of publication bias when the number of estimated suppressed studies was above 3 (i.e., $R_0^+ > 3$), regardless of the number of observed studies. Regarding estimator L_0^+, there is not a clear recommended cutoff, as in the simulation of Duval and Tweedie (2000a) it was found that the null hypothesis of $k_0 = 0$ was rejected under different values of L_0^+ and R_0^+. Therefore, a challenge of this paper is to find plausible thresholds for the L_0^+ estimator. The simulation study of Kromrey and Rendina-Gobioff (2006) concluded that the Trim and Fill (R_0^+) method yields conservative Type I error rates and low power.

The main goal of this study is to explore by means of a simulation study the performance of (an extension of) the traditional tests for detecting publication bias when primary studies include multiple effect sizes and when different types of selection bias may exist. This study explores scenarios that have never been studied before. First, all previous simulations have only considered situations where only one effect size is reported per study. Second, previous studies have never explored the performance of these methods under conditions where selective outcome reporting bias occurs, or where both publication and selective outcome reporting bias occur together. Third, this study is also the first that explores the performance of the three-level version of the Funnel Plot test and Egger's regression test when multiple outcomes are included within studies. We consider it essential that applied researchers have knowledge of the properties of these methods when they apply (and extension of) these techniques in meta-analysis where primary studies include multiple effects.

To achieve this goal, we will follow the next steps: first, we will describe the consequences that different types of selection bias have on the estimated pooled effect size. Next, new plausible regions for the estimator L_0^+ will be examined in order to find appropriate thresholds from which the researcher can conclude that selection bias exist. Third, we also aim to

find the (combination of) method(s) that lead to an adequate control of Type I error rates, while still leading to an acceptable power. This approach of combining tests has been actually proposed before by Ferguson and Brannick (2012), although they chose three specific tests, namely the Fail Safe Number, the Egger Regression test and the Trim and Fill, while in our proposal we want to know how many of these methods have to detect the presence of publication bias so the likelihood of a false positive is 10% or below. To that end, we will count how many of these six methods suggest the presence of selection bias, and explore whether we can define a cutoff number to decide on publication bias. The approach of combining several methods to detect publication bias has received some criticisms (Rothstein & Bushman, 2012) because different methods assume different types of publication bias and therefore have different aims. However, it may be the case that researchers want to detect the presence of any type of bias (either small study effects, either the amount of suppressed studies) with the sole purpose of being cautious when interpreting results, so that is why we think it is worth it to explore this option. A fourth step will be to give a general description of how the estimated Type I error rate and power of each of the methods depends on characteristics of the design. Finally, we will describe which method works better under each combination of conditions, and formulate general conclusions.

Method

Data generation

Data were generated in three steps. First, complete meta-analytic datasets (i.e., without selection bias) were created. These datasets consisted of standardized mean differences (Cohen's *d*), that were simulated based on the following meta-analytic three-level model, with normally distributed residuals (Van den Noortgate et al., 2013, 2015):

$$d_{jk} = \gamma_{00} + r_{jk} + v_{jk} + u_{0k} \quad (12)$$

where γ_{00} is the overall effect size, and the other elements have been previously described. Afterwards, Cohen's *d* (and their sampling variances) were corrected using a small sample correction factor, resulting in Hedges' g (Hedges, 1991). In a second step, we derived for each full dataset six datasets in which selection bias was present. To that end, we deleted a percentage of effect sizes based on their *p*-values, that is, the larger the *p*-value, the larger the probability of deletion. These six datasets differ from each other in the pattern of bias (3 levels), and in the degree of bias (2 levels). In the first pattern of selection bias generated, the likelihood of an effect size being reported in a study depended on its *p*-value, which is the so-called selective outcome reporting bias. In line with previous simulation studies (Begg & Mazumdar, 1994; Kromrey & Rendina-Gobioff, 2006), the probability of inclusion was calculated using the following weight function:

$$P\left(g_{jk}\ is\ included\right) = \exp\{-bp_{jk}{}^{a}\} \quad (13)$$

where p_{jk} is the two-sided *p*-value associated to the effect on outcome *j* in study *k.* In conditions with a small degree of selection bias, $a = 1.5$ and $b = 4$, whereas in conditions with large selection bias, $a = 3$ and $b = 4$. For the second selection bias pattern, full studies were suppressed based on the *p*-value of their mean effect size. To that end, the mean differences of all within-study outcomes were averaged and the *p*-value of the averaged mean difference was calculated. The probability of inclusion was obtained with a similar weight function:

$$P(study\ k\ is\ included) = \exp\{-bp_{.k}{}^{a}\} \quad (14)$$

where $p_{.k}$ is the two-sided *p*-value associated to the mean effect size of study *k* and the same sets of *a* and *b* values as described above were used. This type of bias corresponds to the standard conceptualization of publication bias, as full studies were removed from the meta-analysis. For the third selection bias pattern, both types of biases were combined: some full studies were

suppressed, and then some effect sizes from the remaining studies were removed as well. For all derived datasets, classical techniques for detecting the presence of publication bias were applied, and the power of these methods was assessed by calculating the proportion of meta-analyses in which the methods correctly detected selection bias. However, it is also important to know how likely it is that any of these methods spuriously detected the presence of any selection bias (i.e., the Type I error rate), by applying these methods on unbiased datasets. To make a fair comparison between the biased and unbiased datasets in terms of power, we had to make sure they had the same size. To that end, a third step was to calculate how many studies and/or effect sizes were removed in the derivation of the biased datasets and then the same amount of full studies and/or effect sizes were deleted completely at random from the original complete dataset. This means that for each initially generated full data, twelve (six biased and six unbiased) datasets were derived.

Conditions

On top of the existence of bias, the different patterns of selection bias and the size of selection bias (as described above), various other factors were manipulated. The number of studies (*k*) within the complete meta-analytic dataset (i.e., without selection bias) could be 15, 30 or 70. These values are approximately equal to the minimum, median and maximum number of studies typically included in meta-analyses in the field of psychology (Rubio-Aparicio, Marín-Martínez, Sánchez-Meca, & López-López, 2017). Because three effect sizes were generated within each study, the total amount of effect sizes could be 45, 90 or 210, respectively. The total number of subjects included in each primary study (*n*) was a number extracted from a normal distribution with mean 50, 100 or 150 and standard deviation of 30. These three mean sample sizes were selected because according to the systematic review of Rubio-Aparicio et al. (2017), the minimum mean sample size reported by primary studies is 17 and the maximum is 211,

which makes an average of 114. Therefore, the condition where the mean sample size is 100 represents standard primary studies, while the conditions where the mean sample size is 50 and 150 represent primary studies where the mean sample size is below and above the average, respectively. A standard deviation of 30 was selected in order to generate variability in the sample sizes of the primary studies, because this variable is the one introduced directly as a predictor in the Funnel Plot test and indirectly in Egger's Regression test. The population effect size could equal 0, 0.2, 0.5 or 0.8, which corresponds to a null, low, medium and high effect according to Cohen's benchmarks (Cohen, 1988). The between-studies (σ_u^2) and the between-outcomes (σ_v^2) variance could both equal 0.01, 0.06, or 0.11, so the total systematic variance was 0.02, 0.12 or 0.22, generating situations of low, medium or high variability. The selection of these values for the variance components is also based on the results of the revision of Rubio-Aparicio et al. (2017), who found that the median between-studies variance of meta-analyses in the field of psychology was 0.11, the first quartile was 0.06 and the third quartile was 0.18. The combination of these conditions led to a total of 3 x 2 x 2 x 3 x 3 x 4 x 3 = 1,296 conditions, and for each combination we simulated 1,000 datasets, so in total 1,296*1,000 = 1,296,000 datasets were simulated.

Note that inducing any of the three types of selection bias affected the number of studies within the meta-analysis (*k*), the number of outcomes within studies, the value of the population effect size, and the value of the variance components (σ_u^2 and σ_v^2). For the unbiased counterparts, also the number of studies and the number of outcomes within studies was affected, but not the expected mean population effect size and variances. In Table 2, the average or median of the new values obtained after the generation of the three types of selection biases is shown. Thus, in the upcoming sections when we refer to the condition where 15 (45 effect sizes), 30 (90 effect sizes)

or 70 studies (210 effect sizes) comprised the meta-analysis, the actual number of studies and/or effect sizes may be smaller, whereas the estimated pooled effect of the biased datasets is expected to be larger than the population effect. Regarding the between-studies and between-outcomes variance, it has been shown that the between-studies variance can either decrease or increase when publication bias is present (Jackson, 2006).

Evaluation of the simulation results

Each of the methods explored in this study was evaluated by means of Type I error rate (α) and power $(1-\beta)$. The Type I error rate indicates the proportion of meta-analyses in which the presence of any selection bias is spuriously detected. Following the recommendation of Macaskill et al., (2001), we set the nominal Type I error rate value to .10. We used a two-sided test of significance for all methods involving traditional significance tests because this is the default option implemented in the package *metafor* in R. In order to know which estimates of the Type I error rate values indicate a deviation from the nominal value, 95% confidence intervals were constructed around the nominal value. The standard error of α was calculated using the formula $SE(\alpha) = \sqrt{[\alpha(1-\alpha]/I}$, where *I* is the number of iterations and α = .10. Assuming normality, a 95 % confidence interval for the Type I error rate ranges from .0814 to .1185 (.10 ± .0094 * 1.96). Values outside this range were considered too low (conservative) or too high (liberal). Notice that for the Trim and Fill method, it is not possible to set a nominal Type I error rate because the decision of whether any selection bias exists is not based on a *p*-value. In a previous simulation study, Kromrey and Rendina-Gobioff (2006) considered the existence of publication bias if the number of estimated suppressed studies was above 3 (i.e., $R_0^+ > 3$), because Duval and Tweedie (2000a) showed that for the region $R_0^+ > 3$, the power was around .80 when the proportion of suppressed studies was larger than 7. In our study, we have

considered the same region for estimator R_0^+, that in this context represents the number of suppressed effect sizes. The power refers to the proportion of meta-analyses for which selection bias is correctly detected. We consider a statistical power of .80 or larger as adequate (Cohen, 1988).

As mentioned in the previous section, this study has three goals. For the first goal, exploring plausible regions for the L_0^+ estimator of the Trim and Fill method, a Receiver Operating Characteristic (ROC) curve was applied. A ROC curve is a graphical representation that shows the specific Type I error (i.e., False Positive rate) and power (i.e., Sensitivity) for each value of a given classifier. In our context, the results of the ROC analyses allowed us to know which values of the L_0^+ estimator (our classifier) led to an acceptable Type I error rate. The ROC curve analyses were applied separately on the 1,000 datasets generated for each combination of conditions, and in each analysis we selected the value of L_0^+ that led to a Type I error closest to .10 but without exceeding the upper limit of .1185. The results of these analyses were summarized in a dataset, where a specific L_0^+ cutoff value was recommended for each combination of conditions. Afterwards, in order to give general recommendations, we ran an analysis of variance (ANOVA), introducing the simulation design factors as predictor variables for the estimated L_0^+ thresholds.

For the second goal, that consists in exploring how many methods have to detect selection bias to be certain that selection bias is not spuriously detected, we summed up the six variables that indicated whether each method had detected the presence of selection bias (1) or not (0). The total sum led to scores from 0 (no method detected the presence of selection bias) to 6 (all methods detected the presence of selection bias). Afterwards, the Type I error and power was calculated for each of these five new criteria. For the third goal, ANOVAs were performed for

each method to detect which simulation design factors affected the power and Type I error rate to a larger extent, i.e., showing the largest eta-squares. Finally, for describing which method is most suited for each combination of conditions, we calculated the average Type I error rate and power across the three types of selection bias. Afterwards, for each combination of conditions we selected the method that showed an adequate mean Type I error rate and described its power (Bradley, 1978). In the case that two or more methods showed an adequate Type I error rate control, we selected the one that exhibited larger power.

SAS was used to generate and analyze the data. The SAS code for the Funnel Plot test, Egger's Regression test and the Trim and Fill (R_0^+ and L_0^+) method, and the R code of all methods are available upon request from the first author. For Begg's Rank Correlation, the PUB_BIAS SAS macro (Rendina-Gobioff & Kromrey, 2006) can be used.

Results

Effect of selection bias on the estimated pooled effect size

In Table 1, the mean percentage of effect sizes selected for inclusion in the meta-analysis is shown for each type of selection bias, and in Table 2 the mean pooled effect size obtained in the biased datasets is presented. The percentage of included effect sizes was notably lower when selection bias was large and when both publication and selective outcome reporting bias were induced. Therefore, it is not surprising that under these circumstances the estimated pooled effect size was more inflated. However, when the population effect size value was 0, the estimated pooled effect was barely overestimated despite the lower percentage of effect sizes included in the meta-analysis. The largest overestimation of the pooled effect size occurred when the population effect size was 0.2, whereas when the population effect size was larger, the estimated pooled effect was less inflated.

Cutoffs for Trim and Fill L_0^+ estimator

Based on the ROC analyses, we selected for each combination of conditions the value of the L_0^+ estimator that resulted in a Type I error rate as close as possible to the nominal value. The ANOVA done over these L_0^+ cutoffs, with the simulation design factors as independent variables, showed that, regardless of the types of bias, the population effect size (η^2 = .254), the number of effect sizes (η^2 = .471) and its interaction (η^2 = .074) had the largest impact on the L_0^+ cutoffs values, whereas the median number of participants (η^2 = .042), the amount of variance (η^2 = .021), the degree of bias (η^2=.002), and the type of selection bias (η^2=.004) had smaller effects. As can be seen in Table 3, the more effect sizes included in the meta-analysis and the larger the population effect, the larger the optimal cutoff value of L_0^+. These thresholds were applied to calculate the Type I error rate and the power of this method. For instance, for conditions where the population effect size was 0.2 and the number of initial effect sizes was 90, we decided that there was selection bias if the value of L_0^+ was larger than 2, whereas in conditions where the population effect size was 0.8 and the number of initial effect sizes was 210, we concluded that there was selection bias if the value of the L_0^+ was larger than 6. The performance of this method is discussed in the two following sections, together with the Type I error rate and power of the other methods.

Aggregation of methods

When the evidence from all methods for detecting selection bias were put together, we found that if four out of the six methods detected the presence of any selection bias, the Type I error rate was below .10 in all conditions. When three out of the four methods detected the presence of publication bias, the Type I error rate was still too high in conditions of medium and high heterogeneity and large population effect size. In the following section we will refer to this method as the Four-tandem procedure

Effect of the simulation design factors on the mean Type I error rate and power

Table 4 shows the Type I error rate for each method and simulated condition. Eta squared is given only for main effects. Interaction terms were not included because they just showed that the same pattern was visible in all conditions, but to a different degree. The Funnel Plot test, the Trim and Fill (R_0^+) method, and the Four-tandem method were the only procedures that exhibited a mean Type I error rate within the recommended cutoffs. In these methods, the 61.11%, 100%, and 98.77% of the conditions exhibited, respectively, an appropriate Type I error rate. Egger's Regression test and Begg's Rank Correlation (V) had an average Type I error rate around .37, and the 18.36% and the 1.08% of the conditions, respectively, showed a Type I error rate of .10 or below. The Trim and Fill (L_0^+) method and the Begg Rank Correlation (N) exhibited a mean Type I error rate between .10 and .20, and the 40.59% and the 3.24% of the conditions exhibited an adequate Type I error rate.

Looking closer at the Type I error rate in each level of the simulation design factors, we can see that the estimates yielded by Egger's Regression test and by Begg's Rank Correlation (V) were mostly affected by the population effect size value and by the mean sample size of primary studies. These two methods led to too many Type I errors when the population effect size and the number of subjects of primary studies were large. Furthermore, the variability among the Type I error rates was very high. For instance, when the population effect size was 0, the Type I error rate was .117 and .167 for Egger's Regression and Begg's Rank Correlation (V) respectively, whereas when the population effect size was 0.8, the Type I error rate reached values as large as .672 and .640, respectively. Also for intermediate population effect sizes, Type I error rates were not within the recommended cutoffs. The Type I error rates derived from the Funnel plot test were not largely influenced by any of the simulation design factors, and all Type I error rates were within the acceptable range. On the other hand, Begg's Rank Correlation (N) led to liberal

Type I error rates in all conditions, but their variability was low: all the Type I error rates ranged between .14 and .19.

The heterogeneity among effect sizes was the simulation design factor that had the largest influence on the Type I error rate of this method: the more variability, the larger the Type I error rate. The Trim and Fill (R_0^+) method became increasingly liberal with larger population effect sizes and with larger number of studies conforming the meta-analysis, although in none of these conditions the Type I error rate exceeded .05. The Type I error rates yielded by the Trim and Fill (L_0^+) method and the Four-tandem procedure were only affected by the mean sample size, although in opposite directions: the larger the mean sample size, the lower the Type I error rates given by the Trim and Fill (L_0^+) method, whereas the Type I error rate of the Four-tandem procedure test increased when the mean number of subjects in primary studies became larger. In addition, whereas all Type I error rates of the Four-tandem procedure were below .07, the Trim and Fill (L_0^+) method led to too large Type I error rates when the total heterogeneity among effect sizes was almost inexistent, when the population effect size was larger than 0.2, when the number of effect sizes was larger than 90 and when the mean sample size of primary studies was 50.

Table 5 shows the statistical power for each simulated condition and method. The methods that exhibited a larger mean power were the Begg Rank Correlation (V), Egger's Regression test, and the Trim and Fill (L_0^+) method, whereas the Trim and Fill (R_0^+) method had a power below .10, and the Funnel Plot test and the Four-tandem procedure exhibited an average power of .11. The 23.61% of the conditions of the Egger's Regression test, 16.98% of the conditions of Begg's Rank Correlation (V) and 10.03% of the conditions of the Trim and Fill (L_0^+) method showed a power of .80 or larger. However, less than 1% of the conditions of the

Funnel Plot test, Begg's Rank Correlations (N), Trim and Fill (R_0^+) and Four-tandem procedure yielded a power equal or larger than .80.

The simulation design factor that had the largest effect on the average power of almost all methods was the population effect size value. The power of Egger's Regression test, Begg's Rank Correlation (V), the Trim and Fill (R_0^+) method and increased when the population effect size was large (0.8). The value of the population effect size also had an influence on the power of the Trim and Fill (L_0^+) method, but in this case the largest power was observed when the population effect size was 0.2. Nevertheless, there were huge differences in the estimated mean power across these methods. For instance, whereas Egger's regression test and Begg's Rank Correlation (V and N) yielded large power rates (>.7) when the population effect size was large (0.8), the power of the Trim and Fill (R_0^+) method within the same condition was below .12. The power estimates of the remaining methods, namely the Funnel Plot test, the Begg Rank Correlation (N) and the Four-tandem procedure, were unaffected by any of the simulation design factors. The power estimates of the Four-tandem procedure and of the Funnel Plot were extremely low and never exceeded .14, while Begg's Rank Correlation (N) showed slightly higher power estimates, but still did not reach values above .22.

Under which conditions did each method perform better?

As described in the previous section, there are some methods that showed systematically unacceptable Type I error rates across all simulated conditions, such as Begg's Rank Correlation (V and N), or a really low mean power, like the Trim and Fill (R_0^+) method. However, some methods that on average showed inadequate Type I error rates, did show appropriate a Type I error rate under specific combination of conditions, like, for instance, Egger's Regression test and the Trim and Fill (L_0^+) method. On the other hand, the Trim and Fill (R_0^+) method, the Funnel Plot

test and the Four-tandem procedure lacked power, but these methods also showed an increasing power under certain combinations of simulated conditions[1]. Based on the mean Type I error rates and power across the three conditions of selection bias, we have elaborated the classification shown in Table 6. In this table, we indicate which method is better to apply under each combination of conditions based on their mean power, given, first, an adequate mean Type I error rate. Practically no method had a condition where the Type I error rate was adequate and where, at the same time, the power was .80 or above. The only exception was the Trim and Fill (L_0^+) method and Four-tandem procedure, in which less than 1% of the conditions exhibited a Type I error rate equal or below to .10 and a power equal or above .80.

As shown, the Trim and Fill (L_0^+) method and the Four-tandem procedure had the highest power in conditions where there was medium or high variability among effect sizes and the population effect size was large, or when the population effect size was moderate and the number of effect sizes included in the meta-analysis was small. Egger's Regression test was the method that led to a larger power when the population effect size was zero and there was not variability between effect sizes, or when the population effect size was larger and the mean sample size of primary studies was small. The Funnel Plot test, the Four-tandem procedure and the Trim and Fill (R_0^+) method can be used in any combination of conditions, because they always showed adequate Type I error rates. However, the Trim and Fill (R_0^+) method showed a better power when there was no variability among effect sizes and the population effect size was large, while the Four-tandem procedure showed in general better power when the population effect size was 0.5 or higher, and the mean sample size of primary studies was low. The Funnel Plot test did not lead to power estimates above .10, so this method is only recommended in those combination of

[1] The tables with the Type I error rate and power disaggregated by each method and combination of conditions are available upon request from the first author.

conditions where the Trim and Fill (R_0^+) method or the Four-tandem procedure showed a power below .10. Yet, it is important to keep in mind that the power of all methods was, in general, very low: the power of the methods in the cells of Table 6 that are not written in bold or do not contain an asterisk have a mean power lower than .3.

Discussion

The aim of this study was to explore the performance of the classical methods for detecting publication bias in the common situation where primary studies include multiple effect sizes and different patterns of selection bias exist. Before describing the performance of these methods, we have shown how the type of bias induced and the value of the population effect size affect the pooled effect size estimate. For instance, we found that selective outcome reporting bias had a larger influence on the overestimation of the pooled effect size compared to the effect of publication bias. When publication bias was generated, the selection process was based on the study-mean effect size and not on individual effect sizes, so it was more likely that studies selected for inclusion contained some non-significant effect sizes, leading to a smoother bias and to a more symmetrical funnel plot. In contrast, when selective outcome reporting bias was generated, the effect sizes were selected as a function of their *p*-values. Thus, large *p*-values were much less likely to be included, making the effect of the bias much stronger. Regarding the value of the population effect size, it was found that when the population value was zero, the estimated pooled effect was barely inflated despite the large percentage of suppressed effect sizes in that condition. This happened because the censoring process occurred in the center of the funnel plot, where all effect sizes were non-significant, and therefore large negative values below 0 and large positive values above 0 were selected for inclusion in the meta-analysis. This situation led to a symmetrical funnel plot that was almost empty in the center and to an estimated mean effect size close to the real value. In conditions where the population effect size was 0.5 or 0.8, the pooled

effect size was barely overestimated because most of the effect sizes were selected for the meta-analysis: when the population effect size was 0.8, the percentage of included effect sizes was almost 100% so no overestimation of the pooled effect was expected. In contrast, when the population effect size was 0.2, the percentage of included effect sizes was much lower and hence a large overestimation of the pooled effect size was predicted.

We can relate these results to the performance of the methods for detecting selection bias. For instance, all methods showed an increasing power when the total number of studies or effect sizes was high. It is well known that a good way to increase the power is to use a larger number of subjects. Therefore, it is logical that the probability of finding a significant effect was larger in conditions where many effect sizes were included in the meta-analysis. In this line, we have also found that the power of most methods increased for larger population effect size values. Following the same reasoning as before, in conditions with larger population effect size values, there were fewer studies or effect sizes censored and hence more effect sizes were available for the analysis, increasing the power. However, it might appear counterintuitive that Egger's Regression test and Begg's Rank Correlation (V) showed more power in conditions where the population effect size was 0.8, because in this condition there was a small effect of the selection bias. In other words, if the population effect size was large, most effect sizes were included in the meta-analysis and therefore the effect of selection bias was expected to be less pronounced and hence more difficult to detect. Nevertheless, the Type I error rate of these methods in the condition where the population effect size was 0.8 was also quite large, so the high power observed might be just a by-product of the complementary scenarios that lead to too many Type I errors also being observed. Another common pattern observed in all methods was the low estimated power for detecting selection bias when the population effect size was zero, which is

not surprising given that the distribution of effect sizes remained symmetric because effect sizes were only suppressed in the center of the funnel plot.

Besides the general influence that the population effect size value and the number of effect sizes had on the estimated Type I error rate and power, some simulated conditions affected the Type I error rate and power of some methods to a larger extent. For instance, and in line with previous simulation studies (Kromrey & Rendina-Gobioff, 2006; Macaskill et al., 2001; Peters et al., 2006; Sterne et al., 2000), Egger's Regression test and the Begg Rank Correlation (V) became highly liberal when the population effect size, the number of studies, the mean sample size of primary studies and the heterogeneity among effect sizes were larger.

According to our results, Begg's Rank Correlation (N) also led to Type I error rates above the recommended cutoff, whereas Kromrey and Rendina-Gobioff (2006) found that this method resulted in controlled Type I error rates in most of the conditions. This contradictory result might be due to the different amount of between-studies variance generated: while Kromrey and Rendina-Gobioff (2006) did not manipulate the value of the between-studies variance (and was apparently set to zero, although they do not mention it), we generated conditions of small, medium and high heterogeneity. The Type I error rate for this method in condition with small variability among effect sizes was 0.14, which is close to an adequate Type I error rate, and the Type I error rate in conditions of high heterogeneity was almost 0.20. Therefore, our results show that Begg's Correlation Test (N) tends to work better under conditions with small heterogeneity among effect sizes, which goes actually in line with the results showed by Kromrey and Rendina-Gobioff (2006).

The Type I error rates of the Funnel Plot test and the Trim and Fill (R_0^+) method were less influenced by the simulated conditions and were in general within the recommended thresholds. However, as previous simulations have also shown, these methods lacked power (Kromrey &

Rendina-Gobioff, 2006; Macaskill et al., 2001), although power increased in conditions with low heterogeneity, large population effect sizes, and larger numbers of studies and effect sizes. The Trim and Fill (L_0^+) method became too liberal when effect sizes did not vary among outcomes and studies and when the mean sample size of primary sizes was too small. However, one advantage of this method is the good trade-off between Type I error rate and power: in conditions where the Type I error was controlled, this method led to a power notably higher than the power associated with any other method. The Four-tandem procedure proposed in this study worked well in terms of Type I error rates but power was still insufficient. In summary, none of these methods performed systematically well on terms of Type I error rates, and when they did, the power was in general low. Less than 1% of the conditions of the Trim and Fill (L_0^+) method and of the Four-tandem procedure showed adequate Type I error rates and power above .80, whereas none of the other methods had even one condition where both Type I error rate and power were adequate.

The most important limitation of this study is that the conclusions can be generalized only to the conditions simulated. Although we have tried to select representative values for the simulation, the characteristics of a certain meta-analysis might not fit into any of these combinations. For instance, the simulation study only focuses on standardized mean differences (Hedges' *g*) so it is not possible to know whether the conclusion from this study can be generalized to other effect sizes, such as Pearson correlation coefficients or odds ratio. Also, we simulated three patterns of selection bias, but other patterns (i.e., the selection of only one *p*-value per study) or a mix of patterns are possible. Furthermore, we have fixed the number of effect sizes per study to be three, but this might not be a realistic assumption, as some meta-analyses might have a more unbalanced structure (e.g., some primary studies may include only 1 effect size while others report 10 effect sizes) or might have a different number of total effect sizes (less than 45 or more than 210). This will probably affect the results for the L_+^0 estimator of the Trim

and Fill method, because the optimal thresholds are likely to change if there is a different number of total effect sizes in the meta-analysis. Along the same lines, it is important to consider that for selecting a method to detect selection bias, we should look at both the number of studies and the number of total effect sizes for applying Egger's regression test and the Funnel Plot test. However, for use of the Begg's Rank Correlation (N and V) and the Trim and Fill method (R_+^0 and L_+^0), the researcher should solely focus on the total number of effect sizes, because in these methods the nesting of effect sizes within studies is irrelevant. Another limitation of this study is that it focused on assessing rather than correcting for selection bias. The Trim and Fill method can however also be applied to correct for bias (and therefore to test how sensitive the results are to the possible presence of publication bias), although previous research has shown that results are not always accurate (Peters et al. 2007, Terrin et al., 2003). Future work has to be done on how this and other methods for correcting for publication bias (e.g., the PEESE method, selection methods, the *p*-curve method and the *p*-uniform method) can be adapted and applied in the multilevel context.

A final remark is that it is important to remember that this study, as well as previous research focused on this topic, would have not been necessary if publication or selective outcome reporting bias did not exist. Ideally, publication and selective reporting bias is avoided so that its detection and correction are no longer needed. In this regard, several suggestions have been given (e.g., Ioannidis et al., 2014; Thornton & Lee, 2000), such as the creation of registries where researchers can upload their ongoing studies regardless of the results or the change of editorial policies towards a system that encourage editors to publish studies based on their quality and not on their results. So far, and until all these changes have been carried out, we recommend applied researchers to be very careful in using these methods and we encourage researchers to explore the performance of other existing methods to detect and correct for selection bias.

Conclusion

From the results of this simulation study several general conclusions can be extracted. First, we do not recommend the use of Begg's Rank Correlation (V and N) in multilevel meta-analysis because there were only a few simulated conditions in which the Type I error rate was within the acceptable range. Regarding the Trim and Fill (L_0^+) method, its use is adequate under certain conditions. For instance, this method worked well in terms of both Type I error rate and power when the population effect size was moderate to large (0.5 to 0.8), variability among effect sizes was medium or high, there were a lot of effect sizes included in the meta-analysis and the mean sample size of primary studies was large. The Funnel Plot test, the Trim and Fill (R_0^+) method and the Four-tandem procedure showed an appropriate Type I error rate control, although the power was quite low. In conclusion, none of these methods work well and they should be used with caution.

Finally, it is important to mention that other techniques can be used for accounting for dependent effect sizes, like the Robust Variance Estimation method (Hedges et al., 2010). Previous simulation research (Moeyaert, Ugille, Beretvas, Ferron, Bunuan, & Van den Noortgate, 2016) have shown that the meta-analytic three level model and the Robust Variance Estimation method perform very similarly, so it is likely that the results obtained from the three-level Egger's Regression test and three-level Funnel Plot test are similar to the ones that would be obtained if the Robust Variance Estimation method was applied instead.

References

Assink, M., & Wibbelink, C. J. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology, 12*, 154-174.

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

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19,* 211-229.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Copas, J. B. (1999). What works? Selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 162*, 95-109.

Decullier, E., Lhéritier, V., Chapuis, F. (2005). Fate of biomedical research protocols and publication bias in France: retrospective cohort study. *British Medical Journal, 331*, 19.

Duval, S., & Tweedie, R. (2000a). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56,* 455-463.

Duval, S. & Tweedie, R. (2000b). A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89-98.

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. *PLoS One*, *8*(7), e66844.

Egger, M., Davey-Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891-904.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*, 120-128.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502-1505.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

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.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*, 299-332.

Hutton, J. L., & Williamson, P. R. (2000). Bias in meta-analysis due to outcome variable selection within studies. *Applied Statistics*, *49*, 359-370.

Ioannidis, J. P. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Journal of the American Medical Association, 279*, 281-286.

Ioannidis, J. P., Cappelleri, J. C., Lau, J. (1998). Issues in comparisons between meta-analyses and large trials. *Journal of the American Medical Association, 279*, 1089-1093.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18,* 235-241.

Jackson, D. (2006). The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine, 25,* 2911-2921.

Jin, Z. C., Zhou, X. H., & He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine, 34,* 343-360.

Kromrey, J., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. *Educational and Psychological Measurement, 66,* 357-373.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20,* 641-654.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39,* 129-149.

Moeyaert, M., Ugille, M., Beretvas, S. N., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2016). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*. Retrieved from http://www.tandfonline.com/doi/pdf/10.1080/13645579.2016.1252189?needAccess=true

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

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association*, *295*, 676-680.

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

Platt, L., Melendez-Torres, G. J., O'Donnell, A., Bradley, J., Newbury-Birch, D., Kaner, E., & Ashton, C. (2016). How effective are brief interventions in reducing alcohol consumption: do the setting, practitioner group and content matter? Findings from a systematic review and meta-regression analysis. *British Medical Journal Open*, *6*, e011473.

Rendina-Gobioff, G., & Kromrey, J. D. (2006). PUB_BIAS: A SAS macro for detecting publication bias in meta-analysis. 14th Annual Southeast SAS Users Group (SESUG) Conference; Atlanta, GA: Southeast SAS Users Group (SESUG).

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: Comment on Ferguson and Brannick (2012). *Psychological Methods*, *17*, 129-136.

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

Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2017). A methodological review of meta-analyses of the effectiveness of clinical psychology treatments. *Behavior Research Methods*, 1-17.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534-547.

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

Sterne, J. A. & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99-110). Hoboken, NJ: Wiley.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53,* 1119-1129.

Tannock, I. F. (1996). False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. *Journal of the National Cancer Institute*, *88,* 206-207.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22,* 2113-2126.

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology*, *53,* 207-216.

Van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20,* 293-309.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45,* 576-594.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods, 47*, 1274-1294.

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.

Table 1.

Average percentage and number of included effect sizes and studies for each type of selection bias.

	Publication bias			Selective reporting			Both types of bias		
	% selec. ESs	ESs	Studies	% selec. ESs	ESs	Studies	% selec. ESs	ESs	Studies
k = 15									
Small									
0	80.0%	36	12	71.1%	32	15	57.7%	26	12
0.2	80.0%	36	12	75.5%	34	15	62.2%	28	12
0.5	93.3%	42	14	84.4%	38	15	82.2%	37	14
0.8	100%	45	15	95.5%	43	15	93.3%	42	15
Large									
0	66.6%	30	10	60.0%	27	14	44.4%	20	10
0.2	73.3%	33	11	64.4%	29	14	51.1%	23	11
0.5	93.3%	42	14	80.0%	36	15	75.5%	34	13
0.8	100%	45	15	91.1%	41	15	91.1%	41	15
k = 30									
Small									
0	76.6%	69	23	72.2%	65	29	58.8%	53	23
0.2	80.0%	72	24	74.4%	67	29	63.3%	57	24
0.5	93.3%	84	28	85.5%	77	30	81.1%	73	28
0.8	100%	90	30	94.4%	85	30	94.4%	85	30

Table 1 (*continued*)

	Publication bias			Selective reporting			Both types of bias		
	% selec. ESs	N ESs	N Studies	% selec. ESs	N ESs	N Studies	% selec. ESs	N ESs	N Studies
Large									
0	70.0%	63	21	60.0%	54	28	45.5%	41	20
0.2	73.3%	66	22	64.4%	58	28	52.2%	47	22
0.5	90.0%	81	27	78.8%	71	29	74.4%	67	27
0.8	100%	90	30	92.2%	83	30	91.1%	82	29
k = 70									
Small									
0	78.6%	165	55	71.9%	151	68	58.6%	123	54
0.2	81.4%	171	57	74.8%	157	69	63.3%	133	57
0.5	92.9%	195	65	85.7%	180	69	81.4%	171	65
0.8	98.6%	207	69	94.8%	199	70	93.8%	197	69
Large									
0	68.6%	144	48	60.5%	127	65	45.7%	96	47
0.2	74.3%	156	52	64.3%	135	66	52.4%	110	51
0.5	90.0%	189	63	79.0%	166	68	74.8%	157	63
0.8	98.6%	207	69	91.9%	193	70	91.4%	192	69

Notes. Small and Large refer to the degree of bias generated; % selec. ESs=percentage of selected effect sizes; N ESs = total number of effect sizes; N Studies= total number of studies; k = initial number of studies; γ_{00}= combined effect size; SE= standard error; σ_u^2= between-studies variance;

σ_v^2 = between-outcomes variance. The between-studies and between-outcomes variances were fixed to 0.06 and the mean sample size (*n*) was fixed to 100. The values 0, 0.2, 0.5, and 0.8 refer to the population effect size.

Table 2.

Mean estimates of the combined effect size and median estimates of the standard error and variance components after the different types of selection bias were generated.

	Publication bias				Selective reporting				Both biases			
k = 15	γ_{00}	SE	σ_u^2	σ_v^2	γ_{00}	SE	σ_u^2	σ_v^2	γ_{00}	SE	σ_u^2	σ_v^2
Small												
0	.000	.096	.075	.053	-.001	.099	.089	.068	-.001	.122	.123	.057
0.2	.237	.087	.062	.053	.244	.092	.076	.061	.287	.105	.088	.052
0.5	.523	.074	.042	.052	.550	.075	.046	.046	.580	.070	.033	.043
0.8	.782	.075	.049	.052	.810	.071	.041	.044	.814	.070	.039	.043
Large												
0	.001	.108	.089	.051	-.001	.114	.115	.074	-.001	.151	.178	.043
0.2	.259	.093	.063	.052	.272	.103	.090	.060	.333	.118	.108	.041
0.5	.536	.072	.037	.052	.584	.073	.039	.040	.619	.067	.026	.033
0.8	.786	074	.047	.052	.825	.068	.037	.039	.832	.067	.035	.037
k = 30	γ_{00}	SE	σ_u^2	σ_v^2	γ_{00}	SE	σ_u^2	σ_v^2	γ_{00}	SE	σ_u^2	σ_v^2
Small												
0	-.003	.068	077	.053	-.004	.070	.088	.069	-.006	.086	.121	.059
0.2	.235	.063	.065	.053	.241	.067	.079	.062	.284	.076	.093	.053
0.5	.520	.053	.046	.055	.549	.054	.049	.050	.577	.052	.039	.045
0.8	.789	.053	.051	.055	.814	.051	.044	.046	.819	.050	.041	.046
Large												
0	-.003	.076	.089	.053	-.004	.081	.116	.073	-.006	.106	.175	.047

Table 2. (*continued*)

	Publication bias				Selective reporting				Both biases			
k = 30	γ_{00}	SE	σ_u^2	σ_v^2	γ_{00}	SE	σ_u^2	σ_v^2	γ_{00}	SE	σ_u^2	σ_v^2
0.2	.257	.068	.069	.053	.271	.074	.096	.062	.332	.086	.115	.043
0.5	.536	.052	.040	.055	.583	.053	.043	.045	.620	.049	.030	.038
0.8	.792	.053	.049	.055	.830	.049	.039	.042	.836	.048	.037	.041
k = 70	γ_{00}	SE	σ_u^2	σ_v^2	γ_{00}	SE	σ_u^2	σ_v^2	γ_{00}	SE	σ_u^2	σ_v^2
Small												
0	-.001	.045	.080	.053	-.001	.046	.092	.069	-.001	.057	.125	.057
0.2	.233	.042	.068	.053	.241	.044	.082	.063	.281	.050	.096	.054
0.5	.520	.035	.048	.054	.550	.035	.050	.049	.578	.034	.041	.045
0.8	.789	.035	.052	.054	.814	.034	.045	.047	.819	.033	.043	.046
Large												
0	-.001	.051	.093	.053	-.001	.054	.119	.074	-.001	.070	.178	.049
0.2	.254	.045	.072	.054	.268	.049	.098	.064	.328	.057	.119	.045
0.5	.535	.035	.043	.054	.584	.035	.046	.044	.618	.033	.035	.037
0.8	.793	.035	.050	.054	.831	.032	.041	.042	.838	.032	.037	.041

Notes. Small and Large refer to the degree of bias generated*; k* = initial number of studies; γ_{00}= combined effect size; SE= standard error; σ_u^2= between-studies variance; σ_v^2= between-outcomes variance. The between-studies and between-outcomes variances were fixed to 0.06 and the mean sample size (*n*) was fixed to 100. The values 0, 0.2, 0.5, and 0.8 refer to the population effect size.

Table 3.

Optimal values for the L_0^+ estimator of the Trim and Fill method.

Population effect size	Number of effect sizes	L_0^+ cutoffs
0	45 *(20 – 45)*	2
(-0.01 - 0.01)	90 *(41 – 90)*	2
	210 *(96 – 207)*	3
0.2	45 *(20 – 45)*	2
(0.20 - 0.33)	90 *(41 – 90)*	2
	210 *(96 – 207)*	3
0.5	45 *(20 – 45)*	2
(0.50 - 0.62)	90 *(41 – 90)*	3
	210 *(96 – 207)*	4
0.8	45 *(20 – 45)*	2
(.80 - .84)	90 *(41 – 90)*	3
	210 *(96 – 207)*	6

Notes. Values in parenthesis represent possible values of the conditions in real settings, because when selection bias exists, the estimated overall effect size increases and the number of studies and effect sizes decrease. These values are based on the values showed in Table 1 and 2.

Table 4.

Mean Type I error rate for each simulated condition and method.

	ERT	FP	BG (V)	BG (N)	TF-R_0^+	TF-L_0^+	FT
Mean	**.372**	.098	**.377**	**.170**	*.041*	**.124**	*.057*
Total variance							
0.02	**.258**	.097	**.288**	**.140**	*.039*	**.146**	*.052*
0.12	**.403**	.098	**.402**	**.180**	*.041*	.116	*.059*
0.22	**.455**	.099	**.442**	**.192**	*.042*	.109	*.060*
η^2	.066	.004	.063	.407	.005	.064	.059
Population ES							
0	.117	.093	**.167**	**.164**	*.034*	.087	*.062*
0.2	**.204**	.093	**.232**	**.162**	*.035*	**.129**	*.056*
0.5	**.495**	.099	**.470**	**.173**	*.044*	**.132**	*.053*
0.8	**.672**	.106	**.640**	**.185**	*.049*	**.148**	*.057*
η^2	.466	.125	.524	.067	.147	.131	.042
Number of studies							
15 (45 ESs)	**.309**	.092	**.319**	**.170**	*.025*	.114	*.054*
30 (90 ESs)	**.362**	.098	**.367**	**.170**	*.041*	**.134**	*.057*
70 (210 ESs)	**.444**	.104	**.446**	**.173**	*.056*	**.124**	*.060*
η^2	.029	.091	.040	.002	.572	.017	.031

Table 4. (*continued*)

	ERT	FP	BG (V)	BG (N)	TF-R_0^+	TF-L_0^+	FT
Mean sample size							
50	**.155**	.099	**.220**	**.148**	*.041*	**.164**	*.047*
100	**.416**	.096	**.408**	**.177**	*.041*	.114	*.059*
150	**.545**	.098	**.504**	**.187**	*.040*	.094	*.065*
η^2	.246	.007	.204	.215	.000	.219	.261

Notes. The nominal Type I error rate is .10. Too conservative values (< .0814) are represented in italic and too liberal values (> .1185) are indicated in bold. The simulated factor conditions of 'type of bias' and 'degree of bias' are not included here because in these datasets, no bias was generated. η^2= eta squared; ESs= effect sizes; ERT=Egger's Regression Test; FP=Funnel Plot test; BG (V)= Begg's Rank Correlation using the variance; BG (N)= Begg's Rank Correlation using the sample size; TF-R_0^+= Trim and Fill method using R_0^+ estimator; TF-L_0^+= Trim and Fill method using L_0^+ estimator; FT= Four-tandem procedure.

Table 5.

Mean power for each simulated condition and method.

	ERT	FP	BG (V)	BG (N)	TF-R_0^+	TF-L_0^+	FT
Mean	.462	.109	.481	.196	.061	.330	.107
Degree of bias							
Small	.448	.103	.463	.188	.050	.319	.095
Large	.476	.115	.499	.205	.073	.341	.119
η^2	.002	.028	.005	.025	.030	.003	.027
Type of bias							
Publication bias	.451	.099	.459	.190	.047	.179	.078
Selective Reporting	.447	.108	.471	.187	.060	.429	.113
Both	.489	.120	.513	.211	.077	.382	.131
η^2	.004	.056	.008	.036	.030	.149	.090
Total variance							
0.02	.353	.114	.420	.179	.085	.302	.115
0.12	.493	.109	.499	.203	.060	.338	.107
0.22	.540	.104	.524	.208	.040	.350	.099
η^2	.061	.012	.028	.052	.070	.005	.007
Population ES							
0	.143	.101	.210	.207	.051	.090	.076
0.2	.297	.105	.329	.191	.014	.548	.102
0.5	.634	.123	.636	.203	.069	.421	.138

Table 5. (*continued*)

		ERT	FP	BG (V)	BG (N)	TF-R_0^+	TF-L_0^+	FT
Population ES	0.8	.773	.108	.749	.184	.112	.262	.113
	η^2	.614	.052	.686	.026	.244	.372	.091
Number of studies								
	15 (45 ESs)	.361	.098	.391	.185	.047	.244	.082
	30 (90 ESs)	.450	.107	.468	.193	.066	.338	.103
	70 (210 ESs)	.575	.123	.584	.211	.071	.408	.136
	η^2	.074	.081	.089	.037	.022	.057	.089
Mean sample size								
	50	.327	.121	.408	.200	.068	.443	.137
	100	.481	.105	.484	.192	.060	.294	.096
	150	.579	.101	.550	.197	.057	.252	.089
	η^2	.104	.054	.048	.003	.004	.084	.081

Notes. η^2=eta squared; ES= effect size; ERT=Egger's Regression Test; FP=Funnel Plot test, BG (V)= Begg's Rank Correlation using the variance; BG (N)= Begg's Rank Correlation using the sample size; TF-R_0^+= Trim and Fill method using R_0^+ estimator; TF-L_0^+= Trim and Fill method using L_0^+ estimator; FT= Four-tandem procedure.

Table 6.

Classification of methods for each combination of conditions based on their mean power estimates, given Type I error rate control.

PES	*k*	ESs	*n*	Total Var. = 0.02	Total Var. = 0.12	Total Var. = 0.22
0	15	45	50	Egger's Regression	Egger's Regression	Egger's Regression
			100	Egger's Regression	Funnel Plot	Funnel Plot
			150	Egger's Regression	Funnel Plot	Funnel Plot
	30	90	50	Egger's Regression	Egger's Regression	Egger's Regression
			100	Egger's Regression	Funnel Plot	Funnel Plot
			150	Egger's Regression	Funnel Plot	Funnel Plot
	70	210	50	Egger's Regression	Egger's Regression	Egger's Regression
			100	Egger's Regression	Funnel Plot	Funnel Plot
			150	Egger's Regression	Funnel Plot	Funnel Plot
0.2	15	45	50	Trim and fill L_0^+*	Trim and fill L_0^+*	Trim and fill L_0^+
			100	Trim and fill L_0^+*	Trim and fill L_0^+*	Trim and fill L_0^+
			150	Trim and fill L_0^+*	Trim and fill L_0^+*	Trim and fill L_0^+
	30	90	50	Egger's Regression	Egger's Regression	Egger's Regression
			100	Funnel Plot	Funnel Plot	Funnel Plot
			150	Funnel Plot	Funnel Plot	Funnel Plot
	70	210	50	Egger's Regression	Egger's Regression	Egger's Regression
			100	Funnel Plot	Funnel Plot	Funnel Plot
			150	Funnel Plot	Funnel Plot	Funnel Plot

Table 6. (*continued*)

PES	*k*	ESs	*n*	Total Var. = 0.02	Total Var. = 0.12	Total Var. = 0.22
	15	45	50	Egger's Regression	Four-tandem	Four-tandem
			100	Funnel Plot	Trim and fill L_0^+	Trim and fill L_0^+*
			150	Funnel Plot	Trim and fill L_0^+	Trim and fill L_0^+*
	30	90	50	Egger's Regression*	Four-tandem	**Trim and fill L_0^+**
0.5			100	Trim and fill R_0^+	Trim and fill L_0^+*	Trim and fill L_0^+*
			150	Trim and fill R_0^+	Trim and fill L_0^+	Trim and fill L_0^+*
	70	210	50	**Egger's Regression**	Four-tandem	Four-tandem
			100	Trim and fill R_0^+	Four-tandem	Four-tandem
			150	Trim and fill R_0^+	Trim and fill L_0^+*	**Trim and fill L_0^+**
	15	45	50	Trim and fill R_0^+	Trim and fill R_0^+	Funnel Plot
			100	Funnel Plot	Funnel Plot	Funnel Plot
			150	Funnel Plot	Funnel Plot	Trim and fill L_0^+
	30	90	50	Trim and fill R_0^+	Trim and fill R_0^+	Four-tandem
0.8			100	Trim and fill R_0^+	Trim and fill R_0^+	Trim and fill L_0^+
			150	Funnel Plot	Trim and fill R_0^+	Trim and fill L_0^+
	70	210	50	Four-tandem	Four-tandem	**Trim and fill L_0^+**
			100	Trim and fill R_0^+	Trim and fill R_0^+	Trim and fill L_0^+
			150	Trim and fill R_0^+	Trim and fill R_0^+	Trim and fill L_0^+

Notes. PES= Population effect size; Total Var. = Total variance*;* *k* = number of studies; *n* = mean sample size of primary studies; ESs= effect sizes. The methods in each cell exhibited the largest power in that specific condition and controlled the Type I error rate. Methods in bold

yielded a mean power between .5 and .7. Methods with an asterisk had a mean power between .3 and .5. Methods without any special characteristic (not bold, not asterisk) had a mean power below .3.