Evaluating Meta-Analytic Methods to Detect Selective Reporting in the Presence of Dependent Effect Sizes

Melissa A. Rodgers¹ & James E. Pustejovsky¹

¹ The University of Texas at Austin

Forthcoming in *Psychological Methods*. This paper is not the version of record and may not exactly replicate the final, published version of the article. The final article will be available, upon publication via its DOI.

Selective reporting of results based on their statistical significance threatens the validity of meta-analytic findings. A variety of techniques for detecting selective reporting, publication bias, or small-study effects are available and are routinely used in research syntheses. Most such techniques are univariate, in that they assume that each study contributes a single, independent effect size estimate to the meta-analysis. In practice, however, studies often contribute multiple, statistically dependent effect size estimates, such as for multiple measures of a common outcome construct. Many methods are available for meta-analyzing dependent effect sizes, but methods for investigating selective reporting while also handling effect size dependencies require further investigation. Using Monte Carlo simulations, we evaluate three available univariate tests for small-study effects or selective reporting, including the Trim & Fill test, Egger's regression test, and a likelihood ratio test from a three-parameter selection model (3PSM), when dependence is ignored or handled using ad hoc techniques. We also examine two variants of Egger's regression test that incorporate robust variance estimation (RVE) or multi-level meta-analysis (MLMA) to handle dependence. Simulation results demonstrate that ignoring dependence inflates Type I error rates for all univariate tests. Variants of Egger's regression maintain Type I error rates when dependent effect sizes are sampled or handled using RVE or MLMA. The 3PSM likelihood ratio test does not fully control Type I error rates. With the exception of the 3PSM, all methods have limited power to detect selection bias except under strong selection for statistically significant effects.

Keywords: meta-analysis; selective reporting; small-study effects; publication bias; robust variance estimation

Research synthesis is the practice of combining empirical evidence from multiple studies on a common research question for purposes of drawing generalizations. Using the statistical tools of meta-analysis, research syntheses aim to estimate an average effect size across studies,

characterize the extent of variation in effect sizes, and investigate factors that might explain differences in the average effect sizes (Cooper, Hedges, & Valentine, 2019). Research syntheses play a critical and prominent role in the study of psychological theory, in the development of interventions, and as a basis for informing evidence-based policy and practice across a range of fields (Gurevitch, Koricheva, Nakagawa, & Stewart, 2018).

The validity of findings from a research synthesis hinges on whether the primary study results represent the full scope of relevant evidence. Therefore, a critical threat to validity is selective reporting, which occurs when the availability of results from a primary study is influenced by the statistical significance or magnitude of its findings (Rothstein, Sutton, & Borenstein, 2006). Selective reporting of primary study results can occur at the level of the study, where the choice to write up or publish a study might be more likely if one or more key findings are sta-

Earlier versions of this paper were presented at the 2019 annual convention of the American Educational Research Association in Toronto, Canada and the 2019 annual meeting of the Society for Research Synthesis Methods in Chicago, IL. The authors are grateful to Man Chen, Bethany Hamilton, Megha Joshi, Young Ri Lee, and Devin Dedrick for feedback on this work. Supplementary materials for this article are available on the Open Science Framework at <https://bit.ly/2qnF4dH>.

Correspondence concerning this article should be addressed to Melissa A. Rodgers, The University of Texas at Austin, Educational Psychology Department, 1912 Speedway, MS D5800, Austin, TX 78712. E-mail: melissa.a.rodgers@utexas.edu

tistically significant. Selective reporting can also occur *within* a study, if the choice to report or publish a specific study result depends on its statistical significance. For instance, a researcher might collect five different outcome measures of social-emotional well-being, but selectively report only the subset that tell the most compelling story, in order to improve the chance of a journal publishing the study. This type of selective reporting can be understood as a form of p-hacking (Nuzzo, 2014; Simmons, Nelson, & Simonsohn, 2011).

Reviews of the psychology and education literature have demonstrated that selective reporting is a prevalent problem and that including the unpublished literature in syntheses influences average effect size estimates. In a review of 81 meta-analytic studies published in *Review of Educational Research* and *Psychological Bulletin* between 1986 and 2013, Polanin, Tanner-Smith, and Hennessey (2016) reported that published studies had systematically larger effect sizes than unpublished studies, by an average of 0.18 standard deviations (95% CI [0.10, 0.25]). Even more severe, a review of meta-analyses appearing in the 2013 volumes of 37 special education journals found that published studies tended to have larger effect sizes than unpublished studies, by an average 0.89 standard deviation (95% CI [0.15, 1.65]; Gage, Cook, & Reichow, 2017). Reviews that included only published studies exhibited higher rates of selective reporting than reviews that included unpublished work (Gage et al., 2017). Selective reporting is also a concern within the psychological sciences, where one review reported that 26 (41%) of 64 meta-analytic studies detected some degree of publication bias, despite the majority of studies including some form of unpublished research (Ferguson & Brannick, 2012).

An extensive set of methods have been developed for diagnosing the presence of selective reporting and trying to adjust for the biases that result (Carter, Schönbrodt, Gervais, & Hilgard, 2019; Rothstein et al., 2006). Methods that are readily available in statistical software and commonly used in meta-analyses include the rank correlation test (Begg & Mazumdar, 1994), Trim & Fill (Duval & Tweedie, 2000), and regression test of funnel plot asymmetry (Egger, Smith, Schneider, & Minder, 1997). These methods all seek to detect selective reporting by examining whether effect size magnitude is correlated with study size, or what is called a small-study effect. Another class of methods are selection models, which incorporate explicit assumptions about how the magnitude and statistical significance of effect size estimates relate to the probability that a result is published (Hedges & Vevea, 2005). Although not as widely used in research synthesis, simple forms of selection models have attracted increased attention due to their simplic-

ity, alignment with conventional understanding of how statistical significance is interpreted, and strong performance in recent simulation studies (Carter et al., 2019; McShane, Böckenholt, & Hansen, 2016).¹

Nearly all of the commonly used methods for investigating selective reporting are based on the assumption that each primary study contributes a single effect size estimate to the synthesis. In practice, however, it is very common for syntheses to include primary studies that contribute multiple, statistically dependent effect sizes (Ahn, Ames, & Myers, 2012; Park & Beretvas, 2018; Tipton, Pustejovsky, & Ahmadi, 2019). Dependent effect sizes can occur when primary studies report multiple outcomes, compare multiple treatment groups, or include longitudinal repeated measures (Becker, 2000; Gleser & Olkin, 2009). For example, a recent meta-analysis of differences in executive functioning between bilingual and monolingual adults included 891 effect sizes from 152 studies (Lehtonen et al., 2018). Individual studies contributed a between 1 and 40 effect sizes (median = 4) as a result of reporting multiple measures of executive functioning, comparing multiple types of bilingual or monolingual groups, or both.

Methodologists have developed a variety of techniques to handle dependent effect sizes in meta-analysis (Cheung, 2019; Hedges, 2019). Ad hoc strategies include selecting one effect size or aggregating effect size estimates in order to create a single “synthetic effect” per study. More sophisticated multivariate modeling strategies include generalized least squares (Kalaian & Raudenbush, 1996; Raudenbush, Becker, & Kalaian, 1988), multi-level meta-analysis (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013, 2015), and robust variance estimation (Hedges, Tipton, & Johnson, 2010). In large-scale synthesis projects, researchers are increasingly turning to these multivariate modeling approaches for summarizing effects across studies, describing effect size heterogeneity, and examining potential moderators. However, corresponding tools for investigating selective reporting in syntheses of dependent effect sizes remain lacking.

Despite the breadth of methodological research on handling dependent effect sizes and on techniques for detecting selective reporting, little previous research has considered how to deal with both of these issues simultaneously. Lacking any guidance and left to their own

¹Many methods to detect small study effects or selective reporting also include a function that estimates an adjusted average effect size after accounting for selective reporting. The scope of the present study is limited to methods of detecting selective reporting because the set of adjustment methods is distinct enough and complex enough as to warrant a separate study.

devices, researchers conducting meta-analyses have typically either a) ignored the issue of dependence when investigating selective reporting, using methods that assume independent effect sizes (e.g. Clark, Tanner-Smith, & Killingsworth, 2016; Fisher, Gardella, & Teurbe-Tolon, 2016; Weisz et al., 2017) or b) used ad hoc approaches for handling dependence, such as aggregating effect sizes into one synthetic effect size per study, and then using selective reporting methods that assume independent effect sizes (e.g. Egert, Fukkink, & Eckhardt, 2018; Zheng, Warschauer, Lin, & Chang, 2016). A handful of applied studies have implemented regression tests of funnel plot asymmetry while handling dependent effects using robust variance estimation techniques (Bediou et al., 2018; Coles, Larsen, & Lench, 2019; Stevens, Rodgers, & Powell, 2017). As a novel method, the combination of regression tests and robust variance estimation has yet to be evaluated in methodological studies.

In one of the only methodological studies on detecting selective reporting with dependent effect sizes, Fernández-Castilla et al. (2019) proposed to combine regression tests with multi-level meta-analysis for handling dependence. Using simulations, they demonstrated that this test maintains nominal Type I error rates, but has low power to detect selection (Fernández-Castilla et al., 2019). This multi-level regression test has been implemented in at least one meta-analysis (Lehtonen et al., 2018). Given the novelty of this approach, however, further methodological evaluation of the multi-level Egger's regression test is warranted—particularly under a broader set of data generating conditions than examined in Fernández-Castilla et al. (2019).

Applied meta-analyses often use multiple strategies for handling dependence and multiple methods for investigating potential selective reporting and publication biases (Tipton et al., 2019). For instance, Coles et al. (2019) conducted several tests for selective reporting based on aggregated effect sizes and by using robust variance estimation. Selective reporting was not identified when aggregating, but was detected when dependency was addressed with robust variance estimation. If different methods yield conflicting results, researchers may have difficulty drawing any conclusion about selective reporting bias. Thus, there is an outstanding need to understand and compare the performance of different methods of handling dependent effect sizes when investigating potential selective reporting biases.

The goal of the present study is to begin to address this need. Specifically, we use Monte Carlo simulation to study the performance of commonly used tests for detecting selective reporting in meta-analyses that include primary studies with multiple, dependent effect size estimates. We assess the performance of each test when

used in combination with typical methods for handling dependent effects: ignoring dependencies, aggregating effect sizes to the study level, or sampling a single effect per study. We also examine the performance of novel approaches implemented in Coles et al. (2019), Lehtonen et al. (2018) and several other recent syntheses, which combine a regression test for funnel plot asymmetry with RVE or MLMA methods. Our simulations focus on standardized mean differences, the most common effect size metric in meta-analyses of psychological and educational research. We aim to provide guidance for applied researchers who wish to conduct tests for selective reporting using meta-analytic datasets that include multiple, dependent effect size estimates. For sake of feasibility, we limit the scope of our investigation to methods for diagnosing selective reporting or small-study effects, and so we do not consider methods for estimating bias-corrected average effect sizes after accounting for selective reporting.

The remainder of the paper proceeds as follows. In the next section, we briefly review available methods for handling dependent effect sizes estimates, tests for selective reporting that are currently used in practice, and how these methods are integrated in applied research. We then apply a selection of common methods and tests to empirical data from a recent meta-analysis published in *Psychological Bulletin* (Lehtonen et al., 2018). In the following sections, we describe the simulation study methods and results, which include Type I error rates of common tests in the absence of selective reporting and power to detect selective reporting when present. We conclude with a discussion of limitations, outstanding research needs, and implications for applied meta-analytic research.

Handling dependent effect sizes and testing for selective reporting

In practice, meta-analysts face the dual challenge of handling dependent effect sizes and testing for selective reporting. In order to understand the current range of approaches for dealing with this dual challenge, it is useful to first consider the available methods for addressing each component in isolation. We therefore briefly review currently available methods for handling dependent effect size estimates in meta-analysis. Next, we review available univariate tests of selective reporting or small-study effects, focusing on popular and readily available methods. Finally, we consider the combination of tests of selective reporting or small-study effects with methods for handling dependence.

Handling dependent effect size estimates

Consider a meta-analysis that includes k studies, indexed by $j = 1, \dots, k$, each of which might contribute multiple, statistically dependent effect size estimates. Suppose that study j contributes $n_j \geq 1$ effect size estimates, $T_{1j}, \dots, T_{n_j j}$, with corresponding sampling variances $\sigma_{1j}^2, \dots, \sigma_{n_j j}^2$. Following meta-analytic convention, we assume that the sampling variances are known. Let ρ_{hij} denote the sampling correlation between effect size estimates T_{hj} and T_{ij} within study j . In some applications, it is possible to estimate these correlations from data reported in primary studies. However, it is often the case that primary studies do not provide sufficient information to estimate the correlations (Becker, 2000; Hedges et al., 2010)—a practical challenge that motivated the development of several of the techniques we now describe.

Becker (2000) described four broad strategies that meta-analysts use to handle dependent effect sizes: 1) ignore the dependency and treat the effects as if they were all independent; 2) aggregate effects within each study; 3) create sub-groups or select a single effect size from each study; or 4) use a multivariate model that accounts for within-study dependency. The first strategy, ignoring dependence among effect sizes from a common study, will tend to under-state the uncertainty of meta-analytic estimates and inflate Type I error rates (Becker, 2000). These problems will be more severe when many studies contribute multiple effect sizes and when the dependent effect sizes are highly correlated, so that each effect size contributes less independent information. Additionally, pooled effect size estimates will be more strongly influenced by primary studies that contribute many dependent effect sizes than by those with only one or very few effect sizes (Scammacca, Roberts, & Stuebing, 2014).

Another common strategy for handling dependency is to aggregate effects within each study, yielding a set of “synthetic” effect size estimates. The simplest approach to aggregation is to use the arithmetic mean of the effect sizes and sampling variances for each primary study, taking

$$\bar{T}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} T_{ij} \quad \text{and} \quad \bar{\sigma}_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} \sigma_{ij}^2. \quad (1)$$

The synthetic effect sizes $\bar{T}_j$ and sampling variances $\bar{\sigma}_j^2$, $j = 1, \dots, k$, can then be analyzed using basic meta-analysis or meta-regression models. If the arithmetic mean is used, the aggregated sampling variance ($\bar{\sigma}_j^2$) will only be a rough approximation of the actual sampling variance, as this method assumes that the correlation between the effect sizes is $\rho = 1$. As a result, using the

aggregated sampling variances might tend to overstate between-study heterogeneity and inflate the standard errors of pooled summary effects. Other methods of aggregation involve weighted averages of effect sizes and rely on knowledge of the correlations between effects within each study (Borenstein, Hedges, Higgins, & Rothstein, 2009; Gleser & Olkin, 2009; Marin-Martinez & Sanchez-Meca, 1999). If these correlations are not reported, as is often the case, then a constant correlation among effects is often assumed to calculate a rough approximation of the sampling variance (Borenstein et al., 2009). This might still lead to inaccuracies to the extent that the assumed correlation deviates from the true pattern of within-study dependence.

A third strategy is to select or sample a single effect size from each study, such as by choosing the effect size for the outcome measure that most closely aligns with target outcome construct. More commonly, effect sizes could be classified into sub-groups, where each study contributes only a single effect size estimate per sub-group. The researcher can then apply meta-analytic methods appropriate for independent effect sizes, possibly repeating the analyses within each sub-group of effects. However, running analyses separately for each sub-group has the drawbacks of inflating Type I error, making it more difficult to draw summary conclusions, and precluding the possibility of making statistical comparisons between sub-groups (Scammacca et al., 2014). Selecting a single effect size per study reduces the number of effect sizes available for analysis, which may result in decreased power and loss of information for conducting moderator analysis. Selecting a single effect that is “most representative” of a study also entails making high inference judgments, which might present risks for the reproducibility and objectivity of the synthesis.

A fourth broad strategy is to handle the dependencies among effect size estimates using statistical modeling techniques. Compared to other strategies, using a multivariate model allows for the inclusion of all available effect size information, provides a way to make comparisons between different outcome categories, and yields more accurate characterization of uncertainty (Becker, 2000; Van den Noortgate et al., 2013). Three distinct multivariate approaches have been described: generalized least squares (GLS; Gleser & Olkin, 2009; Kalaian & Raudenbush, 1996; Raudenbush et al., 1988), multi-level meta-analysis (MLMA; Van den Noortgate et al., 2013), and robust variance estimation (RVE; Hedges et al., 2010). We describe these approaches in the context of a general meta-regression model. In the equations that follow, let $\mathbf{x}_{ij}$ be a $1 \times p$ row-vector of predictors corresponding to effect i in study j , which encode information

about one or more potential moderating variables.²

A multivariate GLS model (Kalaian & Raudenbush, 1996; Raudenbush et al., 1988) involves the meta-regression

$$T_{ij} = \mathbf{x}_{ij}\beta + v_{ij} + e_{ij}, \quad (2)$$

where β is a $p \times 1$ vector of meta-regression coefficients, v_{ij} is a random effect, and e_{ij} is the sampling error for effect i in study j , with $\text{Var}(e_{ij}) = \sigma_{ij}^2$. The distinguishing feature of the GLS model is that it incorporates the assumption $\text{Cov}(e_{hj}, e_{ij}) = \rho_{hij}\sigma_{hj}\sigma_{ij}$, and thus requires knowledge of the sampling correlations ρ_{hij} for $h, i = 1, \dots, n_j$, $j = 1, \dots, k$. The initial model proposed by Raudenbush et al. (1988) used a fixed effect assumption, and so omitted the random effect v_{ij} . Later work (Kalaian & Raudenbush, 1996) generalized the model to include random effects that capture within- and between-study heterogeneity. Modern meta-analytic software such as the *metafor* package (Viechtbauer, 2010) allows for a wide range of different assumptions about the variance-covariance structure of the random effects. However, valid estimation of the model (including the variance-covariance structure, the meta-regression coefficients, and their standard errors) relies on knowledge of the sampling correlations, which are often not reported in primary studies. This limits the application of the GLS model for handling dependence, particularly when studies report two or more outcomes on the same sample.

Two more recently developed model-based approaches use all available effect size estimates while avoiding the need for knowledge of the true sampling correlations among effect size estimates within studies. Van den Noortgate et al. (2013) proposed a multi-level meta-analysis (MLMA) model to address the nesting of outcomes within studies, which involves partitioning the variation in effect size estimates into three levels: the sampling error of individual effect sizes, within-study heterogeneity, and between-study heterogeneity. The MLMA model is

$$T_{ij} = \mathbf{x}_{ij}\beta + u_j + v_{ij} + e_{ij}, \quad (3)$$

where u_j is a study-level random effect, v_{ij} is a within-study random effect, and e_{ij} is the sampling error for effect size i in study j . Unlike with the GLS model in Equation (2), the MLMA model assumes that the sampling errors are independent. This assumption is clearly violated when a primary study reports multiple outcomes based on the same sample. However, Van den Noortgate et al. (2013) argued that the multi-level random effects structure accounts for within-study dependence between effect sizes by overestimating the study-level variance. By taking advantage of the multi-level structure,

accounting for sampling covariance through the study-level variance can yield appropriate standard errors for the average effect size estimate (Van den Noortgate et al., 2013). Results of a recent simulation study demonstrated that inferences from the MLMA are appropriate under certain conditions, including when: a) smaller numbers of effect sizes per study or studies are included, b) effect sizes are similar in magnitude within studies, and c) the within- and between-study variances are large (Van den Noortgate et al., 2015). Standard errors tend to be underestimated as the intraclass correlation decreases or number of effect sizes per study increases.

RVE, proposed by Hedges et al. (2010), addresses the issue of dependence among effect sizes without requiring knowledge of the effect size variance-covariance matrix. It does so by estimating the meta-regression based on a “working” model for the dependence structure and then using cluster-robust variance estimation methods to obtain standard errors and confidence intervals for the meta-regression coefficients. The default implementations of RVE, as instantiated in the *robumeta* package for R (Fisher, Tipton, & Zhipeng, 2017) and Stata (Hedberg, 2014), offers a choice of two working models, called “hierarchical” and “correlated” effects models. The hierarchical effects model is identical to Equation 3, but uses moment estimators for the within- and between-study variance components (Hedges et al., 2010). The correlated effects model is simpler and includes only a between-study random effect,

$$T_{ij} = \mathbf{x}_{ij}\beta + u_j + e_{ij}, \quad (4)$$

where $\text{Var}(u_j) = \tau^2$. The sampling variances are treated as dependent, with $\text{Cov}(e_{hj}, e_{ij}) = \rho\sigma_{hj}\sigma_{ij}$, and a moment estimator is used to determine the between-study variance. In contrast to the GLS model, the true sampling correlations among effect size estimates are not required. Instead, RVE uses an assumed, constant value of ρ , with a default of $\rho = 0.8$ in available software implementations. The exact value of the assumed correlation does not typically influence the model estimates, although sensitivity analysis with varying values are recommended and readily available in software (Hedges et al., 2010; Tanner-Smith & Tipton, 2014).

The default implementation of RVE uses a simple, approximately efficient method to estimate the meta-regression coefficients. Rather than using full inverse-variance weighting, RVE uses weighted least squares estimation of Equation (4) with weights given by

²For a summary meta-analysis, the predictor vector would consist solely of an intercept term: $\mathbf{x}_{ij} = 1$. Generally, the predictors may include study-level characteristics, effect-level characteristics (that vary within-study), or both.

$w_{ij} = 1 / \left[n_j \left(\hat{\sigma}_j^2 + \hat{\tau}^2 \right) \right]$ for the correlated effects model or $w_{ij} = 1 / \left(\hat{\sigma}_{ij}^2 + \hat{\tau}^2 + \hat{\omega}^2 \right)$ for hierarchical effects weights. Here, $\hat{\tau}^2$ and $\hat{\omega}^2$ are the method-of-moments estimators of the between-study variance and within-study variance, respectively. The sampling variances and covariances of the meta-regression coefficients are estimated using a cluster-robust method that remains asymptotically consistent even when the working model does not correctly represent the true dependence structure among the effect sizes from each study. Small-sample corrections are incorporated into the variance-covariance estimator, so that it is approximately unbiased when the number of studies is small or moderate (Tipton, 2015; Tipton & Pustejovsky, 2015). Both MLMA and RVE share the advantage of not requiring knowledge of sampling correlations. A strength of RVE is that it has been proven to provide asymptotically consistent standard errors and inferences, whereas the robustness of MLMA rests only on simulation evidence, which may not generalize to all situations encountered in practice.

Considering the four broad strategies for handling dependence, ignoring dependency tends to inflate Type I error rates and produce poor point and standard error estimates. Sampling or aggregating methods are simple to implement and allow for the use of univariate meta-analytic methods, but they may result in poor estimates of the average effect sizes and standard errors. Nonetheless, recent reviews have found that applied meta-analyses with dependent effect sizes were more likely to use one or a combination of the simpler methods, rather than GLS, MLMA, or RVE approaches (Ahn et al., 2012; Tipton et al., 2019).

Testing for selective reporting or small-study effects

Methodologists have proposed a variety of analytic methods to detect selective reporting and small-study effects (Sutton, 2009), but nearly all available tests assume that each study contributes a single independent effect size estimate. The oldest method to diagnose selective reporting is to visually inspect a funnel plot for asymmetry. The funnel plot is a simple scatter plot of effect size estimates versus a measure of their precision, such as sample size, standard error, variance, or the inverse of any of these (Light & Pillemer, 1984; Sterne & Egger, 2001). Visual inspection focuses on the vertical symmetry of the plot. Absence of effect size estimates in the corner of the plot closer to null suggests that non-significant effect sizes may be missing due to selective reporting. Several of the most common formal statistical tests for selective reporting are based on the same premise as the funnel plot: that an asymmetric distribu-

tion of effect sizes might be due to selective reporting of significant results.

Trim & Fill (Duval & Tweedie, 2000) is one of the more popular methods to detect and attempt to correct for selective reporting. This method attempts to create symmetry in the funnel plot by identifying the non-significant and less precise effect sizes that are absent from the side of the funnel plot closer to a null effect. The test iteratively estimates the number of studies that are missing using one of three estimators (R_0 , L_0 , or Q_0). Only one of the estimators, R_0 , allows for a statistical test of the null hypothesis that the number of missing studies is zero. This test uses a non-parametric signed rank test statistic to determine if funnel plot asymmetry is present, under the assumption that effect size estimates are all independent. Past simulations have found that high heterogeneity and large samples sizes tend to inflate the Type I error rate of this test (Carter et al., 2019; Moreno et al., 2009; Peters, Sutton, Jones, Abrams, & Rushton, 2006). We nonetheless include the Trim & Fill method with the R_0 estimator in the simulation study because of its popularity, apparent simplicity, and availability in statistical software.

Another popular method to diagnose potential small-study effects is a regression test, often called Egger's regression. Egger et al. (1997) proposed a test of funnel plot asymmetry based on a meta-regression model with a measure of effect size precision as a predictor:

$$T_j = \beta_0 + \beta_1 x_j + e_j \quad (5)$$

where T_j is the effect size estimate for study j and x_j is the measure of precision. In Egger's original formulation, the predictor is $x_j = \sigma_j$, the standard error of the effect size estimate. Equation (5) is estimated using weighted least squares, with weights equal to the inverse of the sampling variances, $1/\sigma_j^2$. If there is no relationship between effect size and precision (i.e., no small-study effects), then $\beta_1 = 0$. Thus, funnel plot asymmetry is indicated if the coefficient β_1 is significantly different than zero.

A complication with Egger's regression is that for certain effect size metrics, the standard error is naturally correlated with the effect size estimate even in the absence of selective reporting or other sources of asymmetry. Different variants of Egger's regression have been developed to reduce the correlation by using alternative measures of precision, specifically for log odds ratios (Macaskill, Walter, & Irwig, 2001; Moreno et al., 2009; Peters et al., 2006), raw proportions (Hunter et al., 2014), hazard ratios (Debray, Moons, & Riley, 2018), and standardized mean differences (Pustejovsky & Rodgers, 2019). By reducing the correlation between

the effect size estimate and its precision, the modified predictors maintain close-to-nominal Type I error rates in the absence of funnel plot asymmetry.

A further and more fundamental complication with Egger's regression is that it cannot be interpreted as a conclusive test for selective reporting. Rather, it is a test for funnel plot asymmetry, which may have other possible explanations besides selective reporting. Other mechanisms, such as high between-study heterogeneity or study-design differences, could also lead to smaller, less precise studies having larger effects, creating asymmetry in the funnel plot. The Trim & Fill test shares the same interpretative limitation. Thus, these tests should be interpreted as tests of funnel plot asymmetry or small-study effects, rather than selective reporting per se.

Another class of methods, called selection or weight function models (Hedges, 1992; Hedges & Vevea, 2005; Vevea & Woods, 2005), can be used to more directly diagnose selective reporting. These models make explicit assumptions about how the probability of observing a study result depends on its statistical significance. Among the different forms of these models, the 3-parameter selection model (3PSM) is the simplest and has shown promise in several recent methodological studies (Carter et al., 2019; Hedges & Vevea, 2005; McShane et al., 2016; Pustejovsky & Rodgers, 2019). We therefore focus on it here and in the subsequent simulations.

The 3PSM is formulated in two parts: a model for the data-generation process and a model for the selection process (Vevea & Hedges, 1995). In the first part, it is assumed that effect size estimates follow a mixed effects model:

$$T_j = \mathbf{x}_j\beta + u_j + e_j, \quad (6)$$

where $\mathbf{x}_j$ is a set of predictors for study j , u_j is a random effect with mean 0 and variance τ^2 , and e_j is the sampling error, with $\text{Var}(e_j) = \sigma_j^2$. In the second part, it is assumed that the result for study j is observed with some probability π^* if the effect size estimate is in the expected direction and statistically significant at level α . If the effect size estimate is not statistically significant or not in the expected direction, then it is observed with some other probability π^{ns} . We cannot estimate the absolute probability of observing a significant result, but we can estimate $\psi = \pi^{ns}/\pi^*$, the chance of observing a non-significant result relative to the chance of observing a significant one. Thus, the parameters of the 3PSM are the meta-regression coefficients β , between-study heterogeneity τ^2 , and relative chance of observing a non-significant effect size estimate ψ . Maximum likelihood methods are used to estimate the model parameters.

After fitting the 3PSM, a likelihood ratio test can be used

to detect selective reporting. The test evaluates the null hypothesis that there is no selection of significant results ($\psi = 1$), in which case the model reduces to the conventional mixed effects model. A significant likelihood ratio test statistic indicates selective reporting on the basis of the α -level statistical significance threshold (Hedges, 1992; Vevea & Hedges, 1995).

Several alternatives to selection models have been proposed, including the test of excess significance (TES; Ioannidis & Trikalinos, 2007), p-curve (Simonsohn, Nelson, & Simmons, 2014), and p-uniform (Van Assen, Van Aert, & Wicherts, 2015). TES involves comparing the observed number of significant effect sizes to the expected number based on power to detect the estimated pooled effect size (Ioannidis & Trikalinos, 2007). The other methods focus on the distribution of p-values associated with statistically significant effect size estimates. Evidence from simulations has indicated that these methods do not perform as well as the 3PSM or other selection models (McShane et al., 2016; Van Aert, Wicherts, & Van Assen, 2016). In particular, the methods do not work well when there is between-study heterogeneity in effect sizes (Van Aert et al., 2016). Because of their inferior performance, and because there is no readily available way to apply these methods to syntheses of dependent effect sizes, we do not consider these techniques in the simulation study.

Testing for selective reporting with dependent effect size estimates

Lacking fully developed methods for detecting selective reporting or small-study effects in syntheses that include dependent effect sizes, meta-analysts have typically applied ad hoc strategies to obtain a data structure that aligns with the assumptions of currently available methods. A common approach is to first define sub-groups of effects based on a particular moderator, aggregate effect sizes to the study level within each category, and then conduct tests for small-study effects within each sub-group. For example, in a recent synthesis examining the effects of professional development on teacher quality and child outcomes in early child education, the researchers aggregated outcomes based on the type of in-service treatment being evaluated (Egert et al., 2018). Zheng et al. (2016) analyzed the effect of learning with laptops by sub-grouping and then aggregating the dependent effect sizes within five different subject areas. Both studies reported multiple funnel plots and applied Egger's regression test within each sub-group.

Even when statistical approaches like RVE and MLMA are used to handle dependency when estimating an overall average effect size or conducting moderator analyses,

researchers have often ignored the dependence and implemented tests for small-study effects that assume independent effect sizes. For example, a recent synthesis of randomized trials on the effects of youth psychotherapy on mental health outcomes (Weisz et al., 2017) used MLMA to handle dependence. Two other recent syntheses (Clark et al., 2016; Fisher et al., 2016) used RVE to address effect size dependency. In all three studies, the authors ignored dependency when conducting Trim & Fill and Egger's regression tests to detect funnel plot asymmetry.

Of the available tests for selective reporting or small-study effects, only Egger's regression can readily be combined with multivariate modeling methods for handling dependency. In a recent simulation study, Fernández-Castilla et al. (2019) extended Egger's regression test to the MLMA model proposed by Van den Noortgate et al. (2013). They referred to this approach as the "Funnel Plot Test" and compared its performance to Trim & Fill, Begg's Rank Correlation test, and the original Egger's regression method, across a range of conditions. They demonstrated that the multi-level version of the regression test maintained Type I error rates at a nominal $\alpha = 0.10$ level across all conditions examined; however, power to detect small-study effects was minimal. This simulation study was also limited to a data-generating model where effect size estimates were conditionally independent, given the true effects for the study—conditions where the MLMA was correctly specified in the absence of selection.

An alternative, which we also evaluate in the subsequent simulations, is to combine the regression test with RVE for handling dependency. Because the regression test is simply a meta-regression, it can be conducted by including a measure of effect size precision as a predictor in Equation (4), estimating the slope of this predictor using weighted least squares, and testing significance using cluster-robust variance estimation methods (i.e., sandwich estimators). We call this the "Egger Sandwich" test, and the approach combined with a multi-level model "Egger MLMA."

A handful of empirical studies have recognized the potential to conduct regression tests of funnel plot asymmetry using RVE and MLMA techniques. Stevens et al. (2017) synthesized effects of math interventions for remedial and special education students, using a sample of 22 studies with 102 effect sizes. When testing for small-study effects, they conducted Egger's regression test within the RVE framework. Lehtonen et al. (2018) synthesized the literature on executive functioning differences between bilingual and monolingual adults using a four-level MLMA model and reported tests for small-study effects using Egger's regression within this

the framework of the MLMA. Bediou et al. (2018) and Coles et al. (2019) both investigated small-study effects using multiple methods. Following conventional practice, they conducted Trim & Fill and Egger's regression tests while ignoring dependence and after aggregating or sub-setting to obtain a single effect per study. Additionally, both studies also used the Egger Sandwich test on the full dataset, including in a model with no other predictors and in a model that included other potential moderating variables. Both studies reported conflicting findings from the different methods used to detect small-study effects. These examples point towards the need for further methodological research on the performance of available methods.

Empirical illustration

As further illustration of the need for research on the performance of selective reporting tests with dependent effect sizes, consider a recent study that appeared in *Psychological Bulletin*. Lehtonen et al. (2018) reported a systematic review and synthesis examining differences between bilingual and monolingual adults in six domains of executive functioning. The original analysis estimated meta-analytic models and investigated selective reporting separately within each of the six domains. For purposes of illustration, we re-analyzed the data from the largest domain, working memory, which included 73 studies with a total of 243 effect sizes. Within this domain, studies contributed between 1 and 18 effect sizes (median = 4) with sample sizes ranging from 12 to 344 participants (median = 48). Lehtonen et al. (2018) found significant small-study effects based on visual inspection of the funnel plots and on a regression test conducted within a multi-level meta-analytic model.

Table 1 presents the results of three tests for selective reporting or small-study effects, when a) effect size dependency is ignored, b) effect sizes are aggregated to the study level, or c) a single effect size is randomly selected from each study. In addition, the table includes the results of the Egger Sandwich test and the combination of Egger's regression test with MLMA as described by Fernández-Castilla et al. (2019). These results were based on the same analytic methods studied in the simulations, as described in the next section.³

When applied to the effect size data without accounting for the dependence structure, the regression test indicated

³All of the regression tests reported in Table 1 were based on a modified covariate that better maintains Type I error, as evaluated in Pustejovsky and Rodgers (2019). The supplementary materials include results based on the conventional Egger's regression test, using the standard error of the effect size estimate as the predictor.

Table 1

Tests of selective reporting in working memory effect sizes from Lehtonen et al. (2018)

	Trim and Fill		3PSM			Regression Tests		
	Num. Missing	p-value	$\hat{\psi}$	χ^2	p-value	Slope ($\hat{\beta}_1$)	SE	p-value
Ignoring	0	0.5	3.833	29.093	< .001	3.405	0.855	< .001
Aggregating	0	0.5	2.201	2.369	0.124	3.201	1.319	0.009
Sampling	0	0.5	1.916	1.928	0.165	6.079	1.785	< .001
RVE						0.861	1.413	0.274
MLMA						1.009	0.815	0.108

Note: Num. Missing = Estimated number of missing studies.

MLMA = Multi-Level Meta-Analysis RVE = Robust Variance Estimation

strong evidence of small-study effects, but Trim & Fill did not indicate small-study effects, and the 3PSM indicated strong evidence that non-significant results were *more likely* to be published than significant ones. When applied to the aggregated data or a subset of data consisting of a single, randomly selected effect size from each study, the Trim & Fill test did not detect small-study effects but Egger's regression test did indicate significant funnel plot asymmetry. Based on a single randomly-selected effect size per study, the 3PSM likelihood ratio test again indicated significant evidence that non-significant results were more likely to be published than significant ones. Finally, when the dependency among effect sizes was handled using RVE or MLMA, the regression test did not indicate asymmetry in the distribution of effect sizes. The discrepancy and variation in the results of these tests further demonstrate the need for a methodological study to evaluate and compare the performance of these approaches.

Simulation methods

We conducted Monte Carlo simulations to investigate the performance of commonly used tests to detect selective reporting in meta-analyses in which dependent effect sizes are ignored or are addressed using aggregating, sampling, RVE or MLMA approaches. Due to its frequent occurrence in applied meta-analyses, we focused on a data structure where some studies assessed multiple, correlated outcomes, as in the "correlated effects" model in Equation (4). In this section, we explain the data generating process and describe the conditions manipulated in this study, along with justification for the range of parameters we evaluated.

Data generating process

We generated meta-analytic datasets based on a two-level multivariate model. To simulate effect size estimates

from a given study, we generated raw, subject level data for a two-group experimental design with multiple, correlated outcome measures. Similar to other recent simulation studies evaluating methods to detect and adjust for selective reporting (Carter et al., 2019; Moreno et al., 2009; Renkewitz & Keiner, 2018), we introduced selective reporting based on the direction and statistical significance of each effect size estimate. Specifically, we censored some or all effect sizes with one-sided p-values that exceeded .025, which is equivalent to censoring p-values that are either a) not directionally consistent with the presumed effect or b) have a two-sided p-value larger than .05. This approach results in effect-size level selection rather than censoring of full studies. Our approach also differs from the censoring process studied in Fernández-Castilla et al. (2019), who introduced selection both at the effect-size level and at the study-level.

We simulated data with variation across studies in the true effect size, the correlation between outcomes, the number of effect sizes, and the number of total participants. For simplicity, we followed the Hedges et al. (2010) conceptualization of the correlated effects model and focused on multiple measures evaluating a single construct. The true effect size was thus constant within a given study but varied between studies according to a normal distribution with mean μ and standard deviation τ . We assumed that outcome variables for a given study were equi-correlated but that the degree of correlation could vary from study to study. Introducing this variability in the correlation between outcomes allowed us to examine the impact of study conditions that depart from the default correlation assumed in RVE ($\rho = 0.8$). We sampled correlations from a beta-distribution with mean $\bar{\rho}$ and standard deviation ϕ . To determine the number of effect sizes and number of participants for each study, we sampled from the empirical distribution of effects in the meta-analysis by Lehtonen et al. (2018). Given the total number of participants, we generated data with equal sample sizes in both groups.

After determining the parameters for a given study (true effect size δ_j , outcome correlation ρ_j , number of outcomes n_j , and sample size N_j), we generated raw, participant-level data from a multivariate normal distribution within each of two groups. Within the control group, the distribution had mean zero; Within the treatment group, the distribution had mean δ_j . For both groups, the outcome distribution had unit variances and common correlation ρ_j . We then used this simulated subject level data to calculate a standardized mean difference for each outcome i within study j . These calculations included Hedges' g (Hedges, 1981) bias-correction and one-sided p-values from t-tests of each individual mean difference.

We then introduced selective reporting into the generated data by excluding some effect sizes if their one-side p-value exceeded .025. We retained all statistically significant effect sizes (i.e., $\pi^* = 1$), but censored effect sizes that were not statistically significant with probability $\pi^{censor} = 1 - \pi^{ns}$. Here, π^{censor} represents the probability of a non-significant effect size being excluded, with $\pi^{censor} = 0$ corresponding to all generated effects being observed and $\pi^{censor} = 1$ corresponding to only significant effects being observed. Using an iterative process, we generated meta-analytic datasets containing k studies, each of which contributed at least 1 effect size estimate after censoring.

Simulation Conditions

The parameter values used in the simulation study are summarized in Table 2. We used a full factorial simulation design with six factors: a) the overall average effect size, b) amount of between-study heterogeneity, c) number of primary studies included in the meta-analysis, d) the mean and e) variance of the correlation between outcomes within the same study, and f) the probability of censoring non-significant effects. This resulted in 192 conditions without selective reporting ($\pi^{censor} = 0$) and 1152 conditions where selective reporting is present with varying probabilities of censoring ($\pi^{censor} > 0$). We briefly explain the rationale for the selected conditions, most of which align with related methodological work (Carter et al., 2019; Moeyaert et al., 2017; Park & Beretvas, 2018; Pustejovsky & Rodgers, 2019; Renkewitz & Keiner, 2018; Van den Noortgate et al., 2015; Wei & Higgins, 2013).

Overall average effect size. The standardized mean difference for a given meta-analytic study was assumed to follow a normal distribution around the true average effect size of $\mu = 0.0, 0.2, 0.5$, or 0.8 , representing conventional standards for no, small, medium and large effect size values (Cohen, 1988).

Heterogeneity between studies. Between-study heterogeneity included four standard deviation values: $\tau = 0.0, 0.1, 0.2$, and 0.4 , representing a homogeneous or constant effect across included studies along with small, medium, or large degrees of heterogeneity. These values are consistent with benchmarks suggested by Pigott (2012) for the I^2 measure of heterogeneity.

Number of primary studies. We generated meta-analytic datasets with 20, 50, or 80 primary studies included, reflecting the range observed in three recent reviews of meta-analyses in education and psychological research. Meta-analyses published in top-tier journals over the past two decades reported averages of 55 and 79 included studies [Polanin et al. (2016); Ahn et al. (2012); respectively]. This is also broadly consistent with a review of meta-analyses published in 2016, which reported that the average number of primary studies ranged from 69 to 99 across meta-analyses from education, psychology, and medicine (Tipton et al., 2019).

Correlation between outcomes. We varied the correlation between outcomes at the primary study level by sampling from a beta-distribution with mean ($\bar{\rho}$) and standard deviation (ϕ). We selected two means ($\bar{\rho}$) of 0.4 and 0.8 , and two standard deviations (ϕ) of 0.0001 and 0.05 , where the former standard deviation represents near-zero variability and the latter represents moderate variability from study to study. The default correlation in the implementation of RVE is $\rho = 0.8$. Combined with near zero standard deviation, the conditions with $\bar{\rho} = 0.8$ thus reflects the RVE assumption that effect sizes within a study are equi-correlated. We included conditions with $\bar{\rho} = 0.4$ and $\phi = 0.05$ to evaluate performance when the RVE working model was not congruent with the true data-generating process. None of these conditions are fully consistent with the MLMA model, and so varying the average correlation values allows use to evaluate its robustness under stronger ($\bar{\rho} = 0.8$) or more moderate ($\bar{\rho} = 0.4$) degrees of mis-specification. This is in contrast to the approach of Fernández-Castilla et al. (2019), who generated data consistent with the MLMA model (i.e., assuming zero sampling correlation between outcomes from the same study).

Primary Study Sample Sizes. Primary study sample sizes in psychological and educational research tend to be smaller than in other research fields, with a positively skewed distribution (Fraley & Vazire, 2014). To ensure the distribution of study samples in the simulations was reflective of psychological research practices, we randomly sampled from an empirical distribution of study-level total sample sizes reported in a recent meta-analysis of the effect of bilingualism versus monolingualism on executive functioning (Lehtonen et al., 2018). From the 152 studies with 891 effect sizes included in the meta-

Table 2
Simulation parameters

Factor	Levels
Overall average effect size (μ)	0.0, 0.2, 0.5, 0.8
Between-study heterogeneity (τ)	0, 0.1, 0.2, 0.4
Number of primary studies (k)	20, 50, 80
Correlation between outcomes mean ($\bar{\rho}$)	0.4, 0.8
standard deviation (ϕ)	0.0001, 0.05
Probability of censoring (π^{censor})	0.0, 0.2, 0.4, 0.6, 0.8, 0.9, 1.0

analysis, the average sample size per primary study was 62.85 (SD = 45.04), with a median of 48 and a range of 12 to 343. This sample size distribution is generally similar to other recent meta-studies of educational and psychological research. Park and Beretvas (2018) reviewed meta-analyses from 2010 and 2015 in the *Journal of the Review of Educational Research*, reporting a median study sample size of 171, with a range from 66 to 702. However, some primary studies included in this distribution likely used multi-level samples, such as from students nested within schools, which have power similar to much smaller studies from single-level samples. In psychology, Fraley and Vazire (2014) examined a random sample of studies from social-personality research journals between 2006 to 2010 and reported median sample size of 104. The sample size distribution in the present study aligns with the smaller average sample sizes reported in journals focused more broadly on social psychology (Fraley & Vazire, 2014).

Number of Correlated Outcomes per Primary Study. Similar to primary study sample sizes, empirical meta-analyses tend to have a positively skewed distribution for the number of effect sizes per study (Ahn et al., 2012). We sampled the number of participants and the observed number of outcomes jointly from the Lehtonen et al. (2018) meta-analysis, in order to avoid imposing an artificial distribution on these study dimensions. In this database, the mean number of effect sizes per study was 5.8 (SD = 4.4) with a median of 4 (range: 1 to 18). This was similar to the distributions in recent meta-reviews in educational research, which reported an average of 3 or more effect sizes per study (Ahn et al., 2012; Park & Beretvas, 2018).

Analytic procedures

All simulations were conducted in R (Version 3.6.3; R Core Team, 2017) using the packages *metafor* (Viechtbauer, 2010) and *robumeta* (Fisher et al., 2017). To evaluate tests for selective reporting based on

aggregating effect sizes to the study level, we elected to use the simplest possible method of aggregating, by taking the simple average of effect size estimates and the simple average of the sampling variances for each study as in Equation (1). For comparative purposes, we also calculated more accurate sampling variances for the aggregated effects by assuming the average correlation between outcomes ($\bar{\rho} = 0.4$ or 0.8) used to simulate the data. Results for this alternative aggregation method are reported in the supplementary materials.

We used the *metafor* function `trimfill()` to run the Trim & Fill test with the R_0 variant after fitting a univariate random effects model. We conducted the 3PSM likelihood ratio test and Egger's regression using custom-written R code.⁴ For the univariate regression tests, we used inverse sampling variance weights of $1/W_j$ and used an additive random effects model with Knapp-Hartung adjustment for calculating standard errors, an approach that performed well in previous simulations of univariate effect sizes (Pustejovsky & Rodgers, 2019; Van Aert & Jackson, 2019). Because the Knapp-Hartung adjustment is still not broadly used in practice (Tipton et al., 2019), we also calculated Type I error results based on conventional large-sample standard errors; the supplementary materials include a comparison of results with and without the Knapp-Hartung adjustment.

For the Egger Sandwich test, we used the *robu()* function from *robumeta* with a correlated effects working model (Hedges et al., 2010), assuming $\rho = 0.8$. This function incorporates small-sample adjustments and Satterthwaite degrees of freedom for hypothesis tests (Tipton, 2015). For the Egger MLMA Test, we used the *metafor* function `rma.mv()` to estimate a three-level model that included between- and within-study random

⁴The *weighttr* package (Coburn & Vevea, 2019) can also be used to estimate the 3PSM. We opted to write custom R code in order to address convergence issues that occur when all of the simulated effect sizes are statistically significant or all effect sizes were not significant. The supplementary materials provide further details about implementation.

effects. For certain combinations of parameters the Egger MLMA test failed to converge. The supplementary materials provide details on non-convergence rates, which have a maximum of 0.26% for Type-I error conditions and 2.14% for power conditions. To determine the statistical significance of small-study effects or selective reporting, we used one-sided tests for each of the methods. The regression tests and Egger Sandwich used the modified covariate that better maintains the Type I error rate than the conventional standard error covariate (Pustejovsky & Rodgers, 2019). We also evaluated potential differences between the weights used in the `robumeta` package versus weighted least squares estimation with full inverse variance weights based on a correlated effects working model. The supplementary materials provide additional details.

Performance criteria

Performance criteria included Type I error rates and power to detect selective reporting. We calculated rejection rates when no selective reporting was present to determine if the Type I error rates exceed a nominal $\alpha = .05$ level. For the 192 conditions where there was no selective reporting, we present results for all tests and all methods of handling dependent effects. For the tests and methods that maintain close-to-nominal Type I error rates, we calculated rejection rates across 1152 combination of conditions with a positive probability of censoring to determine the relative power to detect selective reporting. With 5,000 replications and a nominal alpha of .05, the Monte Carlo standard error in the Type I error and power rates was less than 0.003, indicating very precise estimates of the performance criteria. Complete R code and numerical simulation results are available at <https://bit.ly/2qnF4dH>.

Simulation results

Type I error rates

We focused on Type I error rates of the methods at the conventional level of $\alpha = .05$. All figures show results for conditions where the correlation between outcomes within a primary study were generated with $\bar{\rho} = 0.4$ and $\phi = 0.05$. The supplementary materials include graphs comparing conditions for all three tests when the correlation between dependent outcomes varies from $\bar{\rho} = 0.4$ to 0.8, with zero or moderate between-study heterogeneity. The results are very similar to those reported here.

Some researchers have argued for using a more liberal nominal Type-I error rate of $\alpha = .10$ due to the limited

power of tests for small-study effects or selective reporting (Egger et al., 1997; Sterne & Egger, 2001). The supplementary materials also include figures presenting simulation results when the nominal Type I error rate is $\alpha = .10$. The trends for all three tests and all methods of handling dependence are very similar at both α levels.

Finally, the figures in the main text present results for aggregating effect sizes are based on the simple aggregation method from Equation (1). The supplementary materials include simulation results for the aggregation method using the true correlation between outcomes.

Figure 1 depicts the Type I error rates for the Trim & Fill test when dependence is ignored and when effect sizes are aggregated or sampled (columns) under varying degrees of heterogeneity (rows). Each line depicts rejection rates for meta-analyses with a different number of primary studies ($k = 20, 50, 80$) across varying levels of overall average effect size (horizontal axis). When dependence is ignored (first column), Trim & Fill results far exceed the nominal alpha level—especially as the average effect size, amount of heterogeneity, or number of studies increases. Aggregating (second column) or selecting one effect size per study (third column) maintains Type I error rates under most conditions. Rejection rates gradually increase above the nominal level as between-study heterogeneity increases. Aggregating performs slightly better than sampling when heterogeneity is low and the number of studies is larger ($k = 50, 80$). There are no differences in the Trim & Fill Type I error rates when the true correlation is used for aggregation, regardless of mean effect size, true correlation, and level of heterogeneity; the supplementary materials provide further details.

Figure 2 presents Type I error rates for the likelihood ratio test based on a 3PSM, following the same format as Figure 1. Similar to Trim & Fill, the likelihood ratio test also suffers from severely inflated Type I error rates when dependence is ignored. When aggregating or sampling is used to handle dependent effect sizes, the rejection rates are closer to nominal levels when the sample size is small ($k = 20$) or heterogeneity is large. When the average effect size is large ($\mu = 0.8$) and between-study heterogeneity is low or moderate ($\tau = 0.0, 0.1, 0.2$), aggregating effect sizes leads to Type I error rate inflation above .1. Compared to using simple averaging for aggregation, aggregating effect sizes based on the true correlation leads to better-calibrated Type I error rates. However, the error rates are still above nominal under some conditions, particularly at smaller sample sizes or when average effect sizes are large; the supplementary materials provide further details. Compared to ignoring or aggregating effect sizes, type I error rate inflation is less severe when sampling a single effect size per study,

Figure 1. Type I error rates for Trim & Fill when dependent effects are ignored, aggregated or sampled, by number of primary studies, level of heterogeneity, and mean effect size.

Figure 2. Type I error rates for 3PSM likelihood ratio test when dependent effects are ignored, aggregated, or sampled, by number of primary studies, level of heterogeneity, and mean effect size.

although rates still tend to exceed the nominal .05 level.

Figure 3 presents the Egger's regression tests using the modified covariate. The figure includes two additional columns that present the Type I error rates when handling dependent effects using robust variance estimation (Egger Sandwich) or multi-level meta-analysis (Egger MLMA). Similar to the Trim & Fill and 3PSM methods, when dependence is ignored, the regression tests have inflated Type I error rates at all levels of heterogeneity, average effect sizes, and study sample sizes. Across conditions, the nominal alpha level is well maintained when dependent effects are aggregated, sampled, or handled with RVE or MLMA. When dependent effects are aggregated, rejection rates are slightly elevated if there is any degree of between-study heterogeneity. As presented in the supplementary materials, aggregation using the true correlation can improve the precision of the sampling variance estimates, which decreases the Type I error rate to close to the nominal level. This is particularly evident when $\bar{\rho} = 0.4$ and there is low or moderate between-study heterogeneity ($\tau = 0.1, 0.2$).⁵

When used with Egger's regression test, sampling, RVE, and MLMA have nearly identical rejection rates, all of which are very close to nominal. Even though the MLMA model is mis-specified (i.e., it assumes zero sampling correlation between effect sizes), it nonetheless maintains close-to-nominal Type I error rates across all conditions examined. The MLMA rates are very slightly elevated above the nominal rate when between-study heterogeneity is low or moderate ($\tau = 0.1$ to 0.4) and the number of primary studies is small ($k = 20$), but this small elevation disappears at larger sample sizes. Using RVE with the Egger Sandwich test maintains Type I error at or slightly below the nominal level. The supplementary materials include a more detailed figure comparing Type I error rates for RVE and MLMA.

The supplementary materials also include results for the Egger Sandwich test using two different weighting methods: the semi-efficient weighting scheme implemented in the `robumeta` package versus full inverse variance weights based on the correlated effects working model. Differences between weighting schemes do not substantially influence the Type I error rates, and both approaches yield tests that maintain the nominal alpha level. An acceptable Type I error rate is maintained even when the true correlation between effects is $\rho = .4$, so that the working model is mis-specified. When heterogeneity is low ($\tau = 0.1, 0.2$), the Egger Sandwich test using semi-efficient weights has a slightly lower Type I error rate than the test that uses fully inverse-variance weights, regardless of the number of included studies.

Power

We now consider the relative power of the three tests to detect selective reporting when using aggregating, sampling, RVE and MLMA approaches to address dependency. We do not discuss power for the tests where dependency is ignored because this approach does not adequately control Type I error. However, we did include some methods in the figures even though they do not exactly maintain Type I error under all conditions. In the following figures, gray bands indicate methods with excess Type I error. We opted to include such methods for comparative purposes because the deviations from nominal Type I error are less severe than when ignoring dependence and because some of the methods are commonly used in practice. Thus, power of the Trim & Fill and 3PSM should be interpreted with caution as they have inflated Type I error rates when dependent effects are combined with aggregating or sampling.

Figure 4 presents the power rates (vertical axis) of the tests as a function of the degree of censoring (horizontal axis, running from 0 to 1) for a sample size of $k = 80$ studies, within-study correlation between effect sizes $\bar{\rho} = 0.4$ and $\phi = 0.05$, and varying levels of average effect size (columns) and heterogeneity (rows). This figure compares the aggregation approach to the RVE and MLMA approaches for handling dependent effects. Across degrees of selective publication, the regression tests and Trim & Fill have limited power, particularly when the true effect size is near zero. Adequate power (i.e. the conventional level of 80%) is only obtained with a moderate to large average effect size, low heterogeneity, and strong selective publication censoring ($\pi^{\text{censor}} = 1$). There are differences in power between the regression tests, with the test based on aggregated effect sizes having slightly more power than RVE, but less than MLMA when heterogeneity is low or moderate ($\tau = 0.0, 0.1, 0.2$) and the average effect size is greater than zero. However, the slight advantage is due in part to the inflated Type I error rate of the regression test based on aggregated effect sizes. Using effect sizes aggregated to the study level, the 3PSM likelihood ratio test has substantially higher power than the other tests, under every condition examined. However, this apparent advantage is partially due to the inflated Type I error rates of the likelihood ratio test, which occur when average effects are large ($\mu = 0.8$). At smaller numbers of studies per meta-analysis, the power advantage of the 3PSM likelihood

⁵Additionally, Egger's regression tests while aggregating or sampling did not perform differently when the large-sample standard errors were used instead of the Knapp-Hartung corrected standard errors. See supplementary materials for further details.

Figure 3. Type I error rates for regression tests when dependent effects are ignored, aggregated, sampled, handled using Robust Variance Estimation (RVE) or Multi-Level Meta-Analysis (MLMA), by number of primary studies, level of heterogeneity, and mean effect size.

ratio test is not as large compared to other tests based on aggregated effects (See supplementary materials for figures).

Figure 5 presents the power rates of tests when a single effect size is randomly sampled from each study. Results for RVE and MLMA approaches are also depicted here for comparative purposes. Results are very similar to those in Figure 4. The regression tests based on sampling a single effect size have nearly identical power to the Egger Sandwich test, across all conditions, and the

Egger MLMA continues to have a small power advantage over the other regression tests.

Trends in the relative power of the tests are for meta-analyses involving a smaller number of studies. The supplementary materials provide plots similar to Figures 4 and 5, but with $k = 20$ and $k = 50$ studies per meta-analysis. As would be expected, the power of each of the tests is larger when the number of studies per meta-analysis is larger, and power is stronger for larger mean effect sizes. When the true correlation among effect sizes

Figure 4. Power of tests for selective reporting when dependent effects are aggregated or handled using Robust Variance Estimation (RVE) or Multi-Level Meta-Analysis (MLMA), for samples of $k = 80$ studies and between study correlation of $\rho = 0.4$. Gray bands indicate tests with excess Type-I error.

Figure 5. Power of tests for selective reporting when dependent effects are sampled or handled using Robust Variance Estimation (RVE) or Multi-Level Meta-Analysis (MLMA), for samples of $k = 80$ studies and between study correlation of $\rho = 0.4$. Gray bands indicate tests with excess Type-I error.

is $\bar{p} = 0.8$ (rather than $\bar{p} = 0.4$ as shown here), power generally tends to be lower for all tests, but the relative performance of the tests remains similar. In general, power of the selection reporting tests that maintain Type I error rates tends to improve when there is less between-study heterogeneity, larger mean effect size, or a higher probability of censoring.

Discussion

This study contributes to the limited research base on the performance of tests for selective reporting or small-study effects when combined with common methods of handling dependent effect size estimates. Using Monte Carlo simulation, we generated meta-analytic datasets composed of primary studies with standardized mean differences and other design factors similar to those encountered in meta-analyses of education and psychological research. We introduced censoring based on the sign and statistical significance of individual effects, with the strength of selection varying from none to total censoring of non-significant studies.

For all three tests examined, ignoring dependence results in severely inflated Type I error. When aggregating effect sizes at the study level, the Trim & Fill test maintains close-to-nominal Type I error except when heterogeneity is large ($\tau = 0.4$). For the 3PSM likelihood ratio test, aggregating effects results in elevated Type I error rates when heterogeneity is small or the true mean effect is large ($\mu = 0.8$).⁶ Regression tests based on aggregated effects have slightly inflated Type I error rates when between-study heterogeneity is moderate ($\tau = 0.1$ or 0.2). The slight elevation results from taking the simple average of the effect size variances, which leads to positively biased variances for the aggregated effects. When the sampling variances are calculated using the true correlation between effect size estimates (Borenstein et al., 2009), the Type I error for the regression tests improves to nearly nominal levels. However, this approach to aggregation will often be infeasible in practice because the information needed to calculate correlations among effect size estimates is infrequently reported (Becker, 2000).

Sampling one effect size per study follows a broadly similar pattern to aggregating, but Type I error rates are closer to nominal for the 3PSM and regression tests. This is because randomly selecting a single effect size from each study establishes a subset of independent effect sizes that is representative of the original meta-analytic dataset, and where the variances of the selected effect sizes are accurately estimated. Finally, the Egger Sandwich and Egger MLMA tests, which combine Egger's

regression test with RVE and MLMA respectively, both maintain adequate Type I error.

Under conditions where non-significant effect sizes are censored, power to detect selective reporting is less than ideal, regardless of what method is used to handle dependence. With the exception of 3PSM, the tests lack power to detect selective reporting except under strong censoring (i.e., when non-significant results have a very small chance of being published). When dependent effect sizes are aggregated or sampled, the 3PSM likelihood ratio test has elevated Type I error rates under certain conditions, but also has substantially higher power than the other tests. Trim & Fill has the least amount of power of any of the tests—a trend that is similar to findings from previous methodological studies (Carter et al., 2019; Macaskill et al., 2001).

Our simulations provide the first direct comparison between the Egger MLMA and Egger Sandwich test. Both approaches are acceptable, valid tests (in terms of Type I error). The Egger MLMA test has a small but consistent power advantage relative to the Egger sandwich and other variants of Egger's regression. However, all of these tests need to be interpreted with caution. These statistical methods still have limited power to detect funnel plot asymmetry and, in practice, such asymmetry may have other causes besides selective reporting.

Our findings regarding the performance of the Egger MLMA test are broadly consistent with Fernández-Castilla et al. (2019), who also demonstrated that the test (which they termed the Funnel Plot test) maintains nominal Type I error rates but has limited power to detect small-study effects. The present simulations differ from Fernández-Castilla et al. (2019) in terms of the data-generating processes and censoring mechanisms examined. Specifically, Fernández-Castilla et al. (2019) simulated data based on a correctly specified MLMA model, then used a multi-step censoring process which involved selection at both the effect-size level and the study-level. In contrast, we simulated data from a model with correlation among the effect size estimates and then introduced selection at the level of effect size estimates only. Considering that the MLMA is mis-specified under our data-generating process, it is remarkable that the MLMA nonetheless maintains nominal Type-I error. Under the range of models we examined, inferences based on the MLMA remain robust to mis-specification, consistent with the findings of Van den Noortgate et al. (2015).

⁶The performance of 3PSM likelihood ratio test improves when effect sizes are aggregated using the true correlation among effect sizes rather than using a simple average, although it still has above-nominal Type I error under many conditions.

Limitations and future directions

This simulation study has several limitations, several of which points towards useful directions for future research. First, our approach to introducing selective reporting in the generated data was based on a model where the chance that an effect size estimate is censored depends on whether it passes a single threshold level of statistical significance. This data generating process directly aligns with the underlying assumptions of the 3PSM likelihood ratio test, which may explain why the 3PSM likelihood ratio test exhibits stronger power than alternative tests examined. Future research should explore the performance of these methods under more complex selection models, such as using multiple significance thresholds or censoring based not only on significance but also on the magnitude of effect sizes (Carter et al., 2019) or the primary study sample size (Renkewitz & Keiner, 2018). Using a variety of different selection mechanisms, some of which do not align with the assumptions of the analytic models and tests, will provide a robust evidence base about the relative strengths and weaknesses of funnel plot asymmetry tests and selection models.

Simulation conditions in this study were designed to emulate research syntheses in education and psychology. However, results are only generalizable to meta-analyses with similar conditions, and specifically, those synthesizing standardized mean differences from a two group between-subjects design with multiple correlated effect sizes and with a similar distribution of primary study sample sizes. That ignoring dependence leads to inflated Type I error rates can reasonably be generalized to effect sizes metrics other than standardized mean differences. However, the exact Type I error rates when aggregating, sampling or handling dependent effects with RVE or MLMA may require further study based on the fact that there are some differences in performance of Egger regression across different effect size indices. Further, the performance characteristics of the tests—and power, in particular—may be influenced by the distribution of primary study sample sizes. In practice, meta-analyses often include more complex data structures with varying types of dependency. Future research should examine how methods to handle dependency and test for selective reporting perform under other dependency structures and with sample size distributions informed by the conventions of other research areas.

Our findings highlight the influence of between-study heterogeneity on the performance of tests for selective reporting or small-study effects. The simulations only included unexplained heterogeneity, yet many meta-analyses also investigate the role of moderators in ex-

plaining heterogeneity. To our knowledge, there has been little methodological development of how to test for small-study effects or selective reporting in meta-analyses where moderators partially explain between- or within-study heterogeneity. A recent review of meta-analytic studies in psychology and medicine discusses the connection between publication bias and heterogeneity, but the reported results are limited in scope as they focused on detecting selective reporting only within homogeneous sub-samples of effect size estimates (Van Aert et al., 2019). Future research should consider how to incorporate study- or effect size-level moderating factors when testing for selective reporting in databases of dependent effect sizes.

Finally, the methods selected in this study are not comprehensive of all available selective reporting detection strategies. For instance, we have not considered methods such as p-curve (Simonsohn et al., 2014), p-uniform (Van Aert et al., 2016), or more complex selection models. Future methodological work should consider how such methods could be extended to handle dependent effects. Although the 3PSM likelihood ratio test did not adequately control Type I error, its relatively strong power suggests that tests based on selection models are a promising direction for further development

Implications for practice

Based on the results of this study, we offer a few general recommendations for applied researchers and for future methodological work. First, applied meta-analysts should not ignore dependent effect sizes when conducting tests for small-study effects or selective reporting. All three tests have inflated Type I error rates across all varying study conditions examined when ignoring dependence. In other words, if dependent effect sizes are ignored, these tests are overly likely to detect selective reporting or small-study effects tests when it is not actually present. This finding is worth emphasizing because it is still common for researchers to simply ignore dependency when applying univariate tests, as seen in some recent meta-analyses (Clark et al., 2016; Fisher et al., 2016; Weisz et al., 2017).

As an alternative to ignoring dependence, our findings suggest that conducting a modified regression-based test, using RVE or MLMA to handle dependence, maintains Type I error rates. The RVE and MLMA methods are both relatively simple to run using the `robu()` function in the `robumeta` (Fisher et al., 2017), and `rma.mv()` function in the `metafor` package in R. A measure of precision for each effect size can be included as a covariate in the meta-regression formulas within these functions. We recommend using a modified covariate that

removes artifactual correlation between the standardized mean difference effect size and its related variance, as demonstrated in Pustejovsky and Rodgers (2019). Researchers interested in implementing these methods can access the annotated code we used to analyze the Lehtonen et al. (2018) data, which is included in the supplementary materials (See <https://bit.ly/2qnF4dH>). In addition, both approaches can, in principle, be extended to also incorporate substantive moderator variables, while also testing for small study effects.

In situations where either RVE or MLMA could feasibly be used to handle dependence, we recommend using the Egger MLMA test because it has a power advantage over the Egger sandwich test. In some applications, however, researchers might have reason to prefer to use only RVE or only MLMA when conducting meta-analysis or meta-regression analysis of dependent effects. In such situations, we see it as reasonable to conduct a test for small-study effects that aligns with the overall approach to handling dependence in the other analyses. In other words, if the overall analysis uses RVE to handle dependent effects, then it would be acceptable to simply report the Egger's sandwich test rather than switching approaches and implementing MLMA to handle dependence.

Whichever framework is used, it is important to bear in mind that both all variants of Egger's regression test have two notable limitations. First, if a significant result for the slope coefficient indicates funnel plot asymmetry, which could be a result of mechanisms other than just selective reporting. Second, the method suffers from a lack of power to detect significant funnel plot asymmetry when non-significant effects are censored. In other words, the test may not detect small study effects when in fact they are truly present in the meta-analytic dataset, and researchers should include this as a limitation when interpreting Egger Sandwich or Egger MLMA results.

Alternatively, the 3PSM likelihood ratio test addresses limitations of these modified regression-based tests by more directly testing for selection and by providing increased power. Currently, the 3PSM test can only be applied to univariate datasets, and therefore aggregating or sampling approaches must be applied prior to detecting selective reporting. However, these approaches do not fully control Type I error, and therefore are not yet ready for use in primary or confirmatory analyses. When testing for selective reporting, practitioners should consider taking an exploratory approach by reporting multiple tests, such as the Egger MLMA and Egger Sandwich tests, along with exploratory 3PSM likelihood ratio tests after aggregating or sampling effect sizes. Researchers might also summarize the results of multiple 3PSM tests after drawing repeated random samples of a single effect size from each primary study. Additionally, sensitivity

analyses could be conducted with more complex selection models that include different levels of censoring and study design characteristics specific to the field of study (Coburn & Vevea, 2015; Vevea & Woods, 2005).

In sum, evidence from present study primarily suggests cautionary guidance regarding how applied meta-analysts should investigate selective reporting. Detecting selective reporting is a challenging task, and currently there are few methods that perform well across a wide range of conditions. While methodologists should continue to improve and extend methods of detecting and adjusting for selective reporting, a broad preventative effort may ultimately be more important. Research incentives need to be restructured to encourage documentation of and access to study results regardless of statistical significance. Academic societies, professional societies, and funders should encourage or even require researchers to register their studies prior to conducting them, update results or changes in protocol for ongoing studies, and report final results. Such steps, along with improved editorial practices that emphasize aims and design rather than statistical significance, might one day mitigate the threat of selective reporting and obviate the need to test for selective reporting in meta-analyses.

References

- Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, 82(4), 436–476. <https://doi.org/10.3102/0034654312458162>
- Becker, B. J. (2000). Multivariate meta-analysis. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). Elsevier.
- Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin*, 144(1), 77–110. <https://doi.org/10.1037/bul0000130>
- Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4), 1088. <https://doi.org/10.2307/2533446>
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Multiple outcomes or timepoints within a study. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd, 225–238.
- 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(2), 115–144.
- Cheung, M. W.-L. (2019). *A guide to conducting a meta-analysis with non-independent effect sizes* (Preprint). PsyArXiv. <https://doi.org/10.31234/osf.io/5p7dj>
- Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, 86(1), 79–122. <https://doi.org/10.3102/0034654315582065>
- Coburn, K., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods*, 20(3), 310.
- Coburn, K., & Vevea, J. L. (2019). *WeightR: Estimating weight-function models for publication bias*. Retrieved from <https://CRAN.R-project.org/package=weightR>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd. Hillsdale, NJ: Erlbaum.
- Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin*.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.
- Debray, T. P., Moons, K. G., & Riley, R. D. (2018). Detecting small-study effects and funnel plot asymmetry in meta-analysis of survival data: A comparison of new and existing tests. *Research Synthesis Methods*, 9(1), 41–50.
- Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449), 89–98. <https://doi.org/10.2307/2669529>
- Egert, F., Fukkink, R. G., & Eckhardt, A. G. (2018). Impact of in-service professional development programs for early childhood teachers on quality ratings and child outcomes: A meta-analysis. *Review of Educational Research*, 88(3), 401–433.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109), 629–634.
- 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(1), 120–128. <https://doi.org/10.1037/a0024445>
- Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2019). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education*, 1–20.
- Fisher, B., Gardella, J., & Teurbe-Tolon, A. (2016). Peer cybervictimization among adolescents and the associated internalizing and externalizing problems: A meta-analysis. *Journal of Youth and Adolescence*, 45(9), 1727–1743. <https://doi.org/10.1007/s10964-016-0541-z>
- Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Robumeta: Robust variance meta-regression*. Retrieved from <https://CRAN.R-project.org/package=robumeta>

- Fraley, R. C., & Vazire, S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, 9(10), e109019.
- Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, 83(4), 428–445.
- Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In *The handbook of research synthesis and meta-analysis* (pp. 357–376). New York: Russell Sage Foundation.
- Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, 555(7695), 175.
- Hedberg, E. (2014). ROBUMETA: Stata module to perform robust variance estimation in meta-regression with dependent effect size estimates.
- Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128.
- Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2), 246–255.
- Hedges, L. V. (2019). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 281–298). New York: Russell Sage Foundation.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39–65. <https://doi.org/10.1002/jrsm.5>
- Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 145–174). Chichester, England: John Wiley & Sons.
- Hunter, J. P., Saratzis, A., Sutton, A. J., Boucher, R. H., Sayers, R. D., & Bown, M. J. (2014). In meta-analyses of proportion studies, funnel plots were found to be an inaccurate method of assessing publication bias. *Journal of Clinical Epidemiology*, 67(8), 897–903.
- Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3), 245–253.
- Kalaian, H., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, 1(3), 227.
- Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., Bruin, A. de, & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*.
- Light, R. J., & Pillemer, D. B. (1984). *Summing up*. Harvard University Press.
- Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20(4), 641–654. <https://doi.org/10.1002/sim.698>
- Marin-Martinez, F., & Sanchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology*, 2, 32–38.
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis an evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5), 730–749.
- Moeyaert, M., Ugille, M., Beretvas, S. N., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). 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*, 20(6), 559–572.
- Moreno, S. G., Sutton, A. J., Ades, A., 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). <https://doi.org/10.1186/1471-2288-9-2>
- Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature News*, 506(7487), 150.
- Park, S., & Beretvas, S. N. (2018). Using total sample size weights in meta-analysis of log-odds ratios. *Journal of Experimental Education*.
- Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Jama*, 295(6), 676–680.
- Pigott, T. (2012). *Advances in meta-analysis*. Springer Science & Business Media.

- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2019). *nlme: Linear and nonlinear mixed effects models*.
- Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes a meta-review. *Review of Educational Research*, 86(1), 207–236.
- Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10(1), 57–71. <https://doi.org/10.1002/jrsm.1332>
- Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, 103(1), 111.
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Renkewitz, F., & Keiner, M. (2018). How to detect publication bias in psychological research? A comparative evaluation of six statistical methods. Retrieved from <https://psyarxiv.com/w94ep/>
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.
- Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, 84(3), 328–364.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6), 666–681. <https://doi.org/10.1177/1745691614553988>
- Sterne, J., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, 54(10), 1046–1055.
- Stevens, E. A., Rodgers, M. A., & Powell, S. R. (2017). Mathematics interventions for upper elementary and secondary students: A meta-analysis of research. *Remedial and Special Education*, 0741932517731887. <https://doi.org/10.1177/0741932517731887>
- Sutton, A. J. (2009). Publication bias. In *The handbook of research synthesis and meta-analysis* (pp. 435–445). Russell Sage Foundation.
- Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in stata and spss. *Research Synthesis Methods*, 5(1), 13–30.
- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20(3), 375–393. <https://doi.org/10.1037/met0000011>
- Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 40(6), 604–634. <https://doi.org/10.3102/1076998615606099>
- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, 10(2), 180–194.
- Van Aert, R., & Jackson, D. (2019). A new justification of the hartung-knapp method for random-effects meta-analysis based on weighted least squares regression. *Research Synthesis Methods*, 10(4), 515–527.
- Van Aert, R., Wicherts, J. M., & Van Assen, M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, 11(5), 713–729.
- Van Aert, R., Wicherts, J. M., & Van Assen, M. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLoS One*, 14(4), e0215052.
- Van Assen, M., Van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20(3), 293.
- 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(2), 576–594. <https://doi.org/10.3758/s13428-012-0261-6>
- 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(4), 1274–1294. <https://doi.org/10.3758/s13428-014-0527-2>
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419–435. <https://doi.org/10.1007/BF02294384>
- Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, 10(4), 428–443. <https://doi.org/10.1037/1082-989X.10.4.428>
- Viechtbauer, W. (2010). Metafor: Meta-analysis package for r. *R Package Version*, 2010, 1–0.
- Wei, Y., & Higgins, J. (2013). Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(7), 1191–1205. <https://doi.org/10.1002/sim.5679>
- Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., ... others. (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. *American Psychologist*, 72(2), 79.
- Zheng, B., Warschauer, M., Lin, C.-H., & Chang, C. (2016). Learning in one-to-one laptop environments: A meta-analysis and research synthesis. *Review of Educational Research*, 86(4), 1052–1084.