18

PUBLICATION BIAS

JACK L. VEVEA
University of California, Merced

KATHLEEN COBURN
University of California, Merced

ALEXANDER SUTTON
University of Leicester

CONTENTS

18.1 INTRODUCTION

There is good evidence to suggest that unpublished scientific results may systematically differ from published results, because selectivity may exist in deciding what to publish (see Dickersin, Min, and Meinert 1991, 1992; Song et al. 2000; Dickersin 2005). That phenomenon is frequently referred to as publication bias. For example, researchers may choose not to write up and submit studies with uninteresting or nonsignificant findings, or such studies may not be accepted for publication. Although publication bias refers to whether work is published, unpublished work still available for inclusion in meta-analyses does not technically contribute to bias in those

specific meta-analyses, even though the published studies themselves are a biased sample.

Examples of publication bias are everywhere. Philippa Easterbrook and her colleagues document the role of the perceived importance of findings in determining which to submit for publication (1991). Allan Coursol and Edwin Wagner present evidence of the role of statistical significance in the publication process (1986). Jerome Stern and John Simes offer evidence that significant results are often published more quickly (1997). An-Wen Chan and his colleagues point out that, even if a study is published, there may be selectivity in which aspects are presented; significant outcomes may be given precedent over nonsignificant ones (Chan, Hrobjartsson, et al. 2004; Chan, Krleza-Jeric, et al. 2004). That is, any selection mechanism may operate through suppression of particular results within a study, or all results from a particular sample may be affected. Sven Kepes and his colleagues discuss the distinction (2012). Additionally, research with positive or statistically significant results may be published in more prestigious venues and cited more times, making it more visible and easier to find (Koricheva 2003; Egger and Smith 1998). Indeed, the publication process should be thought of as a continuum and not a dichotomy (Smith 1999). For example, material that has been published with incomplete reporting in a journal may have been circulated with full reporting as a working paper. In keeping with the previous literature, these biases will be referred to simply as publication bias throughout the chapter, although dissemination bias is perhaps a more accurate name for the collection (Song et al. 2000).

In areas where any such selectivity exists, the literature is biased. That is true whether one is reading a single journal article or conducting a synthesis of many. Publication bias is therefore a major threat to the validity not only of meta-analysis and other synthesis methodologies, but also of the research literature itself. Indeed, one could argue that meta-analysis provides a partial solution to the problem, because researchers can at least attempt to identify and estimate the effect of such bias by considering the information contained in the distribution of effect sizes from the available studies. That is the basis of the majority of statistical methods described here. It is important to note that most of these methods have been developed for use with the meta-analytic models advanced in the tradition of Larry Hedges and Ingram Olkin (1985). Many methods for testing and correcting publication bias are not suitable for the psychometric meta-analysis approaches proposed by James Hunter and Frank Schmidt (1990).

Researchers agree that prevention is the best solution to the problem of selectively reported research. Indeed, with advances in electronic publishing making the presentation of large amounts of information more economically viable than traditional paper-based publishing methods, there is some hope that the problem will diminish, if not disappear. Many have suggested that open-access publication can assist with the problem (see, for example, Joober et al. 2012), but much of this advocacy appears in blog entries or in the mission statements of electronic journals. There is still little or no empirical evidence of such an effect. Ridha Joober and his colleagues also point to the possibility that high fees associated with open-access publication could actually lead to publication bias (2012). Moreover, open access does not offer a solution to the suppression of information due to vested economic interests (Halpern and Berlin 2005).

Jesse Berlin and Davina Ghersi, among others, have advocated the use of prospective registries of studies for selecting studies to be included in systematic reviews (2005). The practice provides an unbiased sampling frame guaranteeing the elimination of publication bias (relating to the suppression of whole studies, at least). However, trial registration does not guarantee availability of data, and an obligation to disclose results in an accessible form is also required. Registries exist for randomized controlled trials in numerous medical areas, and there is an expectation that this practice will ultimately reduce publication bias (Zarin et al. 2011). Such a solution will not be feasible for some forms of research, however, including research relating to analysis of observational data, where the notion of a study that can be registered before analysis may be nonexistent. The idea of registries for research in the social sciences has been put forth but it is far from the norm, and controversy surrounds the effectiveness of preregistration in that context (Anderson 2013; Gelman 2013; Humphreys, de la Sierra, and van der Windt 2013; Monogan 2013). The notion of prospectively designing multiple studies with the intention of carrying out a meta-analysis in the future has also been put forward as a solution to the problem (Berlin and Ghersi 2005), but again may be difficult to orchestrate in many situations.

Carrying out as comprehensive a search as possible when obtaining literature for a synthesis will help minimize the influence of publication bias. In particular, this may involve searching for studies not formally published (chapter 6 in this volume; Hopewell, Clarke, and Mallett 2005), as well as using methods other than simple electronic

searches (such as journal browsing and reference chasing). Since the beginning of the internet, the feasibility and accessibility of publication by means other than commercial publishing houses have greatly increased.

Despite researchers' best efforts, at least in the current climate, alleviation of the problem of publication bias may not be possible in many areas of science. In such instances, graphical and statistical tools have been developed to address publication bias within a meta-analysis framework. The remainder of this chapter provides an overview of these methods. If a research synthesis does not contain a quantitative synthesis (for example, if the data being synthesized are not quantitative), publication bias may still be a problem, but methods to deal with it are limited to prevention through registration and rigorous literature searches (Petticrew et al. 2006). Terese Bondas and Elisabeth Hall suggest that careful identification of unpublished studies, such as dissertations, may help, but that has not proven to be consistently effective for quantitative synthesis, so it may be of limited value for qualitative synthesis (2016). Simon Lewin and his colleagues observe that evidence of publication bias in qualitative literature is lacking (2015). They also state that methodological advances are in development, but are not currently available.

18.2 MECHANISMS THAT CAUSE PUBLICATION BIAS

There is considerable discussion in the literature about the precise nature of the mechanisms that lead to suppression of whole studies and other forms of publication bias. These mechanisms may operate on specific results within a particular study (outcome bias) or on the entire study (dissemination bias). Both of these levels can contribute to the overall presence of publication bias (Kepes et al. 2012). If these mechanisms could be accurately specified and quantified, then the appropriate adjustments to a meta-analytic data set would be straightforward. However, measuring such effects is difficult, and the mechanisms vary with data set and subject area.

Evidence is ample that statistical significance, effect magnitude and direction, study size, and other factors can all influence the likelihood of a study being published. Colin Begg and Jesse Berlin address the role of *p*-values and direction of effect (1988). Harris Cooper, Kristina DeNeve, and Kelly Charlton confirm the existence of filters in the research process other than bias against the null hypothesis (1997). Robert Rosenthal and John Gaito present evidence for cliff effects associated with conventional levels of significance (1963, 1964), as do Nanette Nelson, Robert Rosenthal, and Ralph Rosnow (1986). Deborah Barnes and Lisa Bero show that funding source can lead to selection bias (1998). Justin Bekelman, Yan Li, and Cary Gross discuss the role of industry funding (2003). Kathleen Coburn and Jack Vevea mention industry funding and preferences for results that are consistent with current beliefs, trends, and cultural expectations as sources of bias (2015; see also Kepes, Banks, and Oh 2014; Kepes, Bennett, and McDaniel 2014). José Duarte and his colleagues also provide evidence that social preferences can influence publication (2015), citing the work of Stephen Abramowitz, Beverly Gomes, and Christine Abramowitz, that liberal reviewers were less likely to publish research with results favoring conservatives (1975), and of Stephen Ceci, Douglas Peters, and Jonathan Plotkin, that "reverse discrimination" proposals were approved less often (1985).

The sections that follow outline and demonstrate methods to identify and adjust for publication bias. These methods assume different underlying mechanisms for publication bias, and all of those assumptions are wrong. Accordingly, the focus in these sections includes not only an up-to-date overview of available methods, but also attention to the assumptions of each approach. It is not plausible, for example, that publication bias occurs solely because of statistical significance, or that it arises purely from a relationship between effect size and standard error, or that it follows a deterministic pattern, such as elimination of the largest negative effects. For that reason, these methods should be regarded as tools for sensitivity analysis, and triangulation using multiple techniques is essential. Kepes and McDaniel propose reporting a range of estimates across various methods to assess the effect of publication bias (2015). Reporting standards for meta-analysis endorsed by the American Psychological Association (2008) and the Cochrane Collaboration (Higgins and Green 2011) also recommend this approach. Despite these suggestions, evidence indicates that only about 3 percent of meta-analyses use more than two procedures to address publication bias (Ferguson and Brannick 2012; van Enst et al. 2014).

18.3 METHODS FOR IDENTIFYING PUBLICATION BIAS

18.3.1 The Funnel Plot

Since its introduction by Richard Light and David Pillemer in 1984, the funnel plot has been a preferred exploratory

tool for investigating publication bias and, like the forest plot, for presenting a visual summary of a meta-analytic data set (Sterne, Becker, and Egger 2005). In its original form, the funnel plot is a scatterplot with effect estimates on the horizontal axis and sample size on the vertical axis. Sample size is closely related to study precision, which is usually defined as the reciprocal of either the sampling variances or the standard errors of the effect sizes. More recent forms of the funnel plot typically use such a measure of precision (or, alternatively, the reciprocal of precision) in place of sample size. The expectation is that the plot should appear symmetric with respect to the distribution of effect sizes and should resemble a funnel. The effect sizes should be evenly distributed around the underlying true effect size, and show more variability in the smaller studies than the larger ones because of the greater influence of sampling error. This results in a funnel-shaped plot that narrows as study precision increases. If publication bias is present, we might expect some suppression of smaller, unfavorable, and nonsignificant studies that could be identified by a gap in one corner of the funnel or a decrease in density nearer the center of the funnel, inducing asymmetry in the plot.

Figure 18.1 depicts a relatively symmetric funnel plot produced in the traditional manner, using simulated data. However, that mode of presentation goes against the convention of plotting an unknown quantity on the y-axis and a fixed quantity (such as *N*) on the x-axis. Figure 18.2 shows the same plot with the more standard graphics conventions. Funnels in both orientations exist in the literature. The true effect in both plots is 0.5.

It is interesting that study suppression caused by study size, effect size, or statistical significance (one-sided), either individually or in combination, could produce an asymmetric funnel plot. It is also possible for two-sided statistical significance suppression mechanisms (that is, significant studies in either direction are more likely to be published) to could create a tunnel or hole in the middle of the funnel, particularly when the underlying effect size is close to zero. However, in most circumstances, it is implausible that a selection mechanism based on two-tailed *p*-values would function the same way in both tails.

The most appropriate axes for the funnel plot is debated, particularly with respect to the measure of study precision (Vevea and Hedges 1995; Sterne and Egger 2001). Variance (and its inverse) and standard error (and its inverse) are options for use in place of sample size. This choice can affect the appearance of the plot considerably. For instance, if the variance or standard error is used, the

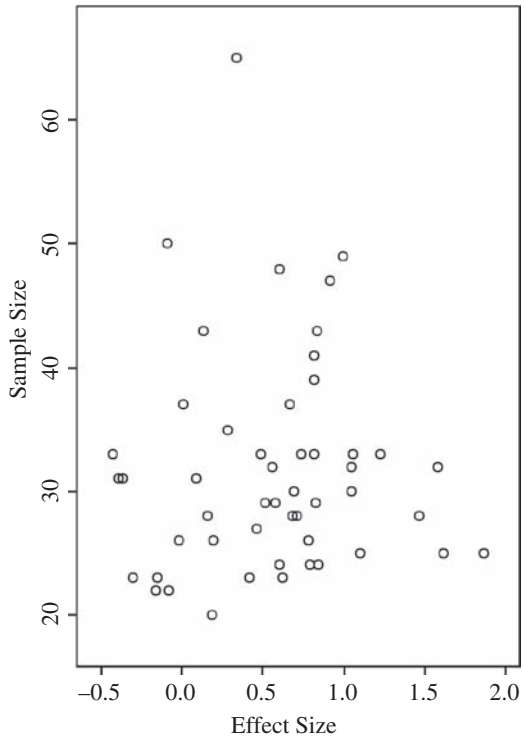

Figure 18.1 Traditional Funnel Plot, Unbiased Data
SOURCE: Author's tabulation.

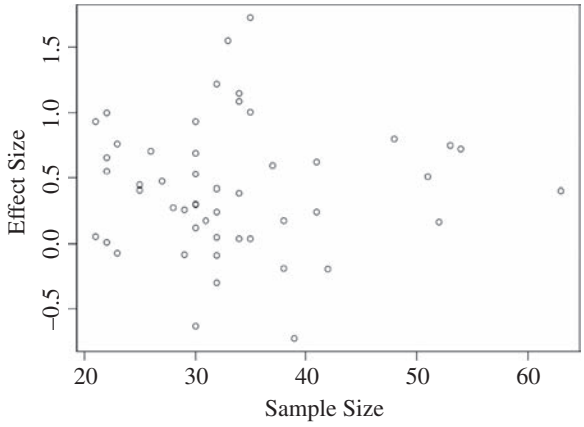

Figure 18.2 Horizontal Funnel Plot, Unbiased Data
SOURCE: Author's tabulation.

distribution of effect sizes covers an expanded range for smaller studies. This gives more plot space to the smaller studies, among which publication bias is more likely to be evident. Jonathan Sterne and Matthias Egger have published comparative plots (2001). Figure 18.3 plots the same effects as figure 18.2, this time against standard error rather than sample size. In figure 18.3, the larger studies appear at the left of the plot, rather than the right, as in figure 18.2, and the range of the plot associated with smaller sample sizes (and larger standard errors) is expanded. Figure 18.4 shows a highly asymmetrical funnel plot, using standard error on the x-axis.

When interpreting funnel plots, the meta-analyst should bear in mind that asymmetry may be due to phenomena other than publication bias. Any external influence associated with both study size and effect size could confound the observed relationship. For example, small studies could be conducted under more carefully controlled experimental conditions than large studies, resulting in differences in effect sizes. In other situations, a higher intensity of the intervention might be possible for the smaller studies, causing their true effect sizes to be larger. Conversely, smaller studies might be carried out under less rigorous conditions; for example, consider a meta-analysis that mixes results from large clinical trials with smaller observational studies that tend to have larger effects. Figure 18.4 actually depicts such a situation; the asymmetry is difficult to miss. Figure 18.5 portrays the same data, but adds information about study type. Neither type of study appears asymmetric, even though the combined distribution is.

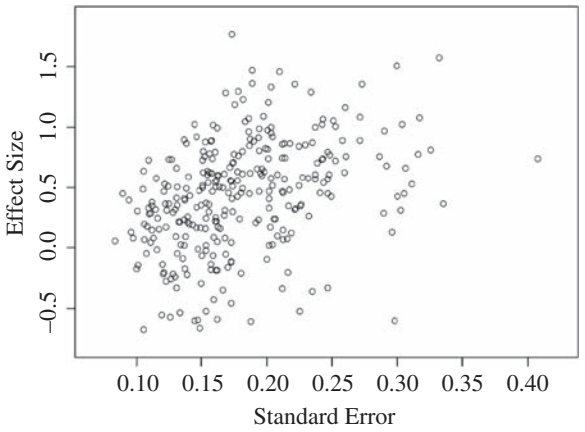

Figure 18.4 Effect Size Against Standard Error, Strong Asymmetry

SOURCE: Author's tabulation.

The utility of the funnel plot has been questioned because of the subjective nature of its interpretation. Norma Terrin, Christopher Schmid, and Joseph Lau find that researchers faced with an assortment of funnel plots cannot correctly identify which plots show bias (2005). Joseph Lau and his colleagues present similar evidence of inconsistent inter-

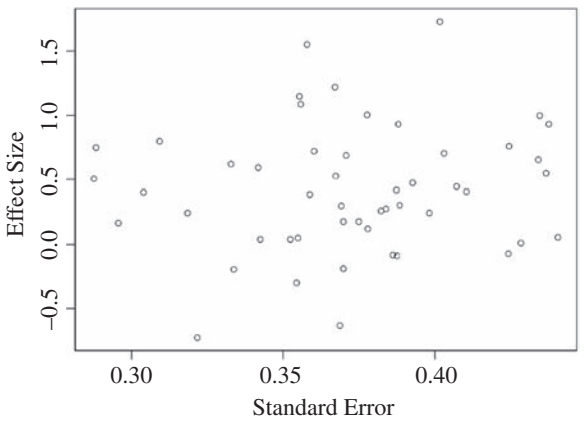

Figure 18.3 Effect Size Against Standard Error, Unbiased Data

SOURCE: Author's tabulation.

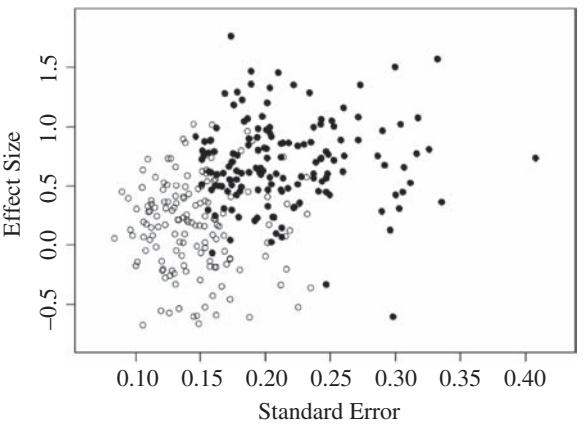

Figure 18.5 Asymmetry Due to Moderator

SOURCE: Author's tabulation.

pretation (2006). Jin-Ling Tang and Joseph Liu (2000), as well as James Hunter and his colleagues (2014), describe problems with interpretation in circumstances where the magnitude of effect sizes is associated with a measure of precision. This is particularly problematic for some outcome measures, such as odds ratios, for which the estimated effect size and its standard error are positively correlated. As a result, in such cases the funnel plot will not be symmetric even in the absence of publication bias. However, Jaime Peters and his colleagues find that the induced asymmetry is small for small and moderate effect sizes (2006).

If variance, standard error, or their inverses are used, it is possible to construct expected 95 percent confidence intervals around the pooled estimate that form guidelines for the expected shape of the funnel under a fixed-effect assumption. Such contour-enhanced funnel plots can also represent other confidence levels. Sterne et al. argue that they can aid in interpretation of the plot (2011). Peters and his colleagues propose that contour-enhanced plots can help the analyst distinguish between asymmetry due to publication bias and asymmetry caused by other factors (2008). Contour guidelines are approximate because they are constructed around the pooled meta-analytic estimate, which may itself be biased. Sometimes the main indication of asymmetry may be differences in the density of plotted points. This frequently occurs well within the bounds of the contour lines. In that case, the contour lines can deflect attention from such changes in density.

18.3.2 Cumulative Meta-Analysis

Cumulative meta-analyses are often used to determine when a meta-analytic effect size appears to stabilize in relation to some variable of interest, such as time of publication. Such a plot consists of a forest plot of meta-analytic results, starting with the oldest study, followed by a meta-analysis of the two oldest studies, and so on, until the final line in the forest plot represents the meta-analysis that includes all of the effect-size estimates. These plots can help identify circumstances in which large effects or results with very small *p*-values are published more quickly than others, a form of publication bias known as time-lag bias.

Recently, the approach has gained a role in the assessment of other forms of publication bias as well. Kepes, Bennett, and McDaniel demonstrate the utility of cumulative meta-analysis (2014). The analyst adds effect sizes one at a time in increasing (or decreasing) order of precision and creates a forest plot of this cumulative meta-analysis. If horizontal drift is present in the plot as studies are added, there is evidence of a relationship between study size and effect size, and therefore the possibility of publication bias. Note, however, that this method, like the funnel plot, cannot distinguish between associations due to publication bias and those due to other causes. A similar approach that bases the meta-analytic estimate on a subset of the most precise studies has also been proposed. Researchers differ on the number or percentage of studies to include for this method. Kepes, Brad Bushman, and Craig Anderson recently used the method using the five most precise studies (2017). In contrast, Tom Stanley, Stephen Jarrell, and Hristos Doucouliagos used the most precise 10 percent of the effect sizes (2010).

18.3.3 Nonparametric Correlation Test

Colin Begg and Madhuchhanda Mazumdar's rank correlation approach makes available a formal test for the presence of funnel plot asymmetry (1994). Along with the other methods presented in this section, it assesses whether a relationship between study size and effect size is present, but does not provide an adjusted estimate of effect size.

The rank correlation test works by estimating a fixed-effect meta-analytic mean, calculating deviations of individual effects from that mean, and standardizing them using the standard error of the deviations from the mean (accounting for both the sampling uncertainty of the individual effects and the standard error of the fixed-effect mean). Subsequently, one calculates a rank correlation between the deviations and their variances using a normalized version of Kendall's tau. Under the null hypothesis of no association between effect and variance, the statistic is tested by reference to the standard normal distribution. As with any hypothesis test, nonsignificance should not be interpreted as a confirmation of the null hypothesis (in this case, that publication bias is absent). Any effect-size scale can be used as long as it is distributed asymptotically normal. The test is available in many statistical packages for meta-analysis.

Begg and Mazumdar's original publication of the method acknowledged that it was underpowered for small meta-analyses (1994). Others echo that finding (Sterne et al. 2000). From comparisons between this and the linear regression test described later in this chapter (Sterne, Gavaghan, and Egger 2000), it appears that the linear regression test is more powerful, though results of

the two tests can sometimes be discrepant. Modifications of the rank correlation test have been proposed to address the power issue, with varying success (for a discussion of these extensions, see Kepes et al. 2012). Because of this concern about power, Jonathan Sterne and his colleagues propose interpreting the results of this test only when a data set includes more than ten effects (2011).

18.4 METHODS FOR ASSESSING THE IMPACT OF PUBLICATION BIAS

Here we examine a variety of approaches for assessing the impact of publication bias on a meta-analysis. Some of these provide an adjusted estimate, a formal test, or both.

For those methods that produce adjusted estimates, meta-analysts likely wish to label the degree of adjustment, or the amount of bias present in their data. Several guidelines for doing so are proposed. Hannah Rothstein, Alexander Sutton, and Michael Borenstein refer to bias as "minimal" when the estimates of effect size are very similar, "modest" when the difference is substantial but the key finding does not change, and "severe" when the key finding is called into question (2005). Kepes, George Banks, and In-Sue Oh refine these definitions, classifying bias as "absent/negligible" when the difference between the unadjusted and adjusted estimates is less than 20 percent, "moderate" when the difference is between 20 percent and 40 percent, and "severe" when the difference is greater than 40 percent (2014). Toward the end of this chapter, we use both guidelines. However, assessing the degree of adjustment is fundamentally subjective, and these guidelines should not be viewed as ironclad.

18.4.1 Fail-Safe *N*

Robert Rosenthal introduced the method that has come to be called the *fail-safe N* (1979). It remains one of the most popular techniques for assessing publication bias today, particularly in the social sciences. The fail-safe *N* addresses the question of how many effect sizes averaging a null value would need to be missing from a meta-analysis to overturn the conclusion that there is a significant effect. Here, significance is defined in terms of inference based on combined *p*-values using the *Z*–score approach (see Stouffer et al. 1949).

Although the fail-safe *N* may be intuitively appealing, it is now generally regarded as valueless. Begg and Berlin argue that the method should be considered nothing more than a crude guide due to a number of shortcomings (1988). Betsy Becker notes that no statistical model underlies the fail-safe *N*, and there is no clear-cut and justifiable criterion for a "large" fail-safe *N* value; she specifically states that the method should be abandoned (2005). One concern is that combining *Z*-scores does not directly account for the sample sizes of the studies. Another is that the choice of zero for the average effect of the unpublished studies is arbitrary; a glance at a typical asymmetric funnel plot will suggest that it is effects below zero that are missing. Hence, in practice, many fewer studies than the number suggested by the fail-safe *N* might be required to overturn the meta-analytic result. Often, then, it has led to unjustified complacency about publication bias. Still another shortcoming is that the method does not adjust or deal with treatment effects—just *p*-values (and thus it provides no indication of the "true" effect size). Satish Iyengar and Joel Greenhouse point out that heterogeneity among the studies is ignored (1988). Robert Orwin notes that the shape of the funnel plot does not influence the method. For these reasons, it is difficult to recommend using the procedure.

It could be argued that, in addition to being hampered by those issues, the fail-safe *N* poses fundamentally the wrong question. Typically, except for meta-analyses that include very few effects, the power to detect even small combined effects is tremendous. Thus it is much more interesting to focus on the question of how many studies would have to be missing for the combined effect to be reduced to a trivial magnitude. Orwin presents an alternative fail-safe *N* that addresses that more interesting question (1983). His method allows the user to specify an average value for missing effects that may or may not be zero, and estimates the number of such effects that would need to be added to the analysis to move the estimated effect below the specified value. This variation is still used with some frequency. However, a glance at the literature that uses the method shows that many who use it accept the default value of zero for the average of the missing effects.

In short, with the possible exception of Orwin's variant, the fail-safe *N* is not a valid method for assessing publication bias.

18.4.2 Methods Based on Observed *p*-Values

The three methods in this category are based not on effect size or study size but on the *p*-values associated with the effects. They have recently gained popularity among meta-analysts, although Blakeley McShane, Ulf Böckenholt,

and Karsten Hansen point out that two of these methods may be viewed as alternative implementations of existing (and more effective) selection models, and all three methods have restrictive assumptions and a series of documented flaws (2016; van Aert, Wicherts, and van Assen 2016; Bruns and Ioannidis 2016; Bishop and Thompson 2016; Ulrich and Miller 2015).

18.4.2.1 *p*-Curve and *p*-Uniform *P*-curve is a method for assessing publication bias, first published by Uri Simonsohn, Lief Nelson, and Joseph Simmons, that has gained popularity (2014). It is based on the notion that, if a given set of studies has evidential value (the average effect size represents a real effect and is not an artifact of bias), the distribution of those one-tailed *p*-values will be right skewed. This means that very small *p*-values (such as $p < .025$) will be more numerous than larger *p*-values if studies have evidential value. If the distribution of significant *p*-values is left skewed and large *p*-values are more numerous than expected, Simonsohn and his colleagues conclude that it is evidence of *p*-hacking—researchers may be striving to obtain *p*-values that fall just below .05.

P-curve uses two tests to assess whether the distribution is right skewed. The first is a binomial test comparing the proportion of observed *p*-values above and below .025; the second is a continuous test that calculates the probability of observing each individual *p*-value under the null hypothesis. The probabilities produced by this second test are then dubbed the studies' "pp" values. These tests for right skew assess what is called the full *p*-curve. To test for "ambitious *p*-hacking," or *p*-hacking to reach below .025 rather than .05, *p*-curve conducts the same tests for right skew on only the observed *p*-values that are below .025, or the "half *p*-curve." If these tests for right skew are not significant, indicating that the studies lack evidential value and no true effect may be present, *p*-curve conducts another pair of binomial and continuous tests to assess whether the studies were underpowered (defined as having power below 33 percent). Simonsohn, Nelson, and Simmons mention that an adjusted effect size can be calculated, and provide supplementary *R* code (R Core Team 2016) on their website (www.p-curve.com) for doing so, but the code is somewhat complicated, and adjusted effect sizes are not discussed here (2014). This method of obtaining an adjusted effect size based on the *p*-curve is an aspect of the *p*-curve approach that functions similarly to *p*-uniform.

P-uniform is also a new method, first published by Marcel van Assen, Robbie van Aert, and Jelte Wicherts (2015). It assumes that the population effect size is fixed—that is, the observed effect sizes are homogeneous with neither systematic nor random heterogeneity. It also assumes that all studies with significant results are equally likely to be published or available for inclusion in a literature review, and that no significant studies are withheld. These are both restrictive assumptions. Most meta-analytic data are heterogeneous to some degree (Gelman 2015; McShane and Böckenholt 2014; McShane and Gal 2015), often despite researchers' best attempts to maintain homogeneity (Klein et al. 2014). In addition, it is extremely unlikely that every significant test has been published and that every nonsignificant test remains unpublished.

Like *p*-curve, *p*-uniform is based on the idea that *p*-values, conditional on a true effect size, are uniformly distributed. If, for example, the hypothesized population effect is 0.50, and the conditional *p*-values for each study are not uniform when calculated under the null hypothesis that the true effect is 0.50, both methods assume that the studies do not reflect the true underlying effect, or that publication bias is present.

P-uniform performs two tests. The first assesses the null hypothesis that the population effect size is zero by transforming the observed significant *p*-values using Ronald Fisher's (1932) method and assessing whether their conditional distribution is uniform. If it is, the test fails to reject the null hypothesis and concludes that there is no evidence of an effect. The second test is a one-tailed test of whether the population effect size equals the effect-size estimate produced by a traditional fixed-effect meta-analysis (van Assen, van Aert, and Wicherts 2015). Again, significant *p*-values are transformed, this time to represent the probability of observing a given effect size conditional on both the fixed-effect average estimate and statistical significance. If this distribution deviates from a uniform distribution, *p*-uniform rejects the null hypothesis and concludes that publication bias may be a threat. Finally, *p*-uniform provides an adjusted effect-size estimate and confidence interval by searching for the population effect size that does meet its qualification—the value where the distribution of conditional *p*-values is uniform. This is similar to the method *p*-curve employs, but the two methods use different algorithms for defining fit to the uniform distribution.

Both *p*-curve and *p*-uniform have several flaws. Robbie van Aert, Jelte Wicherts, and Marcel van Assen point out that these tests perform poorly when meta-analytic data contain *p*-values close to significance levels like .05 (2015). Obviously, many meta-analyses likely con-

tain *p*-values in this range. These methods will always underestimate the true effect when *p*-hacking is present (although, to be fair, many of the other methods in this chapter will as well), and neither method can perform well with heterogeneous data (van Aert, Wicherts, and van Assen 2016). This leads to a recommendation that meta-analysts whose data are heterogeneous should divide their data into homogeneous subgroups prior to estimating the models, but this is often impractical. A traditional random-effects model actually outperforms *p*-curve and *p*-uniform with heterogeneous data, even in the presence of publication bias, the very situation in which both models are designed to work (van Aert, Wicherts, and van Assen 2016). The models assume that all significant studies are published (or otherwise widely available), and are calculated involving only significant *p*-values (McShane, Böckenholt, and Hansen 2016). Finally, Stephan Bruns and John Ioannidis discovered that *p*-curve has difficulty distinguishing between *p*-hacking and the presence of a true effect—the primary purpose for which the method exists (2016). Dorothy Bishop and Paul Thompson confirm their findings (2016), as do Rolf Ulrich and Jeff Miller (2015).

McShane, Böckenholt, and Hansen (2016) note that *p*-curve and *p*-uniform are both a modification of an early selection model presented by Hedges (1984), which forgo maximum likelihood estimation in favor of less efficient alternatives. The benefit of these models is their recent publicity and accessibility, which may increase awareness of publication bias. Beyond that, however, simulations demonstrate that earlier selection models remain more effective under realistic assumptions (McShane, Böckenholt, and Hansen 2016). McShane et al. (2016) also provide mathematical evidence of their ineffectiveness; they argue, based on Jensen's Inequality (Jensen 1906), that *p*-curve and *p*-uniform (as well as the early Hedges 1984 model) will be biased in the presence of heterogeneity.

These models appear to reinvent a wheel first discovered over thirty years ago. The *p*-curve and *p*-uniform methods are certainly superior to some, like the fail-safe *N* or the excess significance test (see following section), and *p*-curve in particular has gained popularity, likely due to its accessibility. Any assessment of publication bias is better than none, and using all methods available is better than using only one. However, meta-analysts should remember that *p*-curve and *p*-uniform are modified versions of simplistic early weight-function models, and should consider using more sophisticated weight-function models as well.

18.4.2.2 Excess Significance Test The excess significance test (or TES), first proposed by John Ioannidis and Thomas Trikalinos (2007), has been the subject of considerable debate. The method is a null hypothesis significance test that takes a given set of studies and asks whether too many are statistically significant or "positive." For example, if a meta-analysis collected three studies and all three were significant with $p < .05$, the excess significance test instructs the meta-analyst to calculate the post hoc power of each study (assuming that the estimated effects are the true effects). The expected number of positive studies is calculated based on the studies' power, and that expected number is compared with the observed number of positive studies using the chi-square statistic. Assuming that each study had 60 percent power, the probability that all three studies would reject the null hypothesis with $p < .05$ works out to the product of the power values, 0.60^3, or 0.22. Guidelines for the TES indicate that a *p*-value less than .10 should be considered significant (Francis 2014). Therefore, given that $0.22 > 0.10$, the meta-analyst will fail to reject the null hypothesis and can conclude that the observed number of positive studies does not exceed the expected.

The simplicity of the excess significance test may initially be appealing. In 2012, Gregory Francis published a series of papers employing the test in various subfields to argue for the presence of publication bias (2012a, b, c, d, e, f, g) and concluded that the results from those subfields should be ignored (Simonsohn 2013). By late 2012, the test of excess significance was attracting more attention, primarily criticism (Balcetis and Dunning 2012; Galak and Meyvis 2012; Piff et al. 2012; Simonsohn 2012), and the test became the focus of a special issue in the *Journal of Mathematical Psychology*.

Criticism of Francis's work and of the excess significance test itself is rooted in a number of important issues. First, as Simonsohn (2013) pointed out, even the presence of publication bias should not result in the rejection of a field of research. The excess significance test does not assess the evidential value (or practical significance) of results. The method is a null hypothesis significance test, implying that it is a form of confirmatory research, despite the fact that it is exploratory at best. Perhaps most notably, Simonsohn writes that the excess significance test actually answers this question: "Has a large enough set of published studies been compiled to reject the obviously false null that all studies, regardless of outcome, would be reported?" (175).

Other criticism of the excess significance test includes the fact that no guidelines are in place for choosing which

tests to examine from a study, especially considering that studies often base multiple tests on the same data, which results in dependencies that affect the outcome of the excess significance test (Johnson 2013). The test does not provide any idea of the magnitude of publication bias or its implications; it also makes assumptions that violate the sequential nature of the publication process (Morey 2013). The excess significance test itself suffers from a substantial lack of power, and simulations demonstrate that it cannot detect even extreme bias without prior knowledge of the true population effect size (Vandekerckhove, Guan, and Styrcula 2013). The test is also likely to perform poorly when effect sizes are heterogeneous (Ioannidis and Trikalinos 2007). Finally, Kepes and Michael McDaniel's simulations revealed that the test for excess significance is not robust to outliers (2015).

In comparison with the many other methods described in this chapter, which have fewer flaws and more redeeming qualities, the excess significance test falls short. Much like the fail-safe *N*, the excess significance test is not useful as a test for publication bias. Researchers should not popularize the test because of its simplicity. Researchers would also be wise to refrain from condemning entire fields of study on the basis of a single severely flawed test.

18.4.3 Trim and Fill

The nonparametric trim and fill method was developed as a simpler alternative to parametric selection models (Duval and Tweedie 2000a, 2000b). It is one of the most popular methods for adjusting for publication bias (Borenstein 2005; Moreno et al. 2009). Also, because the method formalizes the use of a funnel plot, the ideas underlying its statistical calculations can be communicated visually, increasing its accessibility.

The method is based on rectifying funnel plot asymmetry. It assumes that the studies on either the far left-hand or right-hand side of the funnel are suppressed and, therefore, it is a one-sided procedure. First, the method uses an iterative process to determine how many studies would have to be removed, or "trimmed," from one side of the funnel for the remaining effect sizes to be symmetric. It trims the asymmetric effect sizes, then uses one of three estimators to generate, or "fill in," new effects that are mirror images of the remaining ones. The adjusted pooled effect size is then calculated based on this augmented symmetrical data set, which can also be used to calculate an adjusted variance component (Jennions and Moller 2002). Although Eric Weinhandl and Sue Duval (2012) are currently working on allowing trim and fill to include a linear model for the mean effect, it is not yet developed for more than one linear predictor.

Either a fixed- or random-effects meta-analytic model can be used for the iterative trimming and filling parts of this method, and once effect sizes are filled in, either model can be used to obtain adjusted estimates. In this way, the method can accommodate random (or between-studies) heterogeneity, and it can produce an adjusted variance component estimate. The choice between fixed- and random-effects is important. Sue Duval and Richard Tweedie originally advised using a random-effects model for both steps, a process that they referred to as "random-random," because doing so would yield more conservative confidence intervals (2000a, 2000b). However, if publication bias is present and smaller studies are clustered together, the random-effects model (which allows more weight to smaller studies) may be biased; as a result, Alexander Sutton advocates a "fixed-fixed" process (2005). The "fixed-random" process is a compromise, using a fixed-effect model for trimming and a random-effects model to estimate the adjusted effect, although the adjusted estimate may be overly conservative (Peters et al. 2007). Duval recommends the more conservative approaches, although she emphasizes that all three should be estimated and compared (2005).

During the iterative procedure, three possible estimators of the number of missing studies may be employed. Two of these estimators, known as R_0 and L_0, are recommended; the third, Q_0, is merely a linear transformation of one of the others (Duval and Tweedie 2000a, 2000b). If choosing between the two (R_0 and L_0), L_0 sometimes performs better (Jennions and Moller 2002). Others have found that R_0 performs better (Peters et al. 2007). Duval, however, recommends estimating both and comparing the results, especially because the performance of the estimators can depend on the number of observed versus missing effects (Duval 2005; Duval and Tweedie 2000b). Given space constraints, the details of these estimators are not provided here, but thorough examples of their calculation are available elsewhere (Duval and Tweedie 1998; Duval 2005).

Duval and Tweedie initially evaluated this method through simulation, under homogeneous conditions and with a data suppression mechanism matching the model—that is, where the most extreme effect sizes were suppressed (2000a, 2000b). Under those conditions, it performed well. However, other simulations suggest it may perform poorly in the presence of between-study heterogeneity in the absence of any evidence of publication bias (Terrin et al. 2003). Peters and his colleagues conducted

simulations further evaluating the performance of trim and fill, finding that trim and fill underestimates the true effect size in the absence of publication bias (2007). They note that trim and fill is not ideal, in part because it can impute unrealistic effect sizes, although it *can* outperform the unadjusted random-effects model in the presence of publication bias, but should be considered a sensitivity analysis, as originally intended (Peters et al. 2007; Duval and Tweedie 2000a, 2000b). Guido Schwarzer, James Carpenter, and Gerta Rücker (2010) compare trim and fill to the Copas selection model (Copas and Shi 2000) and confirm that trim and fill is more conservative due to inflated standard errors.

All publication bias methods should be regarded as sensitivity analyses; therefore, saying the same of trim and fill is not a slight. Although trim and fill has its share of problems, it is both popular and accessible, and meta-analysts will likely benefit from including it in their arsenal of assessment methods.

18.4.4 Linear Regression Adjustment

In 1997, Matthias Egger and his colleagues described a parametric test for funnel plot asymmetry based on linear regression. The test regresses the standard normal deviate, or the effect sizes divided by their standard errors, on precision (defined as the inverse of the standard error). This regression fits a line to Rex Galbraith's (1994) radial plot, in which the regression line is not constrained to go through the origin. Effect sizes from small studies will have a standard normal deviate that is close to zero regardless of their magnitude, and large studies will produce large standard normal deviates. Therefore, in the absence of publication bias, the regression line will run through the origin. If bias is present, small studies may differ systematically from larger studies, and the line will no longer run through the origin (Egger et al. 1997).

The regression intercept measures the magnitude and direction of asymmetry, and a significant *t*-test on the intercept indicates that asymmetry (and, by extension, publication bias) may be present. A negative intercept indicates that smaller studies have larger effects; a positive intercept indicates that they have smaller effects than expected. The regression model, as proposed above, is equivalent to a weighted meta-regression model, and the regression line can be displayed on a funnel plot for clarity of interpretation.

The model just presented is Egger's linear regression in its original form. Several researchers have proposed extensions or modifications of this model (Macaskill, Walter, and Irwig 2001; Sterne and Egger 2005; Harbord, Egger, and Sterne 2006; Peters et al. 2006; Rücker, Schwarzer, and Carpenter 2008; Deeks, Macaskill, and Irwig 2005). Petra Macaskill, Stephen Walter, and Lesley Irwig propose a variation in which the effect sizes, rather than their standard normal deviates, are regressed on their study size and weighted by their inverse pooled variance (2001). This model and the next three variations reverse the role of the intercept and slope; the slope is expected to be zero in the absence of publication bias. Peters and his colleagues prefer regressing effect sizes on the inverse of their sample size (2006). Sterne and Egger advocate the regression of effect sizes on their standard errors, weighted by their inverse variance (2005). Gerta Rücker, Guido Schwarzer, and James Carpenter describe a variation of this for binary outcome data that has been arcsine-transformed (2008). Jonathan Deeks, Petra Macaskill, and Les Irwig propose a regression of the effect size involving the effective sample size (ESS), defined as $4n_1n_2/(n_1+n_2)$ (2005). The effect size is regressed on the reciprocal of the square root of ESS and weighted by ESS. Finally, Roger Harbord, Egger, Jonathan Sterne recommend regressing the efficient scores (defined as the first derivative of the log-likelihood) against the score variance (Fisher information), for which the intercept is a measure of bias (2006). For binary outcomes, because of the correlation between odds ratios and their standard errors, the original Egger's regression has an inflated type I error rate, and variations are preferable (Moreno et al. 2009).

Clearly, there are several variations of the original Egger's linear regression (Egger et al. 1997). Although these models differ in terms of outcome measure and predictor, and although the role of the intercept and slope occasionally change, the models are not discussed individually for the sake of brevity. A mention of Egger's linear regression or Egger's test here refers to the entire class of models, unless otherwise specified.

If a meta-analytic data set contains systematic heterogeneity due to covariates, these must be considered when using any funnel plot–based assessment. With discrete covariates, separate assessments can be made for each group, although this approach may result in a considerable reduction of power. The meta-analyst can also extend the regression model to include study-level covariates, therefore estimating a mixed-effects weighted regression. In this way, Egger's regression is capable of accommodating some forms of heterogeneity, but it does not incor-

porate random (or between-studies) heterogeneity and cannot estimate a variance component. Egger's regression does, however, produce an adjusted estimate of the average effect size (the slope), although its estimate is biased because its predictor variable is subject to sampling error and therefore violates the assumptions of linear regression (Macaskill, Walter, and Irwig 2001). Although we are not aware of any research on the subject, it is theoretically possible to fit a measurement error model to overcome this bias; exploring such an idea could be promising.

Peters and his colleagues note that Egger's regression is widely used in the medical literature (2006). In the social science literature, the fail-safe *N* is still the most common procedure, despite its deep-seated flaws, but Egger's regression is gaining popularity (Ferguson and Brannick 2012). Egger's regression does suffer from low power and poor performance when the number of studies is small, especially when there are fewer than twenty, or when the treatment effect is large (Moreno et al. 2009; Sterne, Egger, and Smith 2001; Macaskill, Walter, and Irwig 2001). Egger's regression is most powerful with a large number of effect sizes that range widely in terms of study size (Macaskill, Walter, and Irwig 2001). Its problems with power, however, are not unique among publication bias assessment methods, and it is still a useful tool.

18.4.5 PET-PEESE

In 2014, Tom Stanley and Hristos Doucouliagos proposed PET-PEESE, and since then the method has appeared occasionally in the meta-analytic literature (Carter and McCullough 2014; Carter et al. 2015). PET-PEESE is actually an extended modification of Egger's regression (Egger et al. 1997). Stanley and Doucouliagos instruct the meta-analyst first to estimate a regression of effect size on standard error, weighted by the inverse variance (2014). This is the exact model that Sterne and Egger propose (2005). Stanley and Doucouliagos point out that, though the slope of this model is a measure of bias, the intercept is also informative: it represents an estimate of the effect size when the standard error is zero (2014). Therefore, they argue that the intercept is an estimate of a perfectly precise study, or an effect size uninfluenced by publication bias (Stanley 2005). They call this first regression the Precision-Effect Test (or PET). Thus far, PET is a restatement of the fact that, for certain variations of Egger's regression, the intercept is an effect-size estimate adjusted for publication bias. A *t*-test on the intercept, using the null hypothesis that the intercept is zero, indicates whether a true effect is present.

Assuming that a test on the intercept is significant, or that a nonzero effect exists, results in a second problem. The issue with using this adjusted estimate is that, as mentioned previously, the estimate is biased. To avoid this problem, Stanley and Doucouliagos propose a second conditional test (2014). If the intercept from PET is significant, they advise meta-analysts to conduct another regression, this time with effect size predicted by sampling variance rather than standard error. This regression is called the Precision-Effect Estimate with Standard Error (PEESE). PEESE produces an intercept that is *still* biased, but simulations demonstrate that it is less biased than the intercept from PET (Stanley and Doucouliagos 2014). Therefore, they advise that, if the PET test is significant, meta-analysts should estimate PEESE and accept its intercept as an adjusted effect-size estimate. If PET is nonsignificant, there is not enough evidence that the true effect size differs from zero.

A problem with this approach is that bias in the intercept estimate does not vanish when using variance as a predictor rather than its square root. PET-PEESE is not a new technique, although it is described as such; it is a combination of existing variations of Egger's regression. Macaskill, Walter, and Irwig have explained the source of bias in the intercept (2001). The intercept is a biased estimate not because of the choice of predictor, but because of a violation of one of the assumptions for linear regression. Both predictors, variance and standard error, are not fixed; they are random, and are estimated from the observed data. Therefore, using either PET or PEESE, measurement error is inherently present in the independent variable, and the estimate of the intercept will be biased downward—the exact result that Stanley and Doucouliagos (2014) describe. Stanley and Doucouliagos also argue that PET-PEESE outperforms random-effects meta-regression in the presence of publication bias, although it still performs worse overall in the presence of high levels of heterogeneity.

Because it is a combination of two Egger's regression variations, PET-PEESE possesses the same flaws. It has low power and cannot incorporate random or between-studies heterogeneity. When heterogeneity is present, it performs poorly (Stanley and Doucouliagos 2014) and the coverage rate of its confidence intervals is persistently low (Moreno et al. 2009); in the presence of severe bias or when the data are homogeneous, its confidence intervals are too wide (Moreno et al. 2009). Finally, its

effect-size estimate is biased. When we consider the many other methods presented in this chapter that possess more redeeming features, including other variations of Egger's regression, PET-PEESE appears to be a flawed method.

18.4.6 Selection Modeling

Selection models adjust meta-analytic data sets by specifying a model that describes the mechanism by which effect sizes may be suppressed. This model is combined with an effect-size model that describes the distribution of effect sizes in the absence of publication bias.

If the selection model were known, selection methods would be straightforward, but the precise nature of the suppression will almost always be unknown. Instead, selection approaches attempt to estimate the selection model, along with adjusted estimates of the meta-analytic parameters. Although complex to implement, they are recommended over other methods, which can produce misleading results when effect sizes are heterogeneous (Terrin et al. 2003). Selection methods may perform poorly when the number of observed effects is small; an alternative involves specifying selection models of varying severity and estimating the meta-analytic parameters contingent on each hypothetical selection pattern. Jack Vevea and Carol Woods present such an approach (2005; see also table 18.1). These methods, which do not estimate parameter values from the data, are sensitivity analyses by nature, although, of course, all bias assessments are.

Two classes of selection models have been developed: those that model suppression as a function of an effect size's *p*-value, and those that model suppression as a function of a study's effect size and standard error simultaneously. Both are implemented using weighted distributions that represent the likelihood of observing a given effect estimate if it occurs. These methods have gained popularity in the publication bias literature. Descriptions of the more complex selection models are presented here with limited statistical detail. Hedges and Vevea published a comprehensive review of selection models available by the early 2000s that provides a more statistically rigorous account of some of the approaches described here (2005).

Although they do have flaws, namely, their complexity and sample size requirements, both classes of selection model tend to perform well in simulations and allow meta-analysts to evaluate data under a range of selection patterns. Therefore, they are valuable tools in an arsenal of bias assessments.

18.4.6.1 Suppression as a Function of *p*-Value Only Selection models that depend solely on effect sizes'

Table 18.1 Sample Selection Patterns for the Vevea and Woods Method

	Probability of Observing Effect			
p Interval	Moderate One-Tailed Selection	Severe One-Tailed Selection	Moderate Two-Tailed Selection	Severe Two-Tailed Selection
.000–.005	1.00	1.00	1.00	1.00
.005–.010	.99	.99	.99	.99
.010–.050	.95	.90	.95	.90
.050–.100	.90	.75	.90	.75
.100–.250	.80	.60	.80	.60
.250–.350	.75	.50	.75	.50
.350–.500	.65	.40	.60	.25
.500–.650	.60	.35	.60	.25
.650–.750	.55	.30	.75	.50
.750–.900	.50	.25	.80	.60
.900–.950	.50	.10	.90	.75
.950–.990	.50	.10	.95	.90
.990–.995	.50	.10	.99	.99
.995–1.000	.50	.10	1.00	1.00

SOURCE: Author's tabulation.

p-values propose or estimate the likelihood of surviving selection as a function of those *p*-values. Hedges (1984) as well as David Lane and William Dunlap (1978) propose simple selection models that assume all statistically significant effect sizes are observed (for example, $p < .05$ two-tailed, or $p > .975$ or $p < .025$ one-tailed) and all others are suppressed. With this approach, any effect size with a *p*-value < .05 has a probability of one (certainty) of being observed and a probability of zero otherwise. Iyengar and Greenhouse propose somewhat more sophisticated models, assuming that the likelihood of publication is a decreasing function of the *p*-value for studies that are not statistically significant (1988). In the years following, various authors have proposed more sophisticated models.

18.4.6.1.1 Dear and Begg. Keith Dear and Colin Begg introduce a semi-parametric method for assessing publication bias that uses a nonparametric weight function on the two-tailed *p*-value scale (1992). The method, they note, can easily be adapted for one-tailed *p*-values. The weight function is a step function with discontinuities at the alternate individual observed values of *p*. In other words, the Dear and Begg model takes the observed *p*-values of all effect sizes in the data set and orders them. It then includes every other *p*-value as discontinuities in the weight function. For example, if the first four *p*-values of a data set were .001, .01, .03, and .04, the first discontinuity in the weight function would be set at $p = .01$, and weights would be estimated for *p*-values below .01 and *p*-values between .01 and .04. This means there are $k/2$ weight parameters for a meta-analytic data set of size k. The model estimates a weight for each interval that represents the relative probability of surviving the selection process. To identify the model, the weights are constrained to fall between zero and one. However, they are not directly interpretable as probabilities because we lack information about the base rate of publication. No effect has 100 percent probability of publication.

Although the model can provide both an adjusted estimate of the average effect size and a statistical test, Dear and Begg focus on using plots of the weights against *p*-values as a tool for visual assessment (1992). Spikes in the plot indicate that the weight for *p*-values in that particular range is large, meaning that studies with *p*-values in that range are more likely to be published and therefore observed. Valleys or dips in the plot indicate the opposite; studies with *p*-values in those ranges are less likely to be observed.

With this approach, weights for larger (less significant) *p*-values sometimes exceed the weights for the most significant values, making visual assessment of bias difficult. If slight fluctuations in the weights are numerous, identifying the overall pattern may be complicated. Kaspar Rufibach presents an extension of the model that addresses this problem (2011). His approach is identical to Dear and Begg's (1992), except that Rufibach has imposed a constraint, forcing the weights to be a monotone nonincreasing function of *p*-values. Rufibach notes that the constraint improves the performance of estimates, yields more insight into the selection process, and leads to a more realistic weight function. The constraint also makes it easier for meta-analysts to interpret the function from plots.

The Rufibach model provides a useful plot of the weight function, and can be informative (2011). However, as the number of effect sizes in the meta-analysis increases, both the Dear and Begg (1992) and Rufibach models become difficult, if not impossible, to estimate. This problem occurs because, rather than allowing meta-analysts to restrict the number of *p*-value discontinuities, the models determine the number of discontinuities as $k/2$. For a meta-analysis with $k = 20$, this is manageable; for a meta-analysis with $k = 200$, estimating more than one hundred parameters (including a mean and variance component) may be impossible. Furthermore, the difference in assumptions between the two models can point to radically different conclusions (see the example later in this chapter).

In keeping with the importance of triangulation, we encourage the use of these models as part of a toolbox of assessments, but warn meta-analysts that the models may be inestimable under some circumstances.

18.4.6.1.2 Hedges. Hedges proposed a similar model that assumes a step function over *p*-values (1992). His model differs from Dear and Begg's (1992) approach because the analyst must specify steps at perceived milestones in statistical significance. These milestones are based on the perception that a *p*-value of .049 is considerably different from one of .051, that .011 is different from .009, and so on. (Often $p = .50$ is a particularly relevant cut point because it reflects the point at which many effect-size metrics change from positive to negative.) Weights representing the relative likelihood of survival for the intervals are estimated in the context of a random-effects model, and all parameters (weights, the mean effect, and the variance component) are estimated simultaneously by the method of maximum likelihood. The model uses only two-tailed *p*-values, which cannot represent the direction of the effect. (Software for estimating

the two-tailed model is no longer available.) Vevea, Nancy Clements, and Hedges modified the model to use one-tailed *p*-values (1993). Models based on one-tailed *p*-values can still represent a two-tailed selection pattern. A pair of one-tailed *p*-values can define a two-tailed value by employing, for example, .025 and .975 in place of .05. This provides freedom from the constraint that selection must operate identically for positive and negative effects, which is an unlikely phenomenon. One- and two-tailed selection patterns fundamentally reflect the assumption of asymmetry (one-tailed) versus symmetry (two-tailed) of the weight-function model.

The method employs weighted distribution theory: the usual random-effects likelihood is multiplied by the weight for the *p*-value interval of each study, then renormalized. The software first estimates a conventional fixed- or random-effects model. Then the meta-analytic model is reestimated using the weighted likelihood. In addition to the mean and variance component, the model estimates all but one of the weights associated with the *p*-value intervals. To identify the model, the weight for the most significant range of *p*-values is fixed at 1.0. Other weights are interpreted relative to that first weight, and can actually exceed 1.0. Hence, they are not directly interpretable as probabilities. This weighted model provides mean and variance component estimates adjusted for publication bias, as well as estimated weights reflecting the relative likelihood of observing effect sizes in each interval. In addition, a likelihood-ratio test for publication bias compares the conventional model to the adjusted model.

18.4.6.1.3 Vevea and Hedges. Vevea and Hedges (1995) later added the possibility of including study-level covariates to the Vevea, Clements, and Hedges (1993) model. This can remove confounding effects in the distribution of effect sizes if asymmetry in the funnel plot is partly due to the presence of covariates. This weight function model can accommodate a full linear model for the mean effect, including dichotomous and continuous predictors, and can provide estimates of those predictors adjusted for publication bias. A particular advantage is that in some cases, certain classes of effects may remain virtually unaffected by the presence of the selection model, while others may be strongly affected.

The model does have some flaws—in particular, it does not perform as well with smaller meta-analyses, and it cannot estimate weights for ranges of *p*-values in which no observed effect sizes fall. It requires no precise number of effect sizes. Instead, meta-analysts must ensure that there are at least some observed effects in each range of *p*-values they specify, and must keep in mind that weights for intervals with few observed effects will be poorly estimated. Additionally, the model is a selection model, and selection models are often dismissed for their complexity. The model does require users to think about the selection process and to specify some relevant *p*-value breakpoints, but this is not necessarily a flaw. It is unlikely that a phenomenon as complex and multifaceted as publication bias could be adequately handled without some careful consideration.

Despite its flaws, the model has a number of positive features. First and most important, it is capable of handling both random and systematic heterogeneity. Many other assessments cannot accommodate linear models; meta-analysts can still use them on homogeneous subsets, but this may not be practical and, for continuous moderators, may be impossible. Additionally, in terms of performance, a simulation shows that variations of the Hedges model outperform both *p*-curve and *p*-uniform (1992). Such variations have narrower confidence intervals and are robust both to heterogeneity and to differing selection strengths (McShane, Böckenholt, and Hansen 2016). An earlier study by Hedges and Vevea also finds that such models are robust to violations of assumptions about the distribution of random effects—that is, to nonnormal distributions (1996).

Thus far, the Vevea and Hedges method has not seen much use, likely because no user-friendly software has been available (1995). However, Coburn and Vevea have released an *R* package to *CRAN* (the Comprehensive R Archive Network) titled *weightr* (2016a). The program is capable of estimating both the Vevea and Hedges model and the modified Vevea and Woods version described in the following section (2005). The same software is also available through a web-based point-and-click Shiny application (Coburn and Vevea 2016b).

Simulation studies of similar selection models indicate that the Vevea and Hedges model bears promise and will likely perform well under realistic circumstances, whether in the presence of systematic heterogeneity, random heterogeneity, or both (Vevea and Hedges 1995; McShane, Böckenholt, and Hansen 2016). This class of models also appears robust to different patterns of selection, which is a crucial trait given that researchers can never know the true underlying selection pattern. Meta-analysts would be remiss to overlook this model in favor of its simpler counterparts. The release of software will allow the Vevea and Hedges model to see increased use.

18.4.6.1.4 Vevea and Woods. In 2005, Vevea and Woods published a paper presenting a modification of

the Vevea and Hedges (1995) model. Some meta-analysts were disappointed because their data sets were too small to allow estimation of the Vevea and Hedges model. (With small data sets, it is often not possible to estimate weights for more than one or two *p*-value intervals.) There was interest in a method that could allow the user to specify not only *p*-value cut points, but also weights for the *p*-value intervals—a sensitivity analysis tool that would enable the user to explore how the conditional means of a data set might vary under different bias patterns.

The Vevea and Woods model provides this adaptation (2005). With it, there is no need to ensure that the data set is large enough, or even that there are observed effect sizes in every *p*-value interval. The meta-analyst merely specifies the *p*-value cut points of interest and a set of hypothetical weights for the corresponding *p*-value intervals, and the model produces estimates of the adjusted conditional means and variance component under the specified conditions. Because the model is not actually estimating parameter values, the standard errors and confidence intervals are no longer meaningful, nor is the likelihood-ratio test comparing the unadjusted and adjusted models. This does not reduce the impact of the model, however. It is still a valid sensitivity analysis tool that can provide the curious meta-analyst information about how specified selection bias patterns could affect their data, or about how robust their data are to selection bias. When moderators are included in the analysis, the results may show that a subset of effects identified by the linear model are virtually unaffected by any trial bias pattern.

Because the pattern of bias is imposed by the researcher rather than estimated from the data, sometimes the mean and variance component estimates can be adjusted relative to an extreme or unrealistic selection pattern. Kepes and McDaniel record an example of this (2015). To understand why, imagine a case in which the meta-analyst specifies weights of zero for all *p*-value cut points (indicating that no effect sizes can occur). In such a scenario, the estimates will obviously be nonsensical, if the model even converges; estimates may blow up or reduce to zero if extreme selection patterns are imposed. Researchers must remember that they are merely observing the reaction of the estimates to varying scenarios; they should assess the change in estimates *across* scenarios to determine whether their data set is robust to different selection patterns.

The Vevea and Woods (2005) model is a convenient workaround for meta-analysts who wish to implement the Vevea and Hedges (1995) model, but who do not have enough effect sizes to estimate weights. In this way, it is a useful addition to the literature, and helps make selection modeling a feasible option for more researchers.

18.4.6.2 Suppression as a Function of Effect Size and Its Standard Error Another class of models addresses publication bias by assuming a relationship among effect sizes, their standard errors, and the likelihood of their surviving the selection process.

18.4.6.2.1 Copas and Shi. John Copas and Hu Li initially proposed a selection model that functions as a sensitivity analysis in 1997; in subsequent years, Copas and Shi (2000, 2001) published several variations of the model. The method is frequently cited in discussions of selection models, as demonstrated by the fact that the original paper has received more than 340 citations, but it has not seen much practical use. Recently, Schwarzer, Carpenter, and Rücker (2016) created a software package called *metasens* using *R* (R Core Team 2016). The package implements the model and provides guidelines for its interpretation. As a result, the approach is gaining in popularity as more meta-analysts use it (Preston, Ashby, and Smyth 2004; Bennett et al. 2004). Some researchers even advocate a Bayesian implementation, which may avoid the issue of specifying values for the *a* and *b* parameters (Mavridis et al. 2012).

The Copas and Shi selection method (2001) combines two models: a population model that is equivalent to the usual random-effects meta-analytic model, and a model in which the probability of a study being published is a linear function of its reported standard error. There are two parameters in this linear model, the intercept (*a*, or the overall proportion of studies published when the standard error of those studies is zero) and the slope (*b*, or the relationship between standard error and publication). This linear model can be rewritten as a propensity model, where a study is selected for publication if and only if its propensity is greater than zero (Copas and Shi 2001; Copas and Li 1997). A correlation parameter links the observed effect sizes and their estimated propensities. A correlation of zero indicates the complete absence of publication bias, or a case in which effect sizes are published regardless of their standard error, while a positive correlation indicates the presence of bias (Copas and Shi 2001). Therefore, the conditional random-effects model represents the observed effect sizes, given that their propensity score is greater than zero (Copas and Shi 2000).

The selection model involves a total of five parameters—the mean effect size, the variance component, the correlation between effect sizes and propensities, and the slope and intercept (*a* and *b*) for the propensity model. No software to incorporate moderator variables is currently

available, but Copas and Shi indicate that the random-effects population model could easily be replaced with a mixed-effects model (2001). The problem with estimation, however, lies with the *a* and *b* parameters; they are not identified, because not enough information is available (the meta-analyst never knows how many studies remain unpublished). Copas and Shi demonstrate this by proving that the likelihood function for *a* and *b* is almost a plateau, so that using maximum likelihood estimation is near impossible. To solve this problem, they propose entertaining a series of specified values for *a* and *b*, then assessing the impact of those values on the average effect size and variance component. In that way, although all publication bias models should be treated as a sensitivity analysis, their model *must* be; it is a sensitivity analysis by nature, like the Vevea and Woods (2005) approach.

The Copas and Shi (2001) selection model features an algorithm that chooses a range of values for *a* and *b* and then uses maximum likelihood estimation to calculate the average effect size for each pair of values. (On occasions when the model chooses a range that produces uninterpretable results, the user can manually specify a range.) These results demonstrate how the average effect size changes as the likelihood of small studies being published changes. The relationship is easier to observe graphically, and four types of plots aid in its interpretation. The first is a standard contour-enhanced funnel plot. The second is a contour plot of the adjusted effect size against the values of *a* (on the *x*-axis) and *b* (on the *y*-axis), with the values representing no publication bias in the top right (Carpenter et al. 2009). If the contour lines are spread far apart, the adjusted effect size does not change much as the values of *a* and *b* change, and appears robust. The third plot explores this further; it plots the probability of publishing the study with the smallest sample size (on the *x*-axis) against the corresponding adjusted effect size (on the *y*-axis). If this relationship has a slope of zero, the effect size appears to be robust; otherwise, it may be affected by bias (Carpenter et al. 2009). Finally, the fourth plot involves the *p*-values for a likelihood-ratio test that assesses whether selection bias remains. The *p*-values (on the *y*-axis) are plotted against the probability of publishing the smallest-*N* study (again on the x-axis). The point where the plotted curve crosses the horizontal dashed line indicates that the corresponding probability on the x-axis is the most likely probability according to the model (Carpenter et al. 2009).

The Copas and Shi selection model has several positive features (2001). It can accommodate not only random (between-study or unobserved) heterogeneity captured by the variance component but also systematic (or observed) heterogeneity through incorporation of moderators. It can produce adjusted estimates of all the parameters of interest for each specified level of publication bias. In addition, the emphasis on sensitivity analysis encourages meta-analysts to view the model results as flexible, rather than accepting them as truth.

There is a dearth of information from simulations assessing the model's performance. The vast majority of manuscripts exploring the Copas and Shi (2001) model do so empirically, comparing it with other publication bias assessments using a limited set of observed data sets (Carpenter et al. 2009; Mavridis et al. 2012; Schwarzer, Carpenter, and Rücker 2010). Rücker, Carpenter, and Schwarzer (2011) recently presented the results of a small-scale simulation evaluating the Copas and Shi model, but noted that doing so was time-consuming and difficult, and that their simulation neglected to assess extreme heterogeneity and small sample sizes. A thorough simulation of the Copas and Shi model, perhaps along with competing selection models, would be informative.

The Copas and Shi model is a valuable addition to the body of selection models for publication bias, and its increasing popularity is promising (2001). Like the Vevea and Woods (2005) model, that of Copas and Shi does not estimate a selection pattern from the data; it imposes a range of possible patterns and observes the results. Therefore, it may also work well with smaller meta-analyses. Software to implement the model is available, which may encourage its use in the future.

18.4.6.2.2 Rücker. Rücker, Carpenter, and Schwarzer first published the Rücker limit meta-analysis method in 2011. The underlying model is an extended random-effects model that takes account of a possible relationship between effect size and sample size by allowing effect size to depend on standard error. Part of the model is based on earlier simulation work by Rücker, Schwarzer, and Carpenter (2008), which involved artificially inflating the sample size of effect sizes by a given factor of *M*. The model is also based on the original Egger's linear regression (Egger et al. 1997).

The concept of limit meta-analysis begins with the usual random-effects model, with an added parameter *alpha* that represents a small-study effect by allowing effect size to depend on standard error (Rücker, Carpenter, and Schwarzer 2011). The method then considers a situation where the sample size of all observed effect sizes is inflated by a factor of *M*. As *M* approaches positive infin-

ity, so does sample size; the effect sizes become infinitely precise, and variation due to sample size disappears—between-studies heterogeneity is all that remains. Estimating an original Egger's regression (Egger et al. 1997) on the observed effect sizes yields an intercept and slope. The intercept corresponds to the alpha parameter, and the slope is an estimate of the average effect size with a standard error of zero, or the infinitely precise effect. The limit meta-analysis method creates a new data set by transforming the original effect sizes so that they are centered on the slope from the Egger's regression. The slope of an Egger's regression calculated on this centered data set is the estimate of the average effect size, adjusted for small-study effects. A test on the intercept, or alpha, assesses the presence of a small-study effect. Finally, a test for heterogeneity on the centered data addresses the question of whether residual heterogeneity is present after adjustment.

Because it is based on Egger's regression, limit meta-analysis may presumably also incorporate a linear model, although no manuscripts demonstrate this feature. Limit meta-analysis does not produce an adjusted variance component, but does test for the presence of heterogeneity.

Rücker, Carpenter, and Schwarzer (2011) explored the performance of limit meta-analysis, generating and suppressing the data according to the Copas and Shi model, and find that its adjusted estimate was less biased than those from trim and fill and the Copas and Shi model (2001). Of course, the Copas and Shi approach is purely a sensitivity analysis, so its bias depends on the particular parameter settings used in the simulation. Limit meta-analysis was the most conservative of the three. As the size of the small-study effect increases, so does the performance of the limit meta-analysis method in comparison to the usual random-effects model (Rücker, Carpenter, and Schwarzer 2011).

Berlin and Robert Golub briefly explored the performance of the limit meta-analysis method, but further research into its performance would be beneficial (2014). The question of bias also remains. This method relies on the adjusted estimate from Egger's regression, so its overly conservative nature may be due to the same violated assumption that impacts PET-PEESE.

18.4.6.3 Bayesian Approaches After early interest in developing Bayesian approaches to address publication bias, there was a lengthy gap in new developments. Recently, however, there has been a resurgence of activity in Bayesian methods.

M. J. Bayarri and Morris DeGroot (1987) introduced a Bayesian method similar to Hedges's (1984) early approach in that it restricts attention to statistically significant outcomes. Geof Givens, David Smith, and Richard Tweedie developed a method similar to Hedges's (1992) early version of the step-function model (1997). Nancy Silliman (1997) presented, in a random-effects context, Bayesian models that estimate weight functions similar to those that both Hedges and Olkin (1985) as well as Iyengar and Greenhouse (1988) describe. Silliman also developed more complex weight-function models, including one that estimates weights as a step function of *p*-values with unknown cut points between intervals. Daniel Larose and Dipak Dey (1998) offered a similar method, emulating Iyengar and Greenhouse with a random-effects model.

More recently, Dimitris Mavridis and his colleagues developed a Bayesian implementation of the Copas approach (2012). Maime Guan and Joachim Vandekerckhove (2016) describe a method that considers four possible models—no selection, extreme selection with only statistically significant effects, nonsignificant results published with unknown but constant probability, and the Givens and colleagues model (1997). Their approach is to use Bayesian model averaging over the four competing models. Although it could be argued that these four models are not necessarily the best choices, the idea of Bayesian model averaging in this context is intriguing.

Other papers proposing new Bayesian approaches to addressing publication bias are currently under review. Thus, the Bayesian toolbox is likely to be expanded in the near future.

All of these Bayesian approaches have a common shortcoming: to our knowledge, accessible estimation software is not available. Hence, the meta-analyst who is not well versed in Bayesian estimation would find it difficult to employ these methods.

18.5 METHODS TO ADDRESS SPECIFIC DISSEMINATION BIASES

Incomplete data reporting may occur at various levels below the suppression of whole studies. Analyses of specific outcomes or subgroups, for example, may have been conducted but not written up and therefore are not available for meta-analysis. Similarly, all the details of an analysis required for meta-analysis may not be reported. For example, the standard error of an effect size or a study-level covariate, which is required for meta-regression, may not have been published.

This latter category of missing data may be entirely innocent, simply due to a lack of journal space or awareness about the importance of reporting such information.

Such missingness may not be related to outcome, and data can be assumed missing (completely) at random. If that is the case, standard methods for dealing with missing data can be applied to the meta-analytic data set, though this is rarely done in practice (Little and Rubin 1987; Pigott 2001; Sutton and Pigott 2004). More ad hoc methods can also be applied as necessary (Song et al. 1993; Abrams, Gillies, and Lambert 2005). However, meta-analyses in which data are missing completely at random are most likely quite rare.

If the data are missing for less innocent reasons, then it is probably safest to assume data are not missing at random. That is the typical assumption made when addressing publication bias, though it is not often framed as a missing data problem. Outcomes and subgroups may be suppressed under mechanisms similar to those acting on whole studies, so missingness may manifest itself in a similar way, and therefore the methods covered may be appropriate to address it. There may be advantages, however, to developing and applying methods that address specific forms of missing information. This area of research is in its infancy, although Coburn and Vevea have taken steps toward developing models in which the bias pattern may vary with study characteristics (2015).

18.5.1 Outcome Reporting Biases

Outcome reporting bias occurs when a study measures multiple outcomes, and those outcomes that are statistically significant are more likely to be published than those that are not. The issue of outcome reporting bias has received considerable attention in recent years, and empirical research indicates that it is a serious problem, especially for randomized controlled trials in medicine (Hahn, Williamson, and Hutton 2002; Chan, Hrobjartsson, et al. 2004; Chan, Krleza-Jeric, et al. 2004; Chan and Altman 2005). Although few studies examine outcome bias in the social sciences, some evidence indicates that it affects education research (Pigott et al. 2013). A recent survey of psychologists found that at least 63 percent did not report all outcome measures that they assessed (John, Loewenstein, and Prelec 2012). Together, this evidence reinforces the presence and severity of outcome bias.

Although most methods for assessing publication bias are sensitive to outcome reporting bias, they cannot distinguish between that and publication bias from other sources. These methods cannot accommodate patterns of missing data across multiple outcomes measured across all studies, or information across all reported outcomes. A method has been developed that does consider such patterns; it assumes that the most statistically significant outcomes from some fixed number of identically distributed independent outcomes are reported (Hutton and Williamson 2000; Williamson and Gamble 2005). Although its assumptions are unrealistically strict, the model does provide an upper bound on the likely impact of outcome reporting bias, and could help determine whether contacting study investigators for the potentially missing data would be cost effective. However, the bulk of citations of these articles are methodological rather than empirical work, so it appears that employment of the method is not common.

Daniel Jackson, John Copas, and Alexander Sutton developed a selection model for a specific application—success rates of surgery for emergency aneurysm repair—to address outcome reporting bias (2005). The model relies on a specific outcome that it assumes was reported without bias to learn about an outcome that obviously was not. Assuming no other sources of bias at any level, the selection model was identifiable and yielded adjusted estimates. Of course, the assumption that no other bias exists is very restrictive, and Jackson and his colleagues report that a model capable of incorporating both outcome- and study-level bias is in development.

18.5.2 Subgroup Reporting Biases

Subgroup reporting bias is similar to outcome reporting bias in that it involves the omission of one or more uninteresting or nonsignificant subgroup analyses. Either some subgroup results are published and others are not, or all subgroup results may be excluded (Hahn et al. 2000). Subgroup bias has received little attention in the research literature, although there are indications that it exists in medical research (McIntosh and Olliaro 2000). Others note that the prevalence of outcome bias implies the existence of subgroup bias as well (Hahn et al. 2000).

Seokyung Hahn and his colleagues (2000) suggest a sensitivity analysis approach, similar to the Jane Hutton and Paula Williamson (2000) approach for outcome reporting bias, which involves data imputation for missing subgroup analyses under the assumption that the unpublished analyses were nonsignificant. This is a useful start; however, in general, the issue of subgroup bias is underresearched. In the meantime, research guidelines such as PRISMA (preferred reporting items for systematic reviews and meta-analyses) recommend that meta-analysts report all analyses, regardless of significance, and indicate

whether they were planned (Liberati et al. 2009). Doing so may help reduce the impact of subgroup bias.

18.5.3 Time-Lag Bias

Time-lag bias occurs when research with large effect sizes or significant results tends to be stopped earlier than originally planned, published more quickly, or both (Hopewell et al. 2007)—in other words, when the speed of publication depends on the direction or strength of the results (Jadad and Rennie 1998). When such bias operates, the first studies to be published will often show systematically greater effect sizes than subsequently published investigations (Trikalinos and Ioannidis 2005). The cumulative effect size then diminishes over time.

The Proteus effect, named by Thomas Trikalinos and John Ioannidis (2005), is a similar time-related phenomenon in which the exciting findings of the first published study are followed by a series of equally exciting, contradictory studies, while intermediate, less exciting studies are published later on. In the case of the Proteus effect, because large effect sizes in opposite directions are published most quickly, the cumulative effect size may actually increase over time. The effect is named after Proteus, a god who rapidly transformed himself into different figures (Trikalinos and Ioannidis 2005).

Evidence that time-lag biases exist in the field of genetic epidemiology is strong (Ioannidis et al. 2001). It also appears in child psychiatry (Reyes et al. 2011), in clinical trial research (Clarke and Stewart 1998), and in management and industrial-organizational psychology (Banks, Kepes, and McDaniel 2012; Kepes et al. 2012), among others.

Methods for the assessment of time-lag biases are thoroughly described elsewhere (Trikalinos and Ioannidis 2005) and are not included in this chapter.

18.6 EXAMPLES

18.6.1 Data Sets

In this section, we illustrate some of the available methods using two empirical meta-analytic data sets that differ in size: one that is large (containing more than four hundred effect sizes), and one that is small (containing fewer than twenty). Both data sets are available as supplementary material. The large data set is from a social science meta-analysis, consisting of standardized mean differences. The small data set comes from a medical meta-analysis, and consists of log risk ratios.

We use *R* version 3.2.4 for most analyses (R Core Team 2016). As we describe our results, we include relevant sections of *R* code in the text so that interested readers can replicate the examples.

18.6.1.1 Psychotherapy Efficacy The first data set is from a well-known meta-analysis performed by Mary Smith, Gene Glass, and Thomas Miller on the efficacy of psychotherapy (1980). We use a subset of the original data that consists of studies in which the psychotherapy effects being compared include both behavioral and systematic desensitization treatments for phobias. The phobias themselves are also divided into two groups, one consisting of patients suffering from "complex" (multiple) phobias and one of those suffering from "simple" (only one) phobias. The original data set included some effect sizes that modern-day meta-analysts might consider implausibly large; for example, one study reported a standardized mean difference of 25.33. To avoid complications that such huge effects can induce, we deleted five cases with effect sizes larger than 4.0. Of the 489 effect sizes that remain, 216 employ behavioral treatments and 273 employ desensitization therapies. Positive effect sizes indicate effectiveness of psychotherapy.

This is the same data set that Vevea and Hedges (1995) used to demonstrate the use of a linear model for estimating effect size in the presence of publication bias (one of the models included in this chapter). They find that a funnel plot of the effect sizes demonstrated typical one-tailed selection, and the selection model resulted in a reduction of the mean effect size by as much as 25 percent, in the case of desensitization treatment of complex phobias. These results indicate that publication bias does affect this data set.

We read the psychotherapy efficacy data set into *R* with

```
glass <- read.csv("data glass.csv",
header=TRUE)
```

and create variables for the effect sizes and sampling variances:

```
glass_y <- glass$g
```

and

```
glass_v <- glass$v
```

We also create variables for the three moderators we will be using: whether the therapy was behavioral modification,

```
glass_b1 <- glass$b1
```

whether the patients' phobia was simple or complex,

```
glass_b2 <- glass$b2
```

and whether there was an interaction,

```
glass_b3 <- glass$b3
```

which is the product of glass_b1 and glass_b2.

18.6.1.1.1 Funnel Plots. No special software is necessary to create a funnel plot using *R*. For users who prefer to do so, most common meta-analysis packages include a funnel plot function. Wolfgang Viechtbauer's *metafor* includes *funnel()*, which yields both traditional and contour-enhanced funnel plots, with the enhanced plot as the default (2010). Schwarzer's *meta* includes its own *funnel()*, which performs similarly (2016). The funnel plots presented here were created using the basic *R* scatterplot tools, with the margins extended so that there is approximately 5 percent white space between axes and data points. Adding white space aids in the interpretation of asymmetry if data points fall extremely close to the axes. We plot standard error on the *x*-axis and effect size on the *y*-axis.

Figure 18.6 shows the contour-enhanced funnel plot, using *metafor's funnel()* function. This type of plot may be misleading under some circumstances, however, as discussed earlier.

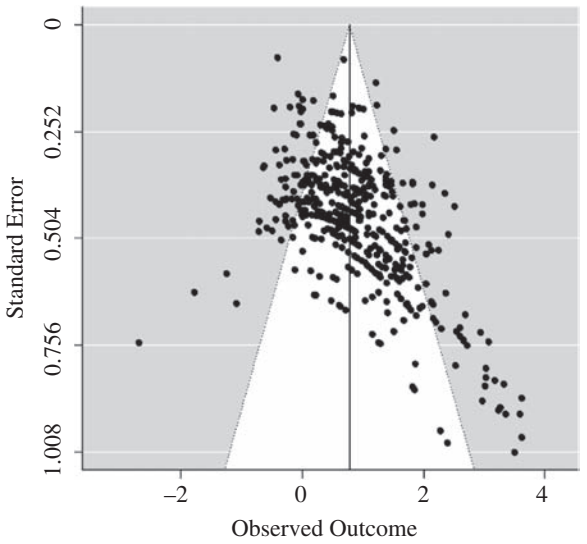

Figure 18.6 Contour-Enhanced Funnel Plot, Psychotherapy Data

SOURCE: Author's tabulation.

We calculated the amount necessary to expand the range of the *x* and *y* axes by 5 percent. Then we created a space for the funnel plot with

```
plot(c(min(sqrt(glass_v))-.0465,
max(sqrt(glass_v))+0.0465),c(min(glass_y)-
0.3145,max(glass_y)+0.3145),type='n',
xlab="Standard Error",ylab="Effect Size")
```

We added the scatterplot points with

```
points(sqrt(glass_v),glass_y)
```

The resulting funnel plot appears in figure 18.7. We computed the mean effect size for this data set using the *metafor* package's *rma()* function:

```
rma(glass_y, glass_v, method='ML')
```

We used a random-effects model and maximum likelihood estimation. This yielded a mean of 0.70 and a variance component of 0.28, which corresponds to I^2 of 66.51 percent. I^2 is fairly large, indicating that the results of some methods—particularly those that cannot accommodate heterogeneity—may be less reliable.

This data set is large, so it is easier to assess asymmetry in the funnel plot. There are many very large effect sizes ($d > 1.00$) with large standard errors (> 0.60) that are not mirrored by small or negative effect sizes; in fact, only *four* effect sizes with standard errors greater than 0.60 fall equally far below. This is an example of asymmetry associated with characteristic one-tailed selection bias. The plot suggests that concern about publication

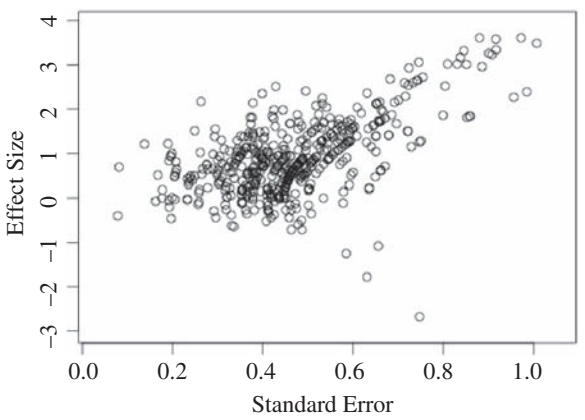

Figure 18.7 Effect Size Against Standard Error, Psychotherapy Data

SOURCE: Author's tabulation.

bias may be appropriate. (The curvature evident in the plot is due to the dependence of standard errors on standardized mean difference; studies with equal sample sizes will have increasingly large standard errors as their effect sizes increase.)

18.6.1.1.2 Cumulative Meta-Analysis. Both *metafor* and *meta* include a function for cumulative meta-analyses; in *metafor* this is *cumul()* and in *meta* it is *metacum()*. Users can create an object to store the results of these analyses and then pass that object to *forest()* using either package. We used *metafor* for our cumulative meta-analyses and forest plots.

The forest plot is not featured here. A forest plot of a cumulative meta-analysis involving 489 effect sizes is more or less impossible to read; the annotations are so tiny as to be invisible, and the average effect estimates resemble a thick black blur. Instead, it is more informative to consider the cumulative meta-analysis itself:

```
glass_cumul <- cumul(glass_rma, order =
order(sqrt(glass_v)))
```

The effect size with the least precision belongs to study 100, with $d = -0.40$. As soon as the second least-precise effect, study 99, is added, the estimate of mean effect jumps to $d = 0.14$. It proceeds from there to $d = 0.49$ with the third least-precise effect and, by the time all effects are included, has drifted all the way to $d = 0.78$. This drift is a sign of a relationship between study size and effect size, which indicates the presence of publication bias.

18.6.1.1.3 Trim and Fill. We used the *trimfill()* function of the *R* package *metafor*. The *R* package *meta* also includes a *trimfill()* function. Users who are interested in separate control of fixed-effect and random-effects models for imputing data and estimating parameters may wish to use *meta*. Duval and Tweedie recommend that trim and fill be used as a sensitivity analysis (2000a, 2000b), so we present the results of trim and fill using both the L_0 and the R_0 estimators, and we consider the possibility of publication bias in favor of both smaller and larger effects.

For each data set, we estimated four trim and fill models. We specified that studies were missing on either the left or the right side of the funnel plot (suppression of smaller or larger effects). In view of the funnel plot, specifying missing effects on the right side makes little practical sense, but we included it for demonstration purposes. We also specified the L_0 or R_0 estimator. The literature is ambiguous about the performance of L_0 versus R_0; see section 18.4.3 for details. We used the *metafor* function *trimfill()*, which can handle either "fixed-fixed" or "random-random" models. (The *meta* package is more versatile in this respect.) The results presented here are "random-random." These factors yielded four separate models.

The models were variations of

```
trimfill(glass_rma, side="left",
estimator="L0")
```

where "L0" was exchanged for "R0" and "left" was exchanged for "right." The results of the trim and fill analyses for the psychotherapy efficacy data set are presented in table 18.2.

Only L_0 added effect sizes, and it did so on the left side of the plot (indicating a suppression of smaller effects). L_0 added 117 effect sizes, reducing the average effect from 0.78 to 0.52 (a difference of 0.26, or 33 percent) and increasing the variance component from 0.29 to 0.58 (a difference of 0.29, or 100 percent).

It seems that this data set may not be robust to publication bias. Nevertheless, an effect persists, even after adjustment.

18.6.1.1.4 Egger's Regression. The package *meta* contains the function *metabias()*, which can be used for both Egger's linear regression and the rank correlation test. The package *metafor* contains *regtest()*, which conducts Egger's linear regression; we used *regtest()*. The function allows users to specify whether they want to estimate a standard Egger's regression, a mixed-effects Egger's regression, or a random-effects Egger's regression. We estimated all three models to compare their conclusions, although because heterogeneity is present in the data set, the mixed- or random-effects models may perform better.

A standard Egger's regression using a weighted regression model

```
regtest(glass_rma, model="lm")
```

was statistically significant, $t(484) = 9.81$, $p < .0001$, indicating that publication bias, or funnel plot asymmetry, is present. The evidence under a random-effects meta-regression model acquired from

```
regtest(glass_rma)
```

was also statistically significant, $z = 12.26$, $p < .0001$.

We incorporated the moderators in this data set as well. There are three dichotomous moderators, as described. Their unadjusted conditional means are presented in the top row of table 18.2.

The mixed-effects variation of Egger's regression was also statistically significant, $z = 11.80$, $p < .0001$. All three variations indicate a relationship between study size and effect size, or that bias may be present.

Table 18.2 Summary of Results for Psychotherapy Data

Method		Overall Mean	BMOD, SP	BMOD, CP
Unadjusted		0.78	0.90	0.63
Trim and fill	left, L_0	0.52 (32.99%, M) *MO*	—	—
	right, L_0	0.78 (0%, A) *MI*	—	—
	left, R_0	0.78 (0%, A) *MI*	—	—
	right, R_0	0.78 (0%, A), *MI*	—	—
PET-PEESE		**–0.04 (105.13%, S) *S***	—	—
Vevea and Hedges	$p = 0.025$	0.68 (12.71%, A) *MI*	—	—
	multiple	0.47 (39.67%, M) *MO*	—	—
	multiple, LM	—	0.65 (27.55%, M) *MO*	0.36 (42.36%, S) *MO*
Vevea and Woods	moderate one-tailed	—	0.78 (13.67%, A) *MI*	0.49 (21.50%, M) *MO*
	severe one-tailed	—	0.56 (37.44%, M) *MO*	**0.21 (65.92%, S) *S***
	moderate two-tailed	—	0.82 (8.78%, A) *MI*	0.57 (9.55%, A) *MI*
	severe two-tailed	—	0.72 (20.11%, M) *MO*	0.49 (21.34%, M) *MO*
Copas and Shi		**0.10 (87.16%, S) *S***	—	—
Rücker		**0.09 (88.45%, S) *S***	—	—
p-uniform		1.06 (36.07%, M) *MO*	—	—

SOURCE: Author's tabulation.

NOTES: Adjusted estimates are reported unless row is labeled "Unadjusted." Percentage adjustment is in parentheses, followed by the Kepes, Banks, and Oh (2012) categorization (A for absent, or < 20 percent adjustment; M for moderate, or adjustment between 20 percent and 40 percent; S for severe, or adjustment > 40 percent). The Rothstein, Sutton, and Borenstein (2005) categorization follows in italics (MI for minimal, or adjustment is similar; MO for moderate, or adjustment is substantial, but key finding remains; S for severe, adjustment that calls the key finding into question). If both categorizations were "Severe," the cell is boldface. BMOD = behavioral modification, SYSDS = systematic desensitization, SP = simple phobia, and CP = complex phobia. ^{1} is the variance component without moderators; ^{2} is the variance component with moderators. Cells marked with a dash either were not or could not be estimated.

18.6.1.1.5 PET-PEESE. We were unable to locate any *R* package capable of implementing PET-PEESE. However, meta-analysts can easily construct the code themselves; PET-PEESE is a pair of linear regressions, which can be estimated using the base *R* function *lm()*.

We estimated the PET regression:

```
pet <- lm(glass_y~sqrt(glass_v),weights =
1/glass_v)
```

followed by the PEESE regression:

```
peese <- lm(glass_y ~ glass_v, weights =
1/glass_v)
```

We stored the estimates from these regressions and kept the PET estimates if PET was nonsignificant; otherwise, we kept PEESE.

PET was nonsignificant, indicating the absence of evidential value, with $p = .56$. This means estimating PEESE is not necessary, and the adjusted estimate of effect size from PET, $d = –0.04$ (see table 18.2) is a result of publication bias. The unadjusted random-effects mean for this data set is 0.78; PET-PEESE reduced the mean by 0.82, or 105 percent. This indicates that the effect size is an artifact of publication bias.

18.6.1.1.6 Nonparametric Correlation Test. The *meta* package contains the function *metabias()*, which can perform the rank correlation test. The *metafor* package contains *ranktest()*. We use *ranktest()* to conduct the analyses. The rank correlation test for funnel plot asymmetry is not model-based, and changing the meta-analytic model does not change the results. The rank correlation results we present here have all used Kendall's tau, but the *R* function can accommodate other correlation coefficients.

The rank correlation test was provided by

```
ranktest(glass_rma)
```

SYSDS, SP	SYSDS, CP	τ_1^2	τ_2^2
0.86	0.70	0.29	0.28
—	—	0.58 (98.98%, S) *MO*	—
—	—	0.29 (0%, A) *MI*	—
—	—	0.29 (0%, A) *MI*	—
—	—	0.29 (0%, A) *MI*	—
—	—	—	—
—	—	0.27 (7.85%, A) *MI*	—
—	—	0.37 (26.28%, M) *MO*	—
0.61 (28.82%, M) *MO*	0.42 (40.63%, S) *MO*	—	0.36 (22.22%, M) *MO*
0.73 (16.76%, A) *MI*	0.56 (20.60%, M) *MO*	—	0.32 (15%, A) *MI*
0.51 (40.26%, S) *MO*	0.28 (59.80%, S) *MO*	—	0.44 (56.07%, S) *MO*
0.78 (8.75%, A) *MI*	0.63 (10.23%, A) *MI*	—	0.27 (5.36%, A) *MI*
0.69 (19.95%, A) *MO*	0.54 (22.87%, M) *MO*	—	0.24 (15.36%, A) *MI*
—	—	—	—
—	—	—	—
—	—	—	—

and was significant, with a Kendall's tau of 0.32 and $p < .0001$. This indicates that we can reject the null hypothesis of no correlation and conclude a danger of publication bias, or of funnel plot asymmetry.

18.6.1.1.7 p-*Curve and* p-*Uniform*. *P*-curve is not available as an *R* package, but Simonsohn, Nelson, and Simmons created a web application (www.p-curve.com) (2014). The application requires users to enter data in the form of the original test statistics rather than effect sizes, likely to encourage meta-analysts to think carefully about which tests are included. We include a caveat here—we are demonstrating *p*-curve using empirical sets of effect sizes and sampling variances, not raw test statistics from the corresponding studies. This exercise is solely for demonstration purposes.

The graph produced by *p*-curve is presented in figure 18.8. The distribution of *p*-values is visibly right skewed. The binomial test for right skew, comparing the proportions of *p*-values below and above .025, was nonsignificant with $p = .13$, but continuous tests, both for the full *p*-curve and for the half *p*-curve, were significant: $z = -3.64$, $p = .0001$ and $z = -4.17$, $p < .0001$, respectively. These results indicate that these studies do contain evidential value; the effect is not completely due to publication bias.

Although not entirely necessary because right skew is present, the binomial test for underpowered studies is significant, $p = .01$. However, the continuous test for underpowered studies is not, with $z = -0.78$ and $p = .22$. Despite the fact that evidential value is present, the continuous test indicates that data may be underpowered, or (according to the application) that the evidential value is inadequate. The studies conducted may have had reduced power; perhaps the researchers did not conduct a priori power analyses.

No *R* package for *p*-uniform is available on the Comprehensive *R* Archive Network, but Robbie van Aert

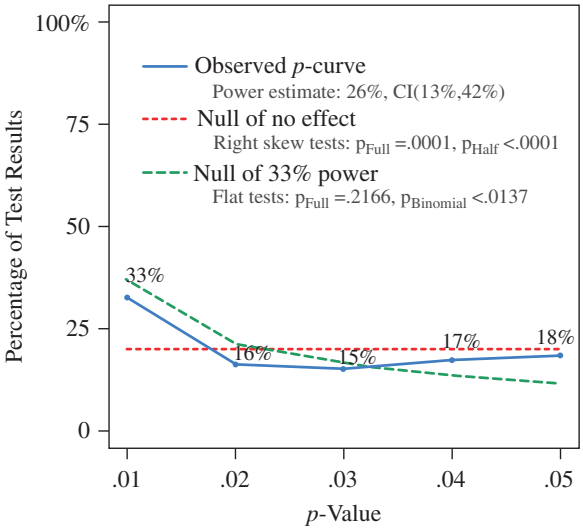

Figure 18.8 *p*-Curve, Psychotherapy Data

SOURCE: Authors' tabulation obtained directly from the *p*-curve application (v. 4.06), available at http://www.p-curve.com/app4/ (accessed January 7, 2019).
NOTE: The observed *p*-curve includes 92 statistically significant ($p<.05$) results, of which 52 are $p<.0.5$. There were 395 additional results entered but excluded from *p*-curve because they were $p>0.5$.

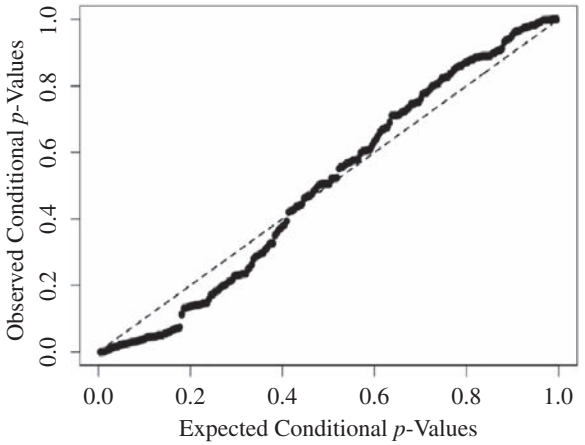

Figure 18.9 *p*-Uniform, Psychotherapy Data

SOURCE: Author's tabulation.

(2015) has uploaded a preliminary version of his package, called *puniform*, on GitHub. GitHub is a less regulated analog of CRAN. To install packages directly to *R* using GitHub, users must first install the *R* package *devtools*, then type *install_github("author/package")*. For *puniform*, this would look like *install_github("RobbievanAert/puniform")*. The function requires that users specify the alpha level of included studies. We entered an alpha level of 0.05. The function also produces a plot of observed conditional *p*-values against expected ones, so users can visually assess deviation from uniformity.

We estimated *p*-uniform:

```
puniform(yi = glass_y, vi = glass_v,
alpha = 0.05,side="right", method="P",
plot=TRUE)
```

The plot of observed versus expected *p*-values is in figure 18.9. There is some deviation from uniformity in the areas of the graph between expected *p*-values of about .60 and 1.00 and between about .10 and .40. The one-tailed test for publication bias was nonsignificant, with $z = -7.96$ and $p > .999$. The adjusted fixed-effect estimate was $d = 1.06$ (see table 18.2), $p < .001$. This is an increase of 0.44, or 71 percent.

P-uniform does not indicate that publication bias is a serious threat for this data set. It is unusual, though, that the mean was adjusted upward, indicating that larger studies were suppressed.

18.6.1.1.8 Excess Significance Test. No *R* package is available to implement the excess significance test. However, it is not complicated to conduct the test in base *R* by calculating the post hoc power for each significant study in your data set and multiplying them. We do not provide the *R* code in text as ours was lengthy and calculating power, of course, depends on the format of your effect sizes.

The product of the power for each significant effect size was 0.00 ($p < 0.10$), so we can reject the null hypothesis and conclude that publication bias may be present.

18.6.1.1.9 Dear and Begg. We faced an estimation problem with the psychotherapy efficacy data set, and were unable to obtain results from either the Dear and Begg (1992) weight-function model or the modified version by Rufibach (2011), because both versions attempted to estimate *n*/2 (or 244) parameters.

18.6.1.1.10 Vevea and Hedges. In 2016, Coburn and Vevea released an *R* package titled *weightr* that can estimate both the Vevea and Hedges (1995) model and the Vevea and Woods (2005) model. We use the package *weightr* to perform these demonstrations.

Meta-analysts who wish to use either the Vevea and Hedges (1995) or the Vevea and Woods (2005) model

will undoubtedly wonder about the *p*-value cut points—how many to specify, whether to specify a one-tailed or two-tailed pattern, which to specify, and so on. The choice of cut points is entirely up to the researcher. We do not recommend any specific pattern over another, nor do we intend for the cut points used here to become a canonical guideline. We cannot emphasize this enough. The researcher must consider the data set and select a series of cut points such that at least some observed effect sizes fall within each interval, and the cut points correspond to *p*-values that may be of psychological relevance. Regarding the latter, it is poor practice to present only the results of a model with a cut point at, say, $p = .129$ if the results drastically differ when the cut point is changed (for instance, to $p = .20$). Do not select cut points to force the model into significance or nonsignificance. Specify several different sets of cut points and observe the change in the adjusted estimates; if there are moderators, calculate conditional means and assess their changes as well. Perhaps most important, present the results of all models specified, to compare the adjusted estimates across models. *Always* report the cut points specified (preferably along with their rationale).

No guideline is in place for the number of cut points meta-analysts should specify for a given *k*—though, of course, it is impossible to estimate more cut points than there are effect sizes, and therefore the number of cut points must be less than *k*. The meta-analyst should survey the distribution of observed *p*-values and ensure at least some observed effect sizes in each specified interval, a fact that can be verified using the table=TRUE argument in *weightr*. If an interval is empty of observed effect sizes or contains only a few effects, this will immediately be evident, because *weightr* will yield a warning and the parameter estimates may be nonsensical or missing due to model nonconvergence.

For these examples, to maintain consistency across data sets, we have specified the same set of *p*-value cut points. Remember: The unadjusted mean for the psychotherapy efficacy data set is $d = 0.78$, and the unadjusted variance component is 0.29. All these results are presented in table 18.2.

First, we specify one *p*-value cut point at $p = .025$.

```
weightfunct(glass_y, glass_v)
```

The $p = .025$ corresponds to the positive tail in a two-tailed test with an alpha of 0.05. This yields a weight for the nonsignificant interval ($.025 < p < 1.00$) of 0.68, indicating that nonsignificant studies are 68 percent as likely to survive selection as significant ones. The mean effect is adjusted downward to 0.68 (a change of 0.10, or 13 percent), and the variance component is also adjusted downward to 0.27 (a change of 0.02, or 7 percent). The likelihood-ratio test comparing the adjusted and unadjusted models is significant, $p < .05$, which indicates that the adjusted model fits the data better, and hence that publication bias is present.

Next, we estimate a more detailed one-tailed selection pattern ($p = .01, .025, .05, .10, .20, .30, .50$, and 1.00). The first four cut points are at *p*-values that correspond to common alpha levels: $p = .10$ is often referred to as *marginal significance*; $p = .025$ corresponds to the positive tail of a two-tailed test at an alpha level of .05; $p = 0.50$ represents the point at which most effect-size measures become negative. There are no cut points after $p = 0.50$ because we wish to specify one-tailed selection; $p = .20$ and $p = .30$ are included because enough observed effects fall in that range for the model to estimate weights.

We enter:

```
weightfunct(glass_y, glass_v, steps=c(0.01,
0.025,0.05, 0.30, 0.50, 1.00))
```

Now the adjusted mean effect size has been reduced even further, to 0.47 (a change of 0.31, or 40 percent), and the variance component has increased to 0.37 (a change of 0.08, or 28 percent). The likelihood-ratio test is still significant, indicating that the adjusted model is a better fit.

Finally, we include a linear model for the mean. We specify the same pattern of one-tailed *p*-value cut points as before. The command is:

```
weightfunct(glass_y, glass_v, mods=~glass_b1
+ glass_b2 + glass_b3, steps=c(0.01, 0.025,
0.05, 0.30, 0.50, 1.00))
```

The likelihood-ratio test comparing this adjusted model to its unadjusted counterpart is still significant, indicating that it is a better fit for the data. We are now presented with not only an adjusted variance component (of 0.36, a change of 0.07 or 24 percent) but also adjusted conditional means for all six groups.

Some conditional means were adjusted more than others, likely based on how much the effect sizes in that particular group are susceptible to publication bias. More specifically, the conditional means for both systematic desensitization and behavioral modification with complex phobias appear to be more heavily affected by publication bias than the other two conditional means (a change of 41 percent and 42 percent, respectively). The other two conditional means were changed by 28 percent and

29 percent. This difference occurs because the less affected subset of effects is systematically larger than other effects, so that the bulk of them fall within the range of the highly significant cut points, where weights tend to be high. The model can adjust some means more than others, as necessary. The results may indicate that more bias is present among these groups of effect sizes.

18.6.1.1.11 Vevea and Woods. The Vevea and Woods (2005) model also requires meta-analysts to specify a series of *p*-value cut points, this time along with a fixed weight for each interval. Because this model does not estimate any weights for the *p*-value intervals, it does not matter how many cut points the meta-analyst specifies. It is possible to specify more cut points than observed effect sizes (that is, the number of cut points can be greater than *k*). It is also possible to specify any weights for those intervals, bearing in mind that for convenience of interpretation the weights should be between 0 and 1. If the specified weights increase as the *p*-value cut points decrease—that is, if *p*-values < .05 are the most likely to survive—this represents traditional one-tailed selection. The model is flexible; researchers can specify two-tailed selection, or any selection pattern, using any *p*-value cut points.

Interpreting the results of the Vevea and Woods (2005) model comes with a caveat. The model does not estimate the pattern of selection; the researcher chooses a pattern of selection and imposes that pattern on the observed effect sizes, then the mean (or set of conditional means) and variance component are adjusted according to the pattern. The idea is to conduct multiple analyses with various weight patterns representing different degrees of selection severity. Some patterns may lead to ludicrous estimates of the mean (or conditional means). But often the mean, or a particular conditional mean, may be relatively unaffected by any reasonable pattern of weights. Under those circumstances, the researcher can be confident that the magnitude of the mean is not principally an artifact of *p*-value based selection. But none of these estimates should be regarded as a true bias-corrected estimate.

The *R* package *weightr* can implement the Vevea and Woods (2005) adaptation. We used *weightr* to conduct the following analyses. We specified the sets of cut points and weights ("moderate" and "severe" one-tailed and two-tailed selection) mentioned in Vevea and Woods, but we emphasize that these are not hard and fast guidelines for severity of selection. They are merely an indication of what severity of selection bias might look like. Bias patterns most likely vary widely both across and within fields, and we are merely using these weights for demonstration purposes. We do not mean for them to become canonical specifications. These sets of weights, and the bias patterns they theoretically represent, are presented in table 18.1.

The results of the analyses appear in table 18.2. The code we used consists of variations on

```
weightfunct(glass_y, glass_v, mods=~glass_b1
+ glass_b2 + glass_b3, steps=c(0.005,
0.010, 0.050, 0.100, 0.250, 0.350, 0.500,
0.650, 0.750, 0.900, 0.950, 0.990, 0.995),
weights=c(1, 0.99, 0.95, 0.80, 0.75, 0.65,
0.60, 0.55, 0.50, 0.50, 0.50, 0.50, 0.50))
```

We replaced the weights vector with the corresponding set of weights for each selection pattern.

Some of the conditional means are affected more than others. All means are most attenuated in the severe one-tailed bias condition, but the means for the group with simple phobias under both treatment conditions are reduced much less than the others. Even though all the conditional means are attenuated to one degree or another, however, these results at least indicate that none of the bias patterns we attempted reversed the direction of the effects, and only two means were reduced below 0.25 for any of the four selection patterns. This implies that, although publication bias has the potential to affect the mean estimates, the changes are only large for specific combinations of treatment and complex phobias. Hence, it is unlikely that publication bias, if present, would overturn conclusions about the positive effects of psychotherapy for these conditions.

18.6.1.1.12 Copas and Shi. The Copas and Shi (2001) model can easily be implemented using the *R* package *metasens* (Schwarzer, Carpenter, and Rücker 2016) and the function *copas()* (Carpenter et al. 2009).

We first estimated a random-effects meta-analysis with maximum likelihood:

```
glass_meta <- metagen(TE = glass_y, seTE =
sqrt(glass_v), method.tau = "ML")
```

Then we estimated the Copas and Shi selection model:

```
cop.glass <- copas(glass_meta)
plot(cop.glass)
summary(cop.glass)
```

The four plots produced by the Copas and Shi (2001) selection model *R* function (Carpenter et al. 2009) are presented in figure 18.10. We begin with the top right plot, the contour plot. The contour lines are straight, indicating

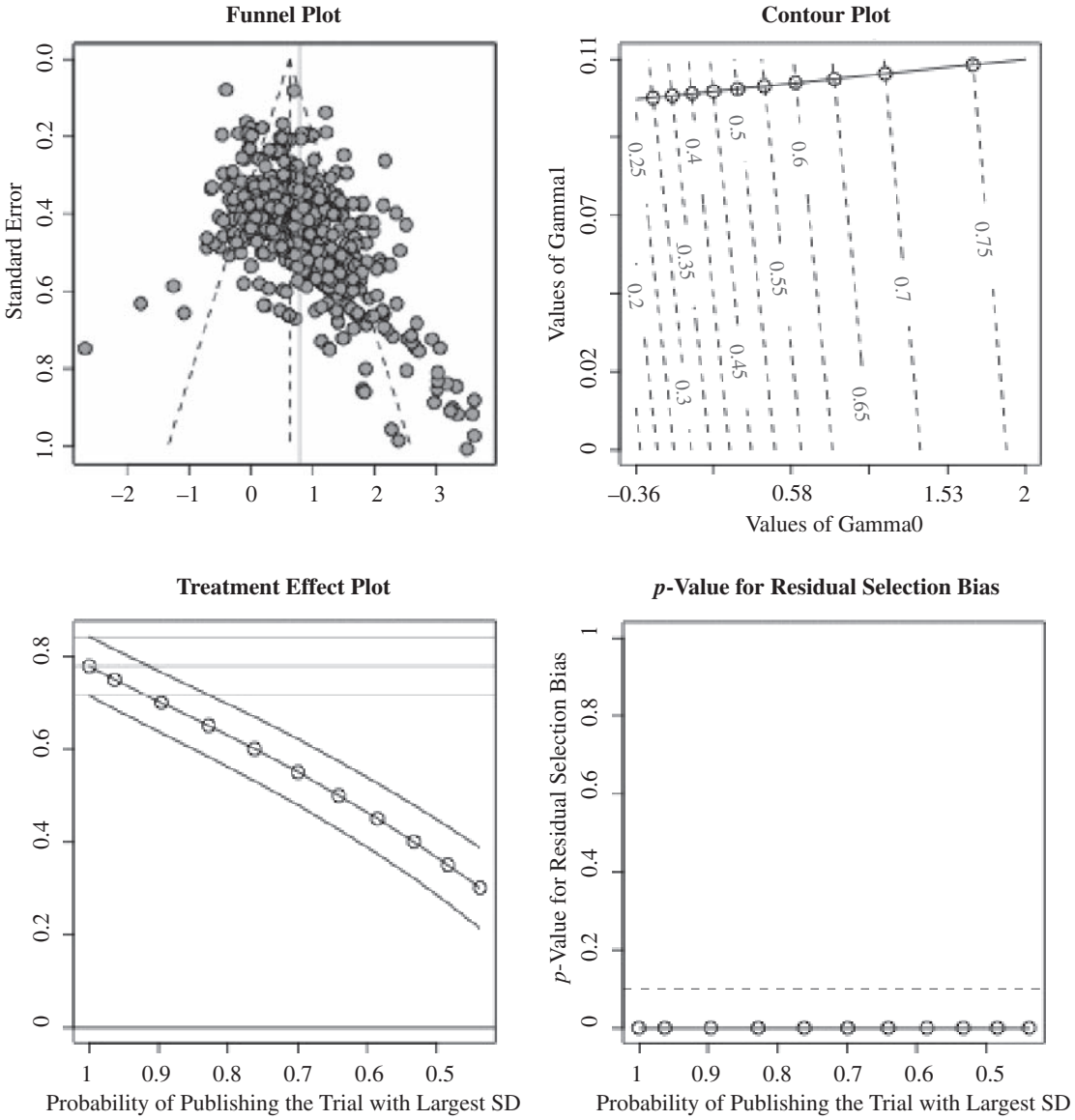

Figure 18.10 Copas and Shi, Psychotherapy Data, $a = -0.36$ to 2, $b = 0$ to 0.11

SOURCE: Author's tabulation.
NOTE: Gamma zero and gamma one represent a and b, respectively.

little difficulty in model estimation, and most of the contour lines are close together, indicating that the data set is not very robust to changes in selection bias patterns. The estimate at the top right of the contour plot, under little to no selection bias, is about $d = 0.75$. The unadjusted random-effects estimate for the psychotherapy data set is $d = 0.78$.

The treatment effect plot indicates that, with no selection bias (a probability of 100 percent), the estimated average effect size is about $d = 0.80$. When the probability of publishing the effect size with the largest standard error reaches 50 percent, the average effect size has plummeted to about 0.30.

The *p*-value plot gives us some problems. The *R* function automatically scanned a range of *a* and *b* values for which the *p*-value associated with residual selection bias never becomes nonsignificant, as it does not cross the horizontal dashed line. This means that the *R* function cannot give us an adjusted estimate. However, the choices for *a* and *b* can be manually altered. The function allows users to specify a range of values for *a* and *b*. We can see the range that the function's algorithm chose by looking at the axes of the contour plot; the algorithm scanned from –0.36 to 2 for *a* (gamma zero in the software nomenclature) and from 0 to 0.11 for *b* (gamma one). We change these and specify a wider range of values—from –2 to 2 for *a* and from 0 to 1 for *b*. This yields the plots in figure 18.11.

It appears that our conclusion was correct; the function was simply not scanning a wide enough range of values by default. Now, if we look at the treatment effect plot, we get an estimated mean effect size for probabilities beyond 20 percent—the effect size has dropped as far as about 0.10. The *p*-value plot shows that, when the probability of publishing an effect size with the largest standard error reaches 10 percent or so, the test for residual selection bias finally becomes nonsignificant. The most likely scenario given these observed data is a probability of 10 percent. This is very strong selection bias.

The adjusted mean effect size that the Copas and Shi (2001) model provides for this most likely scenario is $d = 0.10$ (see table 18.2), a change of 87 percent from the unadjusted $d = 0.78$.

18.6.1.1.13 Rücker Limit Meta-Analysis. To implement the Rücker et al. (2011) limit meta-analysis method, we used the *R* package *metasens* and the function it provides, *limitmeta()*. The *R* function yields a test of heterogeneity, a test of small-study effects (on alpha), a test of residual heterogeneity after accounting for small-study effects, and an adjusted estimate of the average effect size.

We estimated the Rücker limit meta-analysis method:

```
glass_limit <- limitmeta(glass_meta)
summary(glass_limit)
```

The results indicate a significant relationship between effect size and standard error. The test for small-study effects was significant, $Q(1) = 304.35$, $p < .0001$, as was the test for residual heterogeneity, $Q(487) = 1255.21$, $p < .0001$.

The adjusted random-effects estimate for the mean effect is 0.09 (table 18.2), with a 95 percent confidence interval from 0.01 to 0.17. According to the model, the effect size for a study having infinite precision is 0.09. This is a change of 0.69, or 88 percent.

18.6.1.2 Irritable Bowel Syndrome The second data set consists of nineteen trials examining the response rate of patients with irritable bowel syndrome to complementary and alternative medicine (CAM) therapies (Dorn et al. 2007). Only randomized, placebo-controlled trials were included. The CAM response rate was high across trials—more than 40 percent. Risk ratios greater than 1 indicate that patients undergoing CAM therapies had a higher response rate than those undergoing placebo therapies.

Spencer Dorn and his colleagues (2007) assessed publication bias using a funnel plot, Egger's regression, and Begg and Mazumdar's rank correlation test. They concluded that the funnel plot displayed asymmetry and found that Egger's regression was significant ($p = .03$) and the rank correlation was "trend[ing] towards significance" (632), with $p = .06$. They also noted that, of nineteen trials, twelve were statistically significant.

We read the irritable bowel syndrome data set into *R* with

```
ibs <- read.csv("data IBS.csv",
header=TRUE)
```

and create variables for the effect sizes and sampling variances:

```
ibs_y <- ibs$LogRR
```

and

```
ibs_v <- ibs$v
```

(The "header=TRUE" component of the R command is appropriate if the data file contains variable names in the first row, as this one does.)

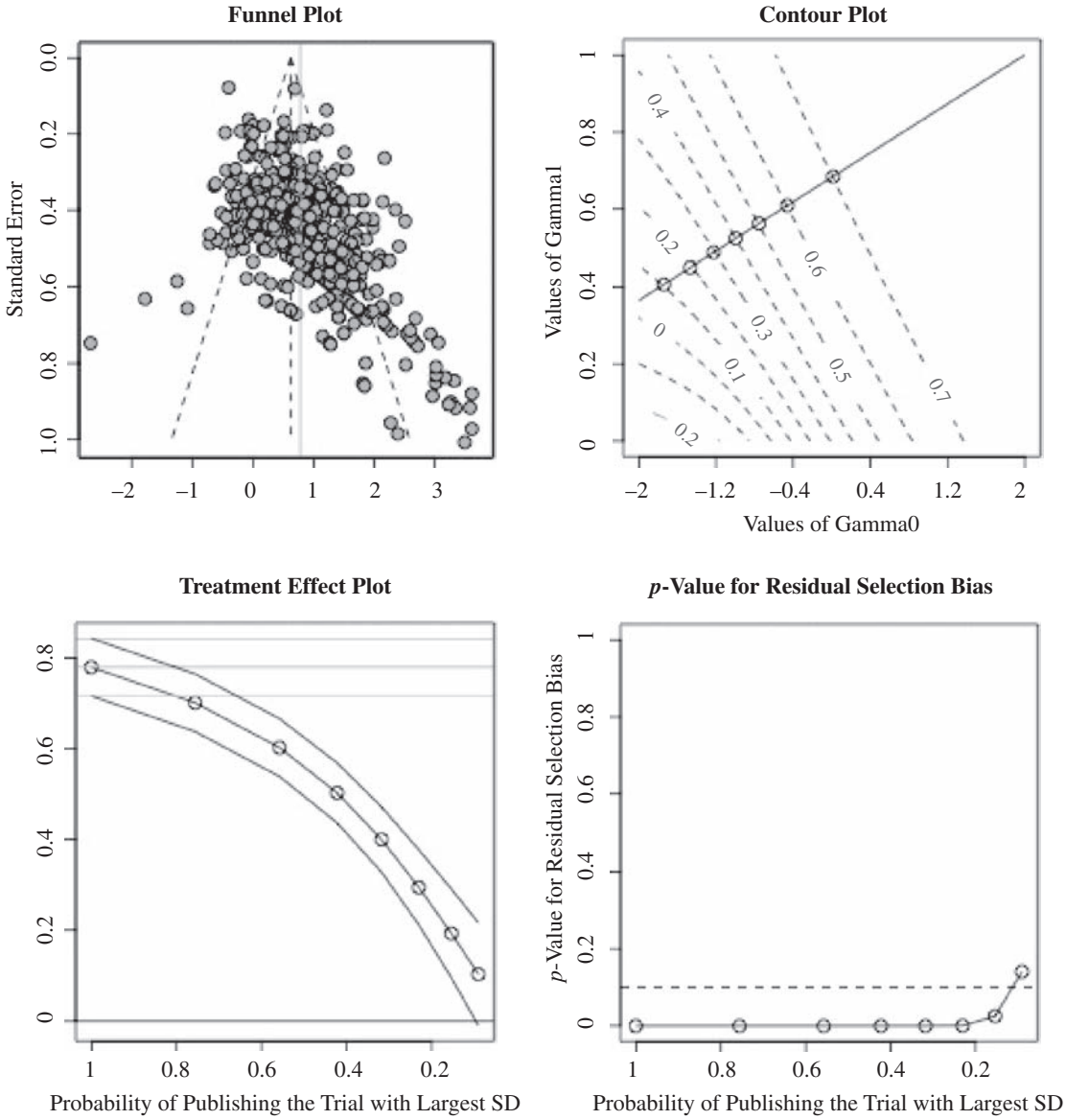

Figure 18.11 Copas and Shi, Psychotherapy Data, $a = -2$ to 2, $b = 0$ to 1

SOURCE: Author's tabulation.
NOTE: Gamma zero and gamma one represent a and b, respectively.

18.6.1.2.1 Funnel Plots. We calculated the amount necessary to expand the range of the *x* and *y* axes by 5 percent. Then we created a space for the funnel plot with

```
plot(c(min(sqrt(ibs_v))-.0193, max(sqrt
(ibs_v))+0.0193),c(min(ibs_y)-.1057,
max(ibs_y) + 0.1057),type='n', xlab=
"Standard Error",ylab="Effect Size")
```

We added the scatterplot points with

```
points(sqrt(ibs_v),ibs_y)
```

The funnel plot for the irritable bowel syndrome data set is featured in figure 18.12. Again, we computed the mean effect size for this data set using the *metafor* package's *rma()* function:

```
rma(ibs_y, ibs_v, method='ML')
```

We used a random-effects model and maximum likelihood estimation. This yielded a mean effect size of 0.42 and a variance component of 0.15 (see table 18.3), which corresponds to a large I^2 (72.90 percent). There are only

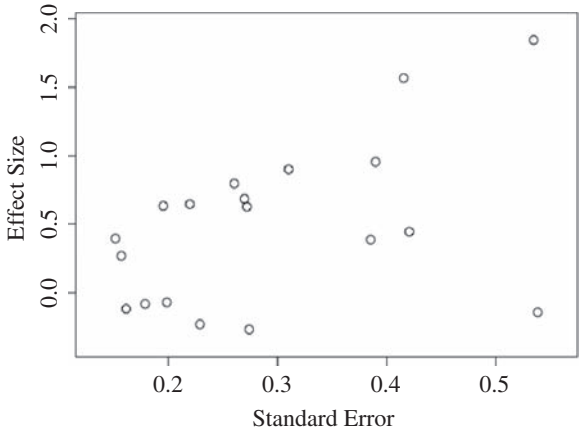

Figure 18.12 Effect Size Against Standard Error, Irritable Bowel Syndrome Data

SOURCE: Author's tabulation.

Table 18.3 Summary of Results for Irritable Bowel Syndrome Data

Method		Overall Mean	τ^2
Unadjusted		0.42	0.15
Trim and fill	left, L0	0.42 (0%, A) *MI*	0.15 (0%, A) *MI*
	right, L0	0.42 (0%, A) *MI*	0.15 (0%, A) *MI*
	left, R0	0.38 (9.09%, A) *MI*	0.18 (20.26%, M) *MO*
	right, R0	0.418 (0%, A) *MI*	0.15 (0%, A) *MI*
PET-PEESE		**–0.23 (155%, S), *S***	—
Vevea and Hedges	$p = 0.025$	0.16 (61.72%, S), *MO*	0.09 (41.18%, S), *MO*
	multiple	—	—
Vevea and Woods	moderate one-tailed	0.32 (24.64%, M) *MO*	0.17 (10.46%, A) *MI*
	severe one-tailed	**0.11 (74.64%, S) *S***	0.23 (49.67%, S) *MO*
	moderate two-tailed	0.37 (11.00%, A) *MI*	0.14 (11.11%, A) *MI*
	severe two-tailed	0.32 (24.16%, M) *MO*	0.11 (26.14%, M) *MO*
Copas and Shi		0.25 (40.19%, S) *MO*	—
Rücker		**0.09 (78.47%, S) *S***	—
p-uniform		0.64 (53.11%, S) *MO*	—

SOURCE: Author's tabulation.

NOTES: Adjusted estimates are reported unless row is labeled "Unadjusted." Percent adjustment is in parentheses, followed by the Kepes, Banks, and Oh (2012) categorization (A for Absent, or < 20% adjustment; M for Moderate, or adjustment between 20% and 40%; S for Severe, or adjustment > 40%). The Rothstein, Sutton, and Borenstein (2005) categorization follows in italics (MI for Minimal, or adjustment is similar; MO for Moderate, or adjustment is substantial, but key finding remains; S for Severe, adjustment that calls the key finding into question). If both categorizations were "Severe," the cell is bolded. Cells with "—" either were not or could not be estimated.

nineteen effect sizes here, so this funnel plot is more difficult to assess. There does appear to be greater density at the top of the funnel than at the bottom, and several large effect sizes (> 0.90) with large standard errors are not mirrored by an equivalent number of smaller effects. Despite the small size of the data set, there are enough signs of asymmetry that bias may still be a concern.

18.6.1.2.2 Cumulative Meta-Analysis. We took the object containing the results of the random-effects meta-analysis, ibs_rma, and created a cumulative meta-analysis:

```
ibs_cumul <- cumul(ibs_rma, order =
order(sqrt(ibs_v)))
```

Then we made a forest plot of the cumulative meta-analysis:

```
forest(ibs_cumul)
```

The forest plot for the irritable bowel syndrome data set appears in figure 18.13. The vertical dashed line represents an effect size—here, log(*RR*)—of 0.00. The least precise study is study 19, with log(*RR*) = 0.40. As more and more precise studies are added, the average effect size drifts to the left a bit, eventually going as far as log(*RR*) = 0.13. However, by the time the most precise studies are added, the average effect size has arrived back where it began, at log(*RR*) = 0.42. This is a very small drift, of about 0.02—not exactly indicative of a relationship between study size and effect size. However, the pattern is unusual. After the first few lines of the plot, the drift is consistently toward larger effects as studies with greater precision are added to the analysis. That would be consistent with publication in the unexpected direction.

18.6.1.2.3 Trim and Fill. Again, we estimated four trim and fill models. The first three were variations of

```
trimfill(ibs_rma, side="left",
estimator="L0")
```

where "L0" was exchanged for "R0." The second two were the same variations of

```
trimfill(ibs_rma, side="right",
estimator="L0")
```

The results of the trim and fill analyses for the irritable bowel syndrome data set are presented in table 18.3. This time, only the R_0 estimator added any additional effects. It imputed one effect on the left side of the funnel plot, reducing the mean from 0.42 to 0.38 (an attenuation of 9.52 percent) and increasing the variance component

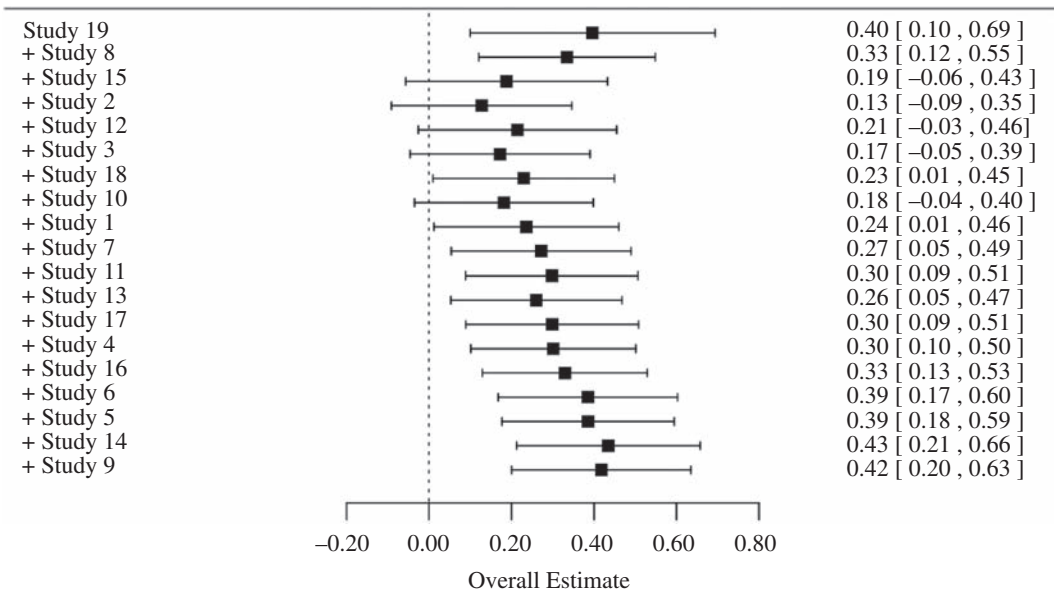

Figure 18.13 Cumulative Meta-Analysis, Irritable Bowel Syndrome Data

SOURCE: Author's tabulation.

from 0.15 to 0.18 (a change of 120 percent). The average effect size has barely been reduced. The trim and fill results for this data set indicate that the irritable bowel data set is very robust to the effects of publication bias.

18.6.1.2.4 Egger's Regression. We estimated a standard Egger's regression of effect size on standard error using weighted regression with multiplicative dispersion

```
regtest(ibs_rma, model="lm")
```

resulted in a test for funnel plot asymmetry on the intercept that was significant, $t(17) = 2.22$, $p = .04$. This indicates that a relationship may exist between study size and effect size.

We also estimated a variation of Egger's regression that predicts effect size with standard error using a random-effects meta-regression model

```
regtest(ibs_rma)
```

A test on the intercept of this model was also significant, $z = 2.64$, $p = .01$. We can reject the null hypothesis and conclude that there may be some evidence of bias.

18.6.1.2.5 PET-PEESE. We estimated PET:

```
pet <- lm(ibs_y ~ sqrt(ibs_v), weights =
1/ibs_v)
```

followed by PEESE:

```
peese <- lm(ibs_y ~ ibs_v, weights = 1/ibs_v)
```

We stored the estimates from these regressions and kept the PET estimates if PET was nonsignificant; otherwise, we kept PEESE.

PET was nonsignificant, indicating the absence of evidential value ($p = .39$). This means that estimating PEESE is not necessary, and the adjusted estimate of effect size from PET, $d = –0.23$ (table 18.3) is due to publication bias. the adjustment is an attenuation of about 156 percent.

18.6.1.2.6 Nonparametric Rank Correlation. The rank correlation returned by

```
ranktest(ibs_rma)
```

was nonsignificant, with Kendall's tau = 0.26 ($p = .13$). This indicates a lack of any significant correlation between effect size and sampling variance, or that no significant evidence of publication bias exists.

18.6.1.2.7 p-*Curve and* p-*Uniform.* The graph *p*-curve produced is presented in figure 18.14. This distribution of *p*-values is also visibly right skewed, and the tests that *p*-curve conducts agree. All three tests for right skew—the binomial ($p = .001$), the continuous full *p*-curve ($z = –4.47$, $p < .0001$), and the continuous half *p*-curve ($z = –2.87$, $p = .00$)—were significant, indicating that the data set contains evidential value.

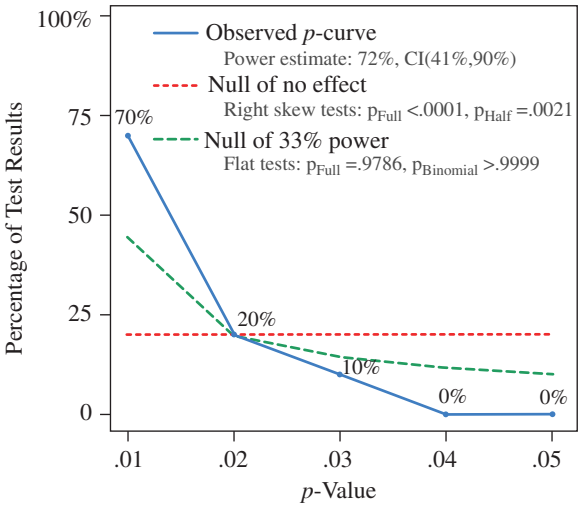

Figure 18.14 *p*-Curve, Irritable Bowel Syndrome Data

SOURCE: Authors' tabulation obtained directly from the *p*-curve application (v. 4.06), available at http://www.p-curve.com/app4/ (accessed January 7, 2019).

NOTE: The observed *p*-curve includes 10 statistically significant ($p<.05$) results, of which 10 are $p<.0.5$. There were 9 additional results entered but excluded from *p*-curve because they were $p>.0.5$.

The binomial test ($p > .999$) and continuous test ($p = .98$) assessing whether the studies are underpowered were both nonsignificant. This indicates that the studies are not underpowered, which makes sense given that right skew is present (a sign of evidential value).

We estimated *p*-uniform:

```
puniform(yi = ibs_y, vi = ibs_v, alpha =
0.05, side="right", method="P", plot=TRUE)
```

The plot of observed versus expected *p*-values for the irritable bowel syndrome data set is in figure 18.15.

This data set does not include many significant *p*-values. (Recall that *p*-uniform analyses only significant *p*-values.) The pattern of deviations from uniformity appears similar to that from the psychotherapy effectiveness data—deviations occur more at the very small and very large parts of the *x*-axis.

The one-tailed test for publication bias was nonsignificant, with $z = –2.61$ and $p = .99$. The adjusted fixed-effects estimate was $d = 0.64$ (0.41, 0.90), $p < .001$ (see

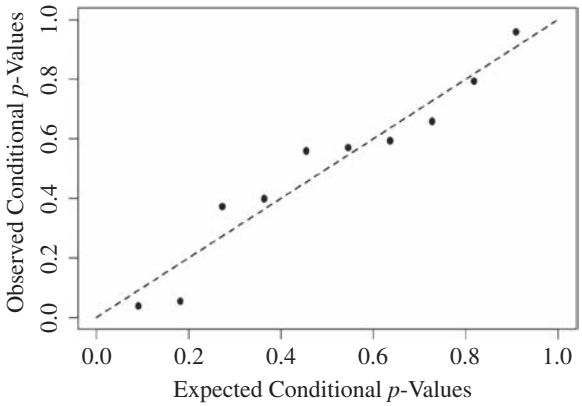

Figure 18.15 *p*-Uniform, Irritable Bowel Syndrome Data

SOURCE: Author's tabulation.

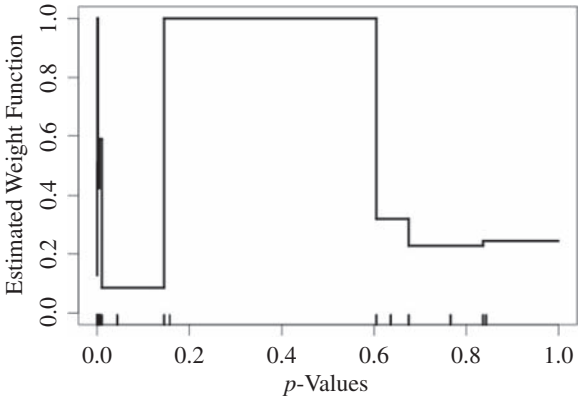

Figure 18.16 Dear and Begg, Irritable Bowel Syndrome Data

SOURCE: Authors' tabulation obtained directly via the *R* package *selectMeta* (v. 1.0.8).

table 18.3). *p*-uniform indicates that publication bias is not a threat for this data set.

18.6.1.2.8 Excess Significance Test. For the irritable bowel syndrome data, the product of the power for each significant effect size was also 0.00. Publication bias may be present.

18.6.1.2.9 Dear and Begg. The Dear and Begg (1992) weight-function model can be implemented in *R* using the *R* package *selectMeta* and its function *DearBegg()*. Meta-analysts must first run *DearBegg()* on their effect sizes and sampling variances, then use the examples in the package manual to create a plot of the resulting weight function. *DearBegg()* itself produces matrices of all the weight estimates for *p*-value intervals, and these must be plotted to be meaningful.

We estimated the Dear and Begg weight-function model:

```
ibs_db <- DearBegg(ibs_y, sqrt(ibs_v),
trace=FALSE)
```

The plot of the Dear and Begg model is presented in figure 18.16.

The weight function has a spike at the far left of the plot, near $p = 0$, indicating that effect sizes with *p*-values in that range are more likely to be observed (a sign of publication bias). However, for observed *p*-values in the range from about .18 to .60, the probability of surviving selection is also high, indicating that nonsignificant studies in that range are also likely to be observed. The spike at the far left, therefore, may not be a matter for concern. It is difficult to determine whether publication bias is a threat based on the plot.

Rufibach's modification can be implemented using *selectMeta*, the same *R* package. The function for Rufibach's modification is *DearBeggMonotone()*, and the results of this function also must be plotted to be meaningful. *R* code to construct plots is featured in the package manual.

We estimated the Rufibach (2011) weight-function model:

```
ibs_db <- DearBeggMonotone(ibs_y,
sqrt(ibs_v), trace=FALSE)
```

The plot of the Rufibach weight-function model is featured in figure 18.17. This weight function indicates that all but the most significant studies have a low probability of surviving publication, and as *p*-value increases the likelihood of surviving decreases even further. For studies with a *p*-value near 1.00, this probability is close to .00. The plot does indicate that publication bias is a concern. The difference between this plot and the one observed for the Dear and Begg method is due to the constraint that the Rufibach model imposes on the weight function—the required monotonicity suppresses the higher weights for the 0.18 to 0.60 *p*-value range.

18.6.1.2.10 Vevea and Hedges. We refresh readers' memory that the unadjusted mean for the irritable bowel syndrome data set is $\log(RR) = 0.42$ and the unadjusted variance component is 0.15.

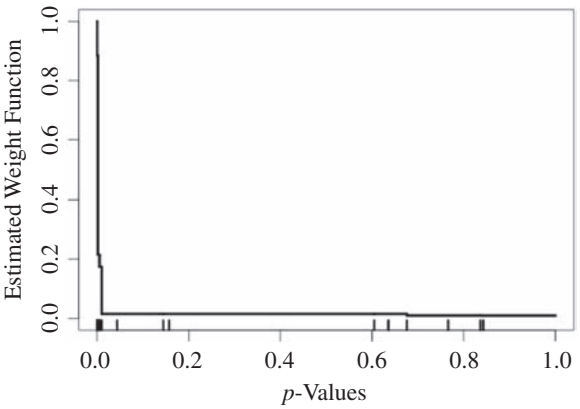

Figure 18.17 Rufibach, Irritable Bowel Syndrome Data

SOURCE: Authors' tabulation obtained directly via the *R* package *selectMeta* (v. 1.0.8).

We begin by specifying one *p*-value cut point, at $p = .05$:

```
weightfunct(ibs_y, ibs_v)
```

The weight for the interval $p < .05$ is, of course, fixed to one. The weight for the interval $0.05 < p < 1.00$ is estimated at 0.19, indicating that nonsignificant studies are 19 percent as likely to survive selection as significant ones. The mean effect size is adjusted downward to 0.16, an attenuation of about 50 percent, and the variance component is also adjusted downward to 0.09, an attenuation of 20 percent (see table 18.3).

Next, we attempt to specify the more detailed one-tailed pattern of *p*-value cut points, but we cannot estimate it—several intervals have no effect sizes. Based on the model that distinguished only between significant and nonsignificant studies, however, the mean effect size was reduced by half. It did not drop below zero, but a reduction this large likely indicates that the data set is not robust to publication bias.

18.6.1.2.11 Vevea and Woods. The results for the irritable bowel syndrome data set are presented in table 18.3. The code we used consists of the same variations that we used for the psychotherapy data. Again, we replaced the weights vector with the corresponding set of weights for each selection pattern. Because weights are fixed, it is irrelevant whether *p*-values actually fall in every interval.

None of the selection bias patterns adjusted the original unadjusted effect size upward. The furthest downward that it is attenuated happens under severe one-tailed selection, where the adjusted effect size reaches 0.11 from its unadjusted 0.42. Despite this reduction, these results are encouraging. Even under the most severe one-tailed bias pattern we created, the average effect size did not become negative or too near zero; there still appears to be a positive effect. This data set does appear to be robust to the effects of publication bias.

18.6.1.2.12 Copas and Shi. We first estimated a random-effects meta-analysis with maximum likelihood:

```
ibs_meta <- metagen(TE = ibs_y, seTE =
sqrt(ibs_v), method.tau = "ML")
```

Then we estimated the Copas and Shi selection model:

```
cop.ibs <- copas(ibs_meta)
plot(cop.ibs)
summary(cop.ibs)
```

The four plots produced by the Copas and Shi (2001) selection model *R* function (Carpenter et al. 2009) are shown in figure 18.18.

The contour plot indicates that the data set is not terribly robust to the effects of selection bias; most of the contour lines are closer together. It also indicates that the model had some difficulty converging, as some of the contour lines curve. The top right estimate, under no selection bias, is about $\log(RR) = 0.40$.

The treatment effect plot shows that, as the probability of publishing the study with the largest standard error decreases, the estimated average effect size decreases as well, moving from about $\log(RR) = 0.40$ under no selection bias to about $\log(RR) = 0.10$ in a situation where studies with the largest standard error are published only about 35 percent of the time.

Finally, the *p*-value plot shows that residual selection bias becomes nonsignificant when the least precise studies are published about 73 percent of the time. This situation includes publication bias, but is far from the extreme of the psychotherapy data set. The Copas and Shi model yields an adjusted estimate in this situation of $\log(RR) = 0.25$, an attenuation of 40 percent (2001; see table 18.3).

18.6.1.2.13 Rücker Limit Meta-Analysis. We estimated the Rücker limit meta-analysis method:

```
ibs_limit <- limitmeta(ibs_meta)
summary(ibs_limit)
```

The results indicate a significant relationship between effect size and standard error. The test for small-study effects was significant, $Q(1) = 14.03$ ($p = .0002$), as was the test for residual heterogeneity, $Q(17) = 48.37$, $p < .0001$.

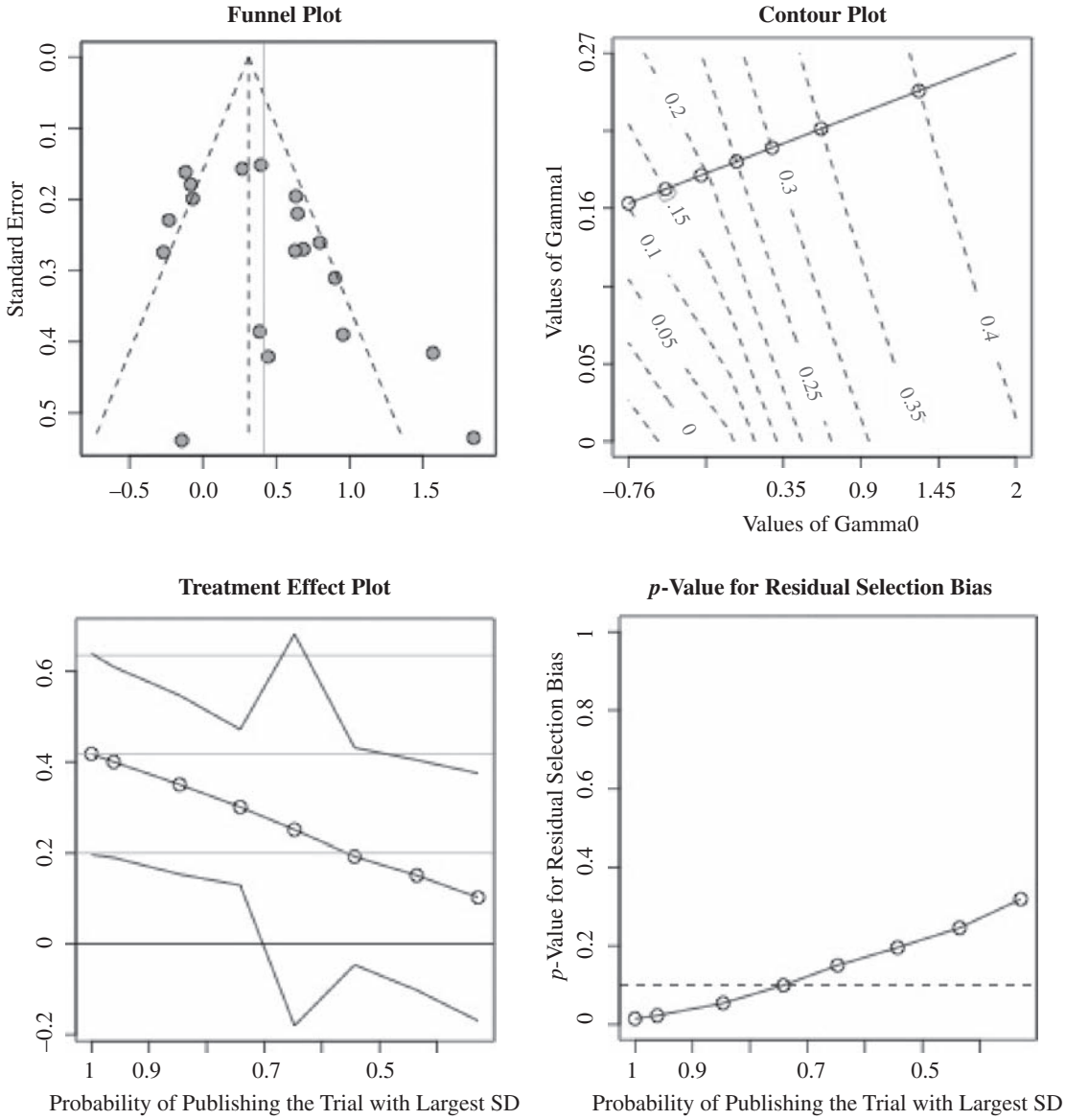

Figure 18.18 Copas and Shi, Irritable Bowel Syndrome Data, $a = -0.76$ to 2, $b = 0$ to 0.27

SOURCE: Author's tabulation.
NOTE: Gamma zero and gamma one represent a and b, respectively.

The adjusted random-effects estimate for the mean effect is 0.09 (see table 18.3), with a 95 percent confidence interval from –0.22 to 0.39. This is a downward adjustment of 78.57 percent.

18.7 DISCUSSION

This chapter provides an overview of the problem of publication bias with a focus on the methods developed to address it in a research-synthesis context. Publication bias is a difficult problem. The mechanisms causing bias are unknown, and the merit of any method to address it depends on the truth of any assumptions the method makes. Therefore, all methods should be viewed as sensitivity analyses, and triangulation across multiple approaches is advisable. Because of the different assumptions, users should not expect triangulation to lead to a consensus across methods and must exercise judgment that includes an assessment of the plausibility of the assumptions. Banks, Kepes, and McDaniel (2012) and Kepes and his colleagues (2012) discuss this issue.

Meta-analysts' toolboxes are now likely to be full of publication bias assessments. The sheer number of possible methods can engender some confusion. As the earlier examples demonstrate, results for a single data set may vary widely across methods, due to the methods' differing assumptions and strategies. Faced with such a range, what should the meta-analyst conclude?

Although all publication bias assessments are sensitivity analyses, some assessments are less sensitive than others. For instance, some methods (such as weight-function methods) are robust to violations of their assumptions; others are not. When conducting a meta-analysis, if the data set does not meet the assumptions of a particular bias assessment, and if that assessment is not robust, the researcher should be less willing to trust its results. On the other hand, if the data set does meet the assumptions, or if the assessment is robust to violations, the researcher should place more credence in its results. This chapter does not provide a specific list of methods that should or should not be included as part of the triangulation process. Instead, the researcher should observe the data set, estimate a range of methods, and assess the body of results, bearing in mind that some results may be more meaningful than others.

Table 18.2 compares the adjusted effect estimates for the psychotherapy data set. One of the more conservative methods (PET-PEESE) gives an adjusted effect size as small as –0.04. At the other end of the spectrum, *p*-uniform adjusts the estimate to 1.06. Other methods (trim and fill and Vevea and Hedges's weight-function model) give estimates that seem more consistent with the funnel plot, ranging from 0.47 to 0.78. The Vevea and Woods model suggests that the data set is robust to the effects of different selection patterns (2005). In no case are the key findings reversed or called into question. The conditional means for complex phobias are attenuated more than simple phobias, due to the smaller magnitude of effects for those groups, which makes them more likely to be affected by weights in the nonsignificant *p*-value ranges. Overall, across all the methods presented in table 18.2, it is plausible that some degree of publication bias is present in this data set. However, with the exception of those methods that can accommodate neither systematic nor random heterogeneity, the key finding is never reversed or called into question.

Table 18.3 compares the adjusted estimates for the irritable bowel syndrome data. PET-PEESE yields a result so extreme that the adjusted finding (–0.23) suggests the treatment is harmful. Once again, *p*-uniform inflates the adjusted effect, likely due to the fact that it disregards nonsignificant effect sizes (in this case, ignoring 50 percent of an already small data set). The Vevea and Woods results yield minimal to moderate adjustment except in the most severe one-tailed scenario (2005). Rücker's method is conservative, producing an estimate similar to the most extreme case of the Vevea and Woods method. The other selection models (Vevea and Hedges, Copas and Shi) reduce the effect dramatically, into the range that Kepes (CITE) define as "severe." It does appear, then, that the true effect may be substantially smaller than estimated in the meta-analysis.

The danger of reliance on a single approach is clear. A responsible analyst here would most likely discount the most extreme results and conclude that, although bias may be a problem, it is not likely to be the primary reason that a positive effect was found. A good sense of what various approaches can and cannot achieve is useful for this triangulation process. Table 18.4 summarizes the characteristics of many methods.

Freely available software has been developed that implements most of the methods described here, with the exception of the Bayesian approaches and methods for outcome, subgroup, and time-lag biases. The new tendency among developers of methods is to make them accessible as open-source packages. *R* (R Core Team 2016) packages implement various models previously inaccessible to typical users. Examples include Wolfgang

Table 18.4 Characteristics of Various Methods

Method	Subjective Interpretation	Tests for Bias	Primarily Visual	Software Available	Linear Model	Adjusted Effect Size(s)	Homogeneity Necessary	Adjusted Variance Component	Based on Relationship Between Study Size and Effect Size	Based on *p*-Values
Funnel plot	X		X	X			X		X	
Cumulative meta-analysis	X		X	X			X		X	
Egger's regression		X		X	X[1]	X	X		X	
Rank correlation		X		X			X		X	
Trim and fill		X		X	X[1]	X	X	X	X	
PET-PEESE		X			X[1]	X	X		X	
Vevea and Hedges		X		X	X	X		X		X
Vevea and Woods				X	X	X		X		X
Dear and Begg	X		X	X						X
Rufibach	X		X	X						X
Copas and Shi			X	X	X[1]	X	X		X	
Limit meta-analysis		X		X	X[1]	X	X		X	
p-curve		X	X	X		X[2]	X			X
p-uniform		X		X		X	X			X
Excess significance test		X					X			X

SOURCE: Author's tabulation.
[1] indicates that the method has the potential to incorporate a linear model, but that software is not readily available to do so.
[2] indicates that the method can yield an adjusted effect size, but software is not readily available to do so.

Viechtbauer's *metafor* (2010), Guido Schwarzer's *meta* package (2016), *metasens* by Guido Schwarzer and his colleagues (2016), and Coburn and Vevea's *weightr* (2016a). Others have made their approaches available through web interfaces. Examples include Coburn and Vevea's Shiny application (2016b), and Uri Simonsohn, Lief Nelson, and Joseph Simmons's web application for *p*-curve (2014).

Future investigation may prove fruitful in several directions. One example is development of methods that simultaneously account for different possible sources of bias (for example, *p*-value as well as magnitude and direction of individual effect estimates). Further development of models that allow various selection patterns for different study designs would be useful (for example, Sutton, Abrams, and Jones 2002). Extension of that idea to account for study characteristics that are not design related (for example, funding source, social preferences, or time) is a developing area (see, for example, Coburn and Vevea 2016b). Dan Jackson points out that little is yet known about the effects of publication bias on the between-studies variance component (2006). Derrick Bennett and his colleagues investigated capture-recapture methods across electronic databases to estimate the number of missing studies, but evidence of further research on that approach is scant (2004). Kepes and McDaniel (2015) mention the need for development of methods that are suitable for psychometric meta-analysis.

Bayesian methods are likely to provide valuable new insights on the publication bias problem. Promising directions could include incorporating Bayesian model averaging or Bayes factors. Bayesian methods also are likely to lead to models that address publication bias for statistical approaches that are more complicated than a standard meta-analysis, such as network meta-analyses.

Publication bias is a pervasive problem in the research literature, and meta-analysis provides a valuable opportunity to assess its impact. This chapter discusses and illustrates a variety of methods that aid in this process. The issue of addressing more nuanced questions about publication is a rapidly developing field, and one that is likely to prove fruitful in the next few years.

18.8 REFERENCES

Abramowitz, Stephen I., Beverly Gomes, and Christine V. Abramowitz. 1975. "Publish or Politic: Referee Bias in Manuscript Review." *Journal of Applied Social Psychology* 5(3): 187–200.

Abrams, Keith R., Clare L. Gillies, and Paul C. Lambert. 2005. "Meta-Analysis of Heterogeneously Reported Trials Assessing Change from Baseline." *Statistics in Medicine* 24(24): 3823–44.

American Psychological Association. 2008. "Reporting Standards for Research in Psychology: Why Do We Need Them? What Might They Be?" *American Psychologist* 63(9): 839–851.

Anderson, Richard. 2013. "Registration and Replication: A Comment." *Political Analysis* 21(1): 38–39.

Balcetis, Emily, and David Dunning. 2012. "A False-Positive Error in Search of Selective Reporting." *i-Perception* 3(3).

Banks, George C., Sven Kepes, and Michael A. McDaniel. 2012. "Publication Bias: A Call for Improved Meta-Analytic Practice in the Organizational Sciences." *International Journal of Selection and Assessment* 20(2): 182–96.

Barnes, Deborah E., and Lisa A. Bero. 1998. "Why Review Articles on the Health Effects of Passive Smoking Reach Different Conclusions." *Journal of the American Medical Association* 279(19): 1566–570.

Bayarri, M. J., and Morris H. DeGroot. 1987. "Bayes Analysis of Selection Models." *Journal of the Royal Statistical Society, Series D (The Statistician)* 36(2/3): 137–46.

Becker, Betsy J. 2005. "Failsafe *N* or File-Drawer Number." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester: John Wiley & Sons.

Begg, Colin B., and Jesse A. Berlin. 1988. "Publication Bias: A Problem in Interpreting Medical Data (with Discussion)." *Journal of the Royal Statistical Society Series A* 151(3): 419–63.

Begg, Colin B., and Madhuchhanda Mazumdar. 1994. "Operating Characteristics of a Rank Correlation Test for Publication Bias." *Biometrics* 50(4): 1088–101.

Bekelman, Justin E., Yan Li, and Cary P. Gross. 2003. "Scope and Impact of Financial Conflicts of Interest in Biomedical Research: A Systematic Review." *Journal of the American Medical Association* 289(4): 454–65.

Bennett, Derrick A., Nancy K. Latham, Caroline Stretton, and Craig S. Anderson. 2004. "Capture-Recapture Is a Potentially Useful Method for Assessing Publication Bias." *Journal of Clinical Epidemiology* 57(4): 349–57.

Berlin, Jesse A., and Davina Ghersi. 2005. "Preventing Publication Bias: Registries and Prospective Meta-Analysis." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Berlin, Jesse A., and Robert M. Golub. 2014. "Meta-Analysis as Evidence: Building a Better Pyramid." *Journal of the American Medical Association* 312(6): 603–06.

Bishop, Dorothy V. M., and Paul A. Thompson. 2016. "Problems in Using P-Curve Analysis and Text-Mining to Detect Rate of P-Hacking and Evidential Value." *PeerJ* 4: e1715.

Bondas, Terese, and Elisabeth O. C. Hall. 2016. "Challenges in Approaching Metasynthesis Research." *Qualitative Health Research* 17(1): 113–21.

Borenstein, Michael. 2005. "Software for Publication Bias." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Bruns, Stephan B., and John P. A. Ioannidis. 2016. "*p*-Curve and *p*-Hacking in Observational Research." *PloS One* 11(2): e0149144.

Carpenter, James R., Guido Schwarzer, Gerta Rücker, and Rita Künstler. 2009. "Empirical Evaluation Showed That the Copas Selection Model Provided a Useful Summary in 80 Percent of Meta-Analyses." *Journal of Clinical Epidemiology* 62(6): 624–31.

Carter, Evan C., Lilly M. Kofler, Daniel E. Forster, and Michael E. McCullough. 2015. "A Series of Meta-Analytic Tests of the Depletion Effect: Self-Control Does Not Seem to Rely on a Limited Resource." *Journal of Experimental Psychology* 144(4): 796–815.

Carter, Evan C., and Michael E. McCullough. 2014. "Publication Bias and the Limited Strength Model of Self-Control: Has the Evidence for Ego Depletion Been Overestimated?" *Frontiers in Psychology* 5: 819.

Ceci, Stephen J., Douglas Peters, and Jonathan Plotkin. 1985. "Human Subjects Review, Personal Values, and the Regulation of Social Science Research." *American Psychologist* 40(9): 994.

Chan, An-Wen, and Douglas G. Altman. 2005. "Identifying Outcome Reporting Bias in Randomised Trials on PubMed: Review of Publications and Survey of Authors." *British Medical Journal* 330(7494): 753.

Chan, An-Wen, Asbjorn Hrobjartsson, Mette T. Haahr, Peter C. Gøtzsche, and Douglas G. Altman. 2004. "Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials. Comparison of Protocols to Published Articles." *Journal of the American Medical Association.* 291(20): 2457–65.

Chan, An-Wen, Karmela Krleza-Jeric, Isabelle Schmid, and Douglas G. Altman. 2004b. "Outcome Reporting Bias in Randomized Trials Funded by the Canadian Institutes of Health Research." *Canadian Medical Association Journal* 171(7): 735–40.

Clarke, Mike, and Lesley Stewart. 1998. "Time Lag Bias in Publishing Clinical Trials." *Journal of the American Medical Association* 279(24): 1952–53.

Coburn, Kathleen, and Jack L. Vevea. 2015. "Publication Bias as a Function of Study Characteristics." *Psychological Methods* 20(3): 310.

———. 2016a. "weightr: Estimating Weight-Function Models for Publication Bias in *R*" (1.0.0). *R* package.

———. 2016b. "The Vevea and Hedges Weight-Function Model for Publication Bias." Computer software. (1.0.0). Accessed December 14, 2018. https://vevealab.shinyapps.io/WeightFunctionModel.

Cooper, Harris M., Kristina M. DeNeve, and Kelly Charlton. 1997. "Finding the Missing Science: The Fate of Studies Submitted for Review by a Human Subjects Committee." *Psychological Method.* 2(4): 447–52.

Copas, John B., and Hu G. Li. 1997. "Inference for Non-Random Samples." *Journal of the Royal Statistical Society, Series B* 59(1): 55–95.

Copas, John, and Jian Qing Shi. 2000. "Meta-Analysis, Funnel Plots and Sensitivity Analysis." *Biostatistics* 1(3): 247–62.

———. 2001. "A Sensitivity Analysis for Publication Bias in Systematic Reviews." *Statistical Methods in Medical Research* 10(4): 251–65.

Coursol, Allan, and Edwin E. Wagner. 1986. "Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias." *Professional Psychology: Research and Practice* 17(2): 136–37.

Dear, Keith B. G., and Colin B. Begg. 1992. "An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis." *Statistical Science* 7(2): 237–45.

Deeks, Jonathan J., Petra Macaskill, and Les Irwig. 2005. "The Performance of Tests of Publication Bias and Other Sample Size Effects in Systematic Reviews of Diagnostic Test Accuracy Was Assessed." *Journal of Clinical Epidemiology* 58(9): 882–93.

Dickersin, Kay. 2005. "Publication Bias: Recognizing the Problem, Understanding its Origins and Scope, and Preventing Harm." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Dickersin, Kay, Yi Min, and Curtis L. Meinert. 1991. "The Fate of Controlled Trials Funded by the NIH in 1979." *Controlled Clinical Trials* 12: 634.

———. 1992. "Factors Influencing Publication of Research Results: Follow-Up of Applications Submitted to Two Institutional Review Boards." *Journal of the American Medical Association* 267(3): 374–78.

Dorn, Spencer D., Ted J. Kaptchuk, Jae Berm Park, Long Thanh Nguyen, Katia M. Canenguez, Bong Hyun Nam, Ko Bo Woods, Lisa A. Conboy, William B. Stason, and Anthony J. Lembo. 2007. "A Meta-Analysis of the Placebo Response in Complementary and Alternative Medicine Trials of Irritable Bowel Syndrome." *Neurogastroenterology and Motility* 19(8): 630–37.

Duarte, José L., Jarret T. Crawford, Charlotta Stern, Jonathan Haidt, Lee Jussim, and Philip E. Tetlock. 2015. "Political Diversity Will Improve Social Psychological Science." *Behavioral and Brain Sciences* 38: e130. DOI: 10.1017/S0140525X14000430.

Duval, Sue. 2005. "The Trim and Fill Method." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Duval, Sue, and Richard Tweedie. 1998. "Practical Estimates of the Effect of Publication Bias in Meta-Analysis." *Australian Epidemiologist.* 5(4): 14–17.

———. 2000a. "A Non-Parametric 'Trim and Fill' Method of Accounting for Publication Bias in Meta-Analysis." *Journal of the American Statistical Association* 95(449): 89–98.

———. 2000b. "Trim and Fill: A Simple Funnel Plot Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis." *Biometrics* 56(2): 455–63.

Easterbrook, Phillipa J., Ramana Gopalan, J. A. Berlin, and David R. Matthews. 1991. "Publication Bias in Clinical Research." *The Lancet* 337(8746): 867–72.

Egger, Matthias, and G. Davey Smith. 1998. "Bias in Location and Selection of Studies." *British Medical Journal* 316(7124): 61–66.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-Analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315(7109): 629–34.

Ferguson, Christopher J., and Michael T. Brannick. 2012. "Publication Bias in Psychological Science: Prevalence, Methods for Identifying and Controlling, and Implications for the Use of Meta-Analyses." *Psychological Methods* 17(1): 120–28.

Fisher, Ronald A. 1932. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Francis, Gregory. 2012a. "Too Good to Be True: Publication Bias in Two Prominent Studies from Experimental Psychology." *Psychonomic Bulletin & Review* 19(2): 151–56.

———. 2012b. "The Same Old New Look: Publication Bias in a Study of Wishful Seeing." *i-Perception* 3(3): 176–78.

———. 2012c. "Response to Author: Some Clarity About Publication Bias and Wishful Seeing." *i-Perception* 3. DOI: 10.1068/i0519ic.

———. 2012d. "Evidence That Publication Bias Contaminated Studies Relating Social Class and Unethical Behavior." *Proceedings of the National Academy of Sciences* 109(25): E1587.

———. 2012e. "Checking the Counterarguments Confirms That Publication Bias Contaminated Studies Relating Social Class and Unethical Behavior." Accessed December 14, 2018. http://www3.psych.purdue.edu/~gfrancis/Publications/FrancisRebuttal2012.pdf.

———. 2012f. "Publication Bias and the Failure of Replication in Experimental Psychology." *Psychonomic Bulletin & Review* 19(6): 975–91. DOI: 10.3758/s13423-012-0322-y.

———. 2012g. "The Psychology of Replication and Replication in Psychology." *Perspectives on Psychological Science* 7(6): 580–89. DOI: 10.1177/1745691612459520.

———. 2014. "The Frequency of Excess Success for Articles in *Psychological Science*." *Psychonomic Bulletin and Review* 21(5): 1180–87.

Galak, Jeff, and Tom Meyvis. 2012."You Could Have Just Asked: Reply to Francis (2012)." *Perspectives on Psychological Science* 7(6): 595–96.

Galbraith, Rex F. 1994. "Some Applications of Radial Plots." *Journal of the American Statistical Association.* 89(428): 1232–42.

Gelman, Andrew. 2013. "Preregistration of Studies and Mock Reports." *Political Analysis* 21(1): 40–41.

———. 2015. "The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective." *Journal of Management* 41(2): 632–43.

Givens, Geof H., David D. Smith, and Richard L. Tweedie. 1997. "Publication Bias in Meta-Analysis: A Bayesian Data-Augmentation Approach to Account for Issues Exemplified in the Passive Smoking Debate." *Statistical Science* 12(4): 221–50.

Guan, Maime, and Joachim Vandekerckhove. 2016. "A Bayesian Approach to Mitigation of Publication Bias." *Psychonomic Bulletin & Review* 23(1): 74–86.

Hahn, Seokyung, Paula R. Williamson, and Jane L. Hutton. 2002. "Investigation of Within-Study Selective Reporting in Clinical Research: Follow-Up of Applications Submitted

to a Local Research Ethics Committee." *Journal of Evaluation in Clinical Practice* 8(3): 353–59.

Hahn, Seokyung, Paula R. Williamson, Jane L. Hutton, Paul Garner, and E. Victor Flynn. 2000. "Assessing the Potential for Bias in Meta-Analysis Due to Selective Reporting of Subgroup Analyses Within Studies." *Statistics in Medicine* 19(24): 3325–36.

Halpern, Scott D., and Jesse A. Berlin. 2005. "Beyond Conventional Publication Bias: Other Determinants of Data Suppression." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Harbord, Roger M., Matthias Egger, and Jonathan A. C. Sterne. 2006. "A Modified Test for Small-Study Effects in Meta-Analyses of Controlled Trials with Binary Endpoints." *Statistics in Medicine* 25(20): 3443–57.

Hedges, Larry V. 1984. "Estimation of Effect Size Under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences." *Journal of Educational Statistics* 9(1): 61–85.

———. 1992. "Modeling Publication Selection Effects in Meta-Analysis." *Statistical Science* 7(2): 246–55.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. San Diego, Calif.: Academic Press.

Hedges, Larry V., and Jack L. Vevea. 1996. "Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model." *Journal of Educational and Behavioral Statistics* 21(4): 299–332.

———. 2005. "Selection Method Approaches." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Higgins, Julian P. T., and Sally Green, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0 [updated March 2011]. London: The Cochrane Collaboration. Accessed December 14, 2018. http://handbook-5-1.cochrane.org.

Hopewell, Sally, Michael Clarke, and Sue Mallett. 2005. "Grey Literature and Systematic Reviews." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Hopewell, Sally, Mike J. Clarke, Lesley Stewart, and Jayne Tierney. 2007. "Time to Publication for Results of Clinical Trials." *Cochrane Database Systematic Reviews* 2 (April): MR000011.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis.* 21(1): 1–20.

Hunter, James P., Athanasios Saratzis, Alex J. Sutton, Rebecca H. Boucher, Robert D. Sayers, and Matthew J. Bown. 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.

Hunter, James P., and Frank L. Schmidt. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Newbury Park, Calif.: Sage Publications.

Hutton, Jane L., and Paula R. Williamson. 2000. "Bias in Meta-Analysis Due to Outcome Variable Selection Within Studies." *Applied Statistics*. 49(3): 359–70.

Ioannidis, John P., Evangelina E. Ntzani, Thomas A. Trikalinos, and Despina G. Contopoulos-Ioannidis. 2001. "Replication Validity of Genetic Association Studies." *Nature Genetics.* 29(3): 306–09.

Ioannidis, John P. A., and Thomas A. Trikalinos. 2007. "An Exploratory Test for an Excess of Significant Findings." *Clinical Trials* 4(3): 245–53.

Iyengar, Satish, and Joel. B. Greenhouse. 1988. "Selection Models and the File Drawer Problem." *Statistical Science* 3(1): 109–35.

Jackson, Daniel. 2006. "The Implication of Publication Bias for Meta-Analysis' Other Parameter." *Statistics in Medicine* 25(17): 2911–21.

Jackson, Daniel, John Copas, and Alexander Sutton. 2005. "Modelling Reporting Bias: The Operative Reporting Rate for Ruptured Abdominal Aortic Aneurysm Repair." *Journal of the Royal Statistical Society Series A* 168(4): 737–52.

Jadad, Alejandro R., and Drummond Rennie. 1998. "The Randomized Controlled Trial Gets a Middle-Aged Checkup." *Journal of the American Medical Association* 279(4): 319–20.

Jennions, Michael D., and Anders P. Moeller. 2002. "Publication Bias in Ecology and Evolution: An Empirical Assessment Using the 'Trim and Fill' Method." *Biological Reviews of the Cambridge Philosophical Society* 77(2): 211–22.

Jensen, Johan Ludwig William Valdemar. 1906. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes." *Acta Mathematica* 30(1): 175–93.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research

Practices with Incentives for Truth Telling." *Psychological Science* 23(5): 524–32.

Johnson, Valen E. 2013. "On Biases in Assessing Replicability, Statistical Consistency and Publication Bias." *Journal of Mathematical Psychology* 57(5): 177–79.

Joober, Ridha, Norbert Schmitz, Lawrence Annable, and Patricia Boksa. 2012. "Publication Bias: What Are the Challenges and Can They Be Overcome?" *Journal of Psychiatry and Neuroscience 37*(3): 149–52.

Kepes, Sven, George C. Banks, Michael McDaniel, and Deborah L. Whetzel. 2012. "Publication Bias in the Organizational Sciences." *Organizational Research Methods* 15(4): 624–62.

Kepes, Sven, George C. Banks, and In-Sue Oh. 2014. "Avoiding Bias in Publication Bias Research: The Value of 'Null' Findings." *Journal of Business and Psychology* 29(2): 183–203.

Kepes, Sven, Andrew A. Bennett, and Michael A. McDaniel. 2014. "Evidence-Based Management and the Trustworthiness of Our Cumulative Scientific Knowledge: Implications for Teaching, Research, and Practice." *Academy of Management Learning & Education* 13(3): 446–466.

Kepes, Sven, Brad J. Bushman, and Craig A. Anderson. 2017. "Violent Video Game Effects Remain a Societal Concern: Reply to Hilgard, Engelhardt, and Rouder (2017)." *Psychological Bulletin,* 143(7): 775–82.

Kepes, Sven, and Michael A. McDaniel. 2015. "The Validity of Conscientiousness Is Overestimated in the Prediction of Job Performance." *PLoS ONE* 10(10): e0141468. DOI: 10.1371/journal.pone.0141468.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr., Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, et al. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project." *Social Psychology* 45(3): 107–12.

Koricheva, Julia. 2003. "Non-Significant Results in Ecology: A Burden or a Blessing in Disguise?" *Oikos* 102(2): 397–401.

Lane, David M., and William P. Dunlap. 1978. "Estimating Effect-Size Bias Resulting from Significance Criterion in Editorial Decisions." *British Journal of Mathematical and Statistical Psychology* 31(2): 107–12.

Larose, Daniel T., and Dipak K. Dey. 1998. "Modeling Publication Bias Using Weighted Distributions in a Bayesian Framework." *Computational Statistics & Data Analysis* 26(3): 279–302.

Lau, Joseph, John P. A. Ioannidis, Norma Terrin, Christopher H. Schmid, and Ingram Olkin. 2006. "The Case of the Misleading Funnel Plot." *British Medical Journal.* 333(7568): 597–600.

Lewin, Simon, Claire Glenton, Heather Munthe-Kaas, Benedicte Carlsen, Christopher J. Colvin, Metin Gülmezoglu, Jane Noyes, Andrew Booth, Ruth Garside, and Arash Rashidian. 2015. "Using Qualitative Evidence in Decision Making for Health and Social Interventions: An Approach to Assess Confidence in Findings from Qualitative Evidence Syntheses (GRADE-CERQual)." *PLoS Medicine* 12(10): e1001895.

Liberati, Alessandro, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John PA Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. 2009. "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration." *PLoS Medicine* 6(7).

Light, Richard J., and David B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research.* Cambridge, Mass.: Harvard University Press.

Little, Roderick J. A., and Don B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

Macaskill, Petra, Stephen D. Walter, and Lesley Irwig. 2001. "A Comparison of Methods to Detect Publication Bias in Meta-Analysis." *Statistics in Medicine* 20(4): 641–54.

Mavridis, Dimitris, Alex Sutton, Andrea Cipriani, and Georgia Salanti. 2012. "A fully Bayesian Application of the Copas Selection Model for Publication Bias Extended to Network Meta-Analysis." *Statistics in Medicine* 32(1): 51–66.

McIntosh, Heather, and Piero Olliaro. 2000. "Artemisinin Derivatives for Treating Severe Malaria." *Cochrane Database Systematic Reviews* 2(2): CD000527.

McShane, Blakeley, and Ulf Böckenholt. 2014. "You Cannot Step into the Same River Twice: When Power Analyses Are Optimistic." *Perspectives on Psychological Science* 9(6): 612–25.

McShane, Blakeley, Ulf Böckenholt, and Karsten Hansen. 2016. "Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes." *Perspectives on Psychological Science* 11(5): 730–49.

McShane, Blakeley, and David Gal. 2015. "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence." *Management Science* 62(6): 1707–18.

Monogan, James E. 2013. "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections." *Political Analysis* 21(1): 21–37.

Moreno, Santiago G., Alex J. Sutton, A. E. Ades, Tom D. Stanley, Keith R. Abrams, Jaime L. Peters, and Nicola J. Cooper. 2009. "Assessment of Regression-Based Methods to Adjust for Publication Bias Through a Comprehensive

Simulation Study." *BMC Medical Research Methodology* 9(1): 2.

Morey, Richard D. 2013. "The Consistency Test Does No—and Cannot—Deliver What Is Advertised: A Comment on Francis (2013)." *Journal of Mathematical Psychology* 57(5): 180–83.

Nelson, Nanette, Robert Rosenthal, and Ralph L. Rosnow. 1986. "Interpretation of Significance Levels and Effect Sizes by Psychological Researchers." *American Psychologist* 41(11): 1299–301.

Orwin, Robert G. 1983. "A Fail-Safe N for Effect Size in Meta-Analysis." *Journal of Educational Statistics* 8(2): 157–59.

Peters, Jaime L., Alexander J. Sutton, David R. Jones, Keith R. Abrams, and Lesley Rushton. 2006. "Comparison of Two Methods to Detect Publication Bias in Meta-Analysis." *Journal of the American Medical Association* 295(6): 676–80.

———. 2007. "Performance of the Trim and Fill Method in the Presence of Publication Bias and Between-Study Heterogeneity." *Statistics in Medicine* 26(25): 4544–62.

———. 2008. "Contour-Enhanced Meta-Analysis Funnel Plots Help Distinguish Publication Bias from Other Causes of Asymmetry." *Journal of Clinical Epidemiology* 61(10): 991–96.

Petticrew, Mark, Matt Egan, Hilary Thomson, Val Hamilton, Renée Kunkler, and Helen Roberts. 2006. "Publication Bias in Qualitative Research: What Becomes of Qualitative Research Presented at Conferences?" *British Medical Journal* 62(6): 552–54.

Piff, Paul K., Daniel M. Stancato, Stéphane Côté, Rodolfo Mendoza-Denton, and Dacher Keltner. 2012. "Reply to Francis: Cumulative Power Calculations Are Faulty When Based on Observed Power and a Small Sample of Studies." *Proceedings of the National Academy of Sciences* 109(25): E1588.

Pigott, Therese D. 2001. "Missing Predictors in Models of Effect Size." *Evaluation and the Health Professions* 24(3): 277–307.

Pigott, Therese D., Jeffrey C. Valentine, Joshua R. Polanin, Ryan T. Williams, and Dericka D. Canada. 2013. "Outcome-Reporting Bias in Education Research." *Educational Researcher* 42(8): 424–32.

Preston, Carrol, Deborah Ashby, and Rosalind Smyth. 2004. "Adjusting for Publication Bias: Modelling the Selection Process." *Journal of Evaluation in Clinical Practice* 10(2): 313–22.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reyes, Magdalena M., Kaitlyn E. Panza, Andrés Martin, and Michael H. Bloch. 2011. "Time-Lag Bias in Trials of Pediatric Antidepressants: A Systematic Review and Meta-Analysis." *Journal of the American Academy of Child & Adolescent Psychiatry* 50(1): 63–72.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin.* 86(3): 638–41.

Rosenthal, Robert, and John Gaito. 1963. "The Interpretation of Levels of Significance by Psychological Researchers." *Journal of Psychology* 55(1): 33–38.

———. 1964. "Further Evidence for the Cliff Effect in Interpretation of Levels of Significance." *Psychological Reports* 15(2): 570. DOI: 10.2466/pr0.1964.15.2.570.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein. 2005. "Publication Bias in Meta-Analysis." In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Rücker, Gerta, James R. Carpenter, and Guido Schwarzer. 2011. "Detecting and Adjusting for Small-Study Effects in Meta-Analysis." *Biometrical Journal* 53(2): 351–68.

Rücker, Gerta, Guido Schwarzer, and James R. Carpenter. 2008. "Arcsine Test for Publication Bias in Meta-Analyses with Binary Outcomes." *Statistics in Medicine* 27(5): 746–63.

Rücker, Gerta, Guido Schwarzer, James R. Carpenter, Harald Binder, and Martin Schumacher. 2011. "Treatment-Effect Estimates Adjusted for Small-Study Effects Via a Limit Meta-Analysis." *Biostatistics* 12(1): 122–42.

Rufibach, Kaspar. 2011. "Selection Models with Monotone Weight Functions in Meta Analysis." *Biometrical Journal* 53(4): 689–704.

Schwarzer, Guido. 2016. "meta: General Package for Meta-Analysis" (4.4-0). *R* package.

Schwarzer, Guido, James Carpenter, and Gerta Rücker. 2010. "Empirical Evaluation Suggests Copas Selection Model Preferable to Trim-and-Fill Method for Selection Bias in Meta-Analysis." *Journal of Clinical Epidemiology* 63(3): 282–88.

———. 2016. "metasens: Advanced Statistical Models to Model and Adjust for Bias in Meta-Analysis" (0.3–0). *R* package.

Silliman, Nancy P. 1997. "Hierarchical Selection Models with Applications in Meta-Analysis." *Journal of the American Statistical Association.* 92(429): 926–36.

Simonsohn, Uri. 2012. "It Does Not Follow: Evaluating the One-Off Publication Bias Critiques by Francis (2012a, b,

c, d, e, f).” *Perspectives on Psychological Science* 7(6) 597–99.

———. 2013. “It Really Just Does Not Follow, Comments on.” *Journal of Mathematical Psychology* 57(5): 174–76.

Simonsohn, Uri., Lief D. Nelson, and Joseph P. Simmons. 2014. “*p*-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143(2): 534.

Smith, Richard. 1999. “What Is Publication? A Continuum.” *British Medical Journal* 318(7177): 142.

Smith, Mary Lee, Gene V. Glass, and Thomas I. Miller. 1980. *The Benefits of Psychotherapy*. Baltimore, Md.: Johns Hopkins University Press.

Song, Fujan, Alison Easterwood, Simon Guilbody, Lelia Duley, and Alexander J. Sutton. 2000. “Publication and Other Selection Biases in Systematic Reviews.” *Health Technology Assessment* 4(10): 1–115.

Song, Fujan, Nick Freemantle, Trevor A. Sheldon, Allan House, Paul Watson, Andrew Long. 1993. “Selective Serotonin Reuptake Inhibitors: Meta-Analysis of Efficacy and Acceptability.” *British Medical Journal* 306(6879): 683–87.

Stanley, Tom D. 2005. “Beyond Publication Bias.” *Journal of Economic Surveys* 19(3): 309–45.

Stanley, Tom D., and Hristos Doucouliagos. 2014. “Meta-Regression Approximations to Reduce Publication Selection Bias.” *Research Synthesis Method* 5(1): 60–78.

Stanley, Tom D., Stephen B. Jarrell, and Hristos Doucouliagos. 2010. “Could It Be Better to Discard 90% of the Data? A Statistical Paradox.” *American Statistician* 64(1): 70–77.

Stern, Jerome M., and R. John Simes. 1997. “Publication Bias: Evidence of Delayed Publication in a Cohort Study of Clinical Research Projects.” *British Medical Journal* 315(7109): 640–45.

Sterne, Jonathan A. C., Betsy J. Becker, and Matthias Egger. 2005. “The Funnel Plot.” In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Sterne, Jonathan A. C., and Matthias Egger. 2001. “Funnel Plots for Detecting Bias in Meta-Analysis: Guidelines on Choice of Axis.” *Journal of Clinical Epidemiology* 54(10): 1046–55.

———. 2005. “Regression Methods to Detect Publication and Other Bias in Meta-Analysis.” In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Sterne, Jonathan A. C., Matthias Egger, and George Davey Smith. 2001. “Investigating and Dealing with Publication and Other Biases in Meta-Analysis.” *BMJ: British Medical Journal* 323(7304): 101.

Sterne, Jonathan A. C., David Gavaghan, and Matthias Egger. 2000. “Publication and Related Bias in Meta-Analysis: Power of Statistical Tests and Prevalence in the Literature.” *Journal of Clinical Epidemiology* 53(11): 1119–29.

Sterne, Jonathan A., Alex J. Sutton, John P. Ioannidis, Norma Terrin, David R. Jones, Joseph Lau, James Carpenter, Gerta Rücker, Roger M. Harbord, Christopher H. Schmid, Jennifer Tetzlaff, Jonathan J. Deeks, Jaime Peters, Petra Macaskill, Guido Schwarzer, Sue Duval, Douglas G. Altman, David Moher, and Julian P. T. Higgins. 2011. “Recommendations for Examining and Interpreting Funnel Plot Asymmetry in Meta-Analyses of Randomised Controlled Trials.” *British Medical Journal* 343(7818): 302.

Stouffer, Samuel A., Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams Jr. 1949. *The American Soldier: Adjustment During Army Life*. Studies in Social Psychology in World War II, vol. 1, edited by Samuel Stouffer and Edward A. Suchman. Princeton, N.J.: Princeton University Press.

Sutton, Alexander J. 2005. “Evidence Concerning the Consequences of Publication and Related Biases.” In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Sutton, Alexander J., Keith R. Abrams, and David R. Jones. 2002. “Generalized Synthesis of Evidence and the Threat of Dissemination Bias: The Example of Electronic Fetal Heart Rate Monitoring (EFM).” *Journal of Clinical Epidemiology* 55(10): 1013–24.

Sutton, Alexander J., and Therese D. Pigott. 2004. “Bias in Meta-Analysis Induced by Incompletely Reported Studies.” In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Tang, Jin-Ling, and Joseph L. Y. Liu. 2000. “Misleading Funnel Plot for Detection of Bias in Meta-Analysis.” *Journal of Clinical Epidemiology* 53(5): 477–84.

Terrin, Norma, Christopher H. Schmid, and Joseph Lau. 2005. “In an Empirical Evaluation of the Funnel Plot, Researchers Could Not Visually Identify Publication Bias.” *Journal of Clinical Epidemiology* 58(9): 894–901.

Terrin, Norma, Christopher H. Schmid, Joseph Lau, and Ingram Olkin. 2003. “Adjusting for Publication Bias in the Presence of Heterogeneity.” *Statistics in Medicine* 22(13): 2113–26.

Trikalinos, Thomas A., and John P. A. Ioannidis. 2005. “Assessing the Evolution of Effect Sizes Over Time.”

In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* edited by Hannah R. Rothstein, Alexander J. Sutton, and Michael Borenstein. Chichester, UK: John Wiley & Sons.

Ulrich, Rolf, and Jeff Miller. 2015. "p-hacking by Post Hoc Selection with Multiple Opportunities: Detectability by Skewness Test?: Comment on Simonsohn, Nelson, and Simmons (2014)." *Journal of Experimental Psychology: General* 144(6): 1137–45.

van Aert, Robbie C. M. 2015. "puniform: Meta-analysis with p-uniform" (0.0.0). *R* package.

van Aert, Robbie C. M., Jelte M. Wicherts, and Marcel A. van Assen. 2016. "Conducting Meta-Analyses Based on *p*-Values: Reservations and Recommendations for Applying *p*-Uniform and *P*-Curve." *Perspectives on Psychological Science* 11(5): 713–29.

van Assen, Marcel A., Robbie C. M. van Aert, and Jelte M. Wicherts. 2015. "Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies." *Psychological Methods* 20(3): 293.

Vandekerckhove, Joachim, Maime Guan, and Steven A. Styrcula. 2013. "The Consistency Test May Be Too Weak to Be Useful: Its Systematic Application Would Not Improve Effect Size Estimation in Meta-Analyses." *Journal of Mathematical Psychology* 57(5): 170–73.

van Enst, W. Annefloor, Eleanor Ochodo, Rob J.P.M. Scholten, Lotty Hooft, and Mariska M. Leeflang. 2014. "Investigation of Publication Bias in Meta-Analyses of Diagnostic Test Accuracy: A Meta-Epidemiological Study." *BMC Medical Research Methodology* 14(1): 70–81.

Vevea, Jack L., Nancy C. Clements, and Larry V. Hedges. 1993. "Assessing the Effects of Selection Bias on Validity Data for the General Aptitude Test Battery." *Journal of Applied Psychology* 78(6): 981–87.

Vevea, Jack L., and Larry V. Hedges. 1995. "A General Linear Model for Estimating Effect Size in the Presence of Publication Bias." *Psychometrika* 60(3): 419–35.

Vevea, Jack L., and Carol M. Woods. 2005. "Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions." *Psychological Methods* 10(4): 428–43.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the *metafor* Package." *Journal of Statistical Software* 36(3): 1–48.

Weinhandl, Eric. D., and Sue Duval. 2012. "Generalization of Trim and Fill for Application in Meta-Regression." *Research Synthesis Methods* 3(1): 51–67.

Williamson, Paula R. and Carol Gamble. 2005. "Identification and Impact of Outcome Selection Bias in Meta-Analysis." *Statistics in Medicine* 24(10): 1547–61.

Zarin, Deborah A., Tony Tse, Rebecca J. Williams, Robert M. Califf, and Nicholas C. Ide. 2011. "The ClinicalTrials.gov Results Database—Update and Key Issues." *New England Journal of Medicine* 364(9): 852–60.