Advances in Social Sciences Research Journal – Vol.2, No.3
Publication Date: March 25, 2015
DoI:10.14738/assrj.23.1044.
Maag, J. W. & Losinski, M. (2015). Thorny issue and Prickly solutions: Publication Bias in Meta-Analytic Reviews in the Social Sciences. *Advances in Social Sciences Research Journal, 2(3)* 242-253.

Thorny Issues and Prickly Solutions: Publication Bias in Meta-Analytic Reviews in the Social Sciences

John W. Maag
University of Nebraska-Lincoln

Mickey Losinski
Kansas State University

Abstract

Two main issues that involve publication bias are the types of studies that get published and the lack of published studies reflecting null results (i.e., the file drawer problem). These issues are of central importance to researchers who conduct systematic (meta-analytic and narrative) reviews of a body of literature because they may result in misleading conclusions being drawn that, in turn, can adversely impact the decisions of policymakers and practitioners who rely on these reviews. Yet without a thorough unbiased critique, a body of literature may not be as evidence-based as the reviewers believe it to be from the extant research analyzed. Therefore, the purpose of this article is to consider issues involved in reporting and evaluating results of meta-analytic reviews, provide an overview of statistical techniques for addressing the file drawer problem (i.e., lack of journals publishing studies with null results), discuss the extent to which the file drawer is actually a "problem," and offer some solutions.

Implementing evidence-based practices (EBPs) has entered the landscape of many social science fields including, but not limited to, corrections, general and special education, mental health, nursing, clinical psychology, and social work (Spring, Neville, & Russell, 2012). In the case of general and special education, the federal government has enacted legislative mandates, such as the Elementary and Secondary Education Act ([ESEA] 2002; formerly the No Child Left Behind Act) and the Individuals with Disabilities Education Act (IDEA, 2004), to ensure that teachers and other school personnel use evidence-based practices. In spite of these mandates, it continues to be a daunting task to develop, validate, and disseminate evidence-based practices to practitioners for use in schools (Cook & Odom, 2013).

One of the primary means for determining whether a particular intervention is evidence-based is through the process of reviewing a body of literature (Banks, Kepes, & Banks, 2012; Cordray & Morphy, 2009). The fields of psychology and education have a long, rich history of journals publishing literature reviews on various topics. For example, the first edition of Psychological Review was published in 1894 while Review of Educational Research began in 1931. There are two types of literature reviews: descriptive and systematic reviews. Descriptive reviews are typically less robust because authors tend to either pick and choose which studies they wish to include in their review or the studies are often evaluated according to the seniority of the authors and the status of the journal in which they are published (Järvholm & Bohlin, 2014). A more rigorous approach is through the use of systematic reviews, of which there are two types: narrative and meta-analytic. Narrative reviews typically rely on the reviewers' interpretations of a body or literature based on such variables as participant characteristics, settings, dependent measures, and experimental design. Conversely, meta-analytic reviews add statistical procedures to the systematic process in order to reach conclusions regarding a

particular body of literature. When properly implemented, meta-analytic reviews can provide a high-yield source of clinically significant information for an evidence-based practice (Mundy & Stein, 2008). In either case, the goal of the reviewers is to conduct a systematic, replicable search of all published, and sometimes unpublished (e.g., dissertations), research studies using objective criteria.

A major problem authors encounter when identifying and collecting research articles, regardless of type of review they conduct, is bias in the obtained (or unobtainable) literature. There are two general types of bias: study level and review level, and they are both interrelated (Banks et al., 2012; Borenstein, Hedges, Higgins, & Rothstein, 2009). Study-level bias occurs when the obtained research articles for a review only report significant findings or when journal editors require authors to omit findings of less importance in order to conserve page numbers (Evangelou, Trikalinos, & Ioannidis, 2005; Song et al., 2010; Sutton, 2009). Review-level bias primarily reflects the inability of reviewers to locate all studies conducted in a body of literature. The main reason for this type of bias is that researchers tend not to submit studies for publication that have null results and journal editors are hesitant to publish null results (Dickersin, 2005; Drotar, 2010; Neuliep & Crandall, 1990). These biases, through publication omission, creates the "file drawer effect" (Rosenthal, 1979) and in the case of meta-analytic techniques, this systematic omission from the literature may distort the omnibus effect size with the exaggerations being strongest when the true effect size approaches zero (Bradley & Gupta, 1997).

A more insidious cause of publication bias may be organizations (particularly for-profit organizations) intentionally withholding findings because the study's insignificant results may shed light on the suspect nature of their product (McDaniel, Rothstein, & Whetzel, 2006). For example, during the 1970s and 1980s the Canters' Assertive Discipline was the most popular approach used by schools (Charles, 2008). However, in a systematic review of the literature, Render, Nell, Padilla, and Krank (1989) discovered how the "evidence" for Assertive Discipline was either misleading or reported selectively.

Regardless of the source of bias, studies not published due to null results or whose results were "cleaned up" by rerunning analyses until expected results were achieved (Ferguson & Heene, 2012), creates a threat to the validity of conclusions drawn from a systematic review. This threat is greater for narrative systematic reviews because, unlike meta-analytic reviews, the former does not involve statistical methods that may be able to account for, or hypothesize the extent of, the file drawer effect. However, statistical approaches are not fool-proof and there is some controversy as to which method is best to account for the "file drawer" and whether it even poses a problem at all for authors conducting systematic reviews. Therefore, the purpose of this article is to: (a) consider issues involved in reporting results of systematic reviews, focusing on those using a meta-analytic approach, (b) review statistical techniques for addressing the file drawer problem, (c) discuss whether the statistical techniques really matter, and (d) offer suggestions for future reviews. The reason this article focuses solely on meta-analytic (versus narrative) reviews is because of the statistical techniques that can be used with this approach. That is not to say that authors of narrative systematic reviews should eschew discussing publication bias, but it is more an issue of interpreting statistical results and conclusions drawn from them because there is no way to account for missing studies or those studies that eliminated reporting null results.

Issues in Reporting and Evaluating Results: Importance of Guidelines

A major issue in evaluating results from meta-analytic reviews has been the lack of clear standards for conducting and presenting results. However, within the past decade the U.S. Department of Education's (2003) What Works Clearinghouse ([WWC] http://ies.ed.gov/ncee/wwc/) has enlisted the service of trained reviewers to synthesize a research-base (Cordray & Morphy, 2009). Results of reviews are then posted to the WWC website as a means of disseminating the findings to practitioners and researchers. However, in certain areas, such as special education, reporting and evaluation are still emerging and there is a need for trained evaluators (Cook & Odom, 2013; Council for Exceptional Children, 2014). In this section, the rational for guidelines is elaborated on and different sets of guidelines are described.

Rationale for Guidelines

Having a set of guidelines is important to determine the extent to which a body of research that was reviewed could be considered evidence-based—especially considering the difference between systematic and descriptive reviews described previously. This distinction is important because descriptive reviews may be conflated with narrative systematic reviews (since neither rely on statistical analysis) to add veracity to an intervention being considered evidence-based. For example, Maag, Losinski, and Katsiyannis (2014) reported that there have been 18 reviews examining the efficacy of pharmacologic agents for treating childhood and adolescent depression. It would be simple to assume that from the shear volume of reviews alone, antidepressants would be effective for treating childhood and adolescent depression, however, many of those reviews were descriptive which may be methodologically flawed given the problems described formerly.

One of the greatest threats of bias in evaluating meta-analytic reviews is that the format, topic, and level of detail authors provide can vary greatly. It becomes difficult to appraise a body of literature when there is no consistent reporting of methodology and analysis. Järvholm and Bohlin (2014) stated that this problem is not as great when analyzing a body of randomized control trial studies because the evaluation process is well formalized. However, they added that observational studies require more judgment (e.g., accounting for confounding variables) and that systematic reviews of those types of studies may vary greatly depending on the topic of the review and types of questions the reviewers ask.

Types and Purposes of Standardized Procedures

The importance of standardized protocols is essential—especially in certain disciplines such as special education that faces numerous experimental complexities and in which a variety of research designed are used including nonrandomized group designs, single-case research designs (SCRDs), and correlational studies (Heward, 2009; Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2005). Therefore, procedures such as PICO (population, intervention, comparison, and outcome) developed by Schünemann, Oxman, and Fretheim (2006) provide reviewers with a formalized system to structure their research question(s).

Many important procedures have been developed for meta-analytic reviews—all of which vary somewhat depending on the topic and type of research design. For example, Stroup et al. (2000) developed guidelines for meta-analysis of observational studies in epidemiology (MOOSE) and Atkins et al. (2004) developed a system called GRADE that ranks the quality of evidence and the strength of recommendations authors of meta-analyses make regarding their findings. Several years later, Liberati et al. (2009), along with a committee of contributors,

developed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). According to these guidelines, there are 12 items that must be addressed in the methods section of meta-analytic reviews that evaluate health interventions: protocol and registration, eligibility criteria, information sources, search strategy, study selection, data collection process, data items, risk of bias in individual studies, summary measures, synthesis of results, risk of bias across studies, and any additional analyses (e.g., subgroup analysis or sensitivity). Each item must be addressed to ensure clarity and transparency of the reviews. The PRISMA guidelines have become the accepted approach for conducting meta-analyses in a number of fields, including education, and also used by a large body of scholarly journals.

STATISTICAL TECHNIQUES FOR ADDRESSING THE FILE DRAWER

Problem: Do They Matter?

Before proceeding, it is important to acknowledge that there are different sources of publication bias besides the "file drawer" problem. However, those sources of publication bias have simple solutions, although they are rarely implemented by journal editors and publishing companies. For example, the "time-lag" problem, which refers to the time it takes for a study to be published from the time it was conducted, can be omitted by editors of journals publishing manuscripts in the order submissions were accepted rather than unilaterally deciding to move up publication of studies deemed more interesting than others (Borenstein et al., 2009).

Nevertheless, the persistent file drawer problem articulated by Rosenthal (1979) approximately 35 years ago still posses a thorny issue for meta-analysts today. The "problem" exists in whether unpublished studies are randomly or systematically omitted from a body of literature. For example, if the unpublished studies were randomly omitted, it is likely that less information will be obtained, wider confidence intervals would be required, and more powerful tests would be necessary, but the missing studies would have no systematic impact on meta-analytic effect sizes (Bornstein et al., 2009). However, if studies are systematically omitted from publication due to having non-significant results, then the conclusions drawn from a meta-analysis will be biased (Rothstein, Sutton, & Borenstein, 2005).

Statistical Methods to Combat the File Drawer Problem

There have been a myriad of creative methods to determine whether a meta-analysis includes as many statistically non-significant results as would be expected from a specific group of effect sizes. Rosenthal's (1979) fail-safe n was one of the first and involves calculating the number of studies averaging null results that would need to be added to the given set of observed effects to bring the overall effect to non-significant levels. Alternative methods for calculating a fail-safe n have also been developed (e.g., Orwin, 1983; Rosenberg, 2005).

Additionally, an adjusted rank correlation test was proposed by Begg and Mazumdar (1994) to address the file drawer problem. In this approach, the correlations between effect sizes have been standardized and their variances are examined. High correlations between the two indicate the possibility of publication bias. The test has been found to be fairly powerful for meta-analyses with at least 75 component studies but only moderately powerful with 25 or fewer studies and thus must be interpreted cautiously in small meta-analyses. The adjusted rank correlation test is similar to funnel plots in which skewness in a graph is identified visually. In addition to rank correlation tests, linear-regression tests like the Egger test (Egger, Davey Smith, Schneider, & Minder, 1997) and Peters' model (Peters, Sutton, Jones, Abrams, & Rushton, 2006) have also been used to assess funnel symmetry (Sterne & Egger, 2005). Further, Moreno et al. (2009) compared several regression methods (Egger et al., 1997) used

to adjust for possible publication bias, and found that all the methods displayed good performance with none consistently outperforming the others.

Perhaps the most popular method for statistically addressing the number of missing studies in a meta-analysis is Trim & Fill (Duval & Tweedie, 2000). Trim & Fill is a nonparametric rank-based augmentation technique to account for asymmetry in a funnel plot. The method "trims" (i.e., removes) the asymmetric smaller studies on the right-hand side of the funnel for which there are no left-hand counterparts. A revised pooled estimate adjusting for missing studies is derived from this reduced dataset and then replaced. Studies assumed to be missing are assigned to the opposite side of the funnel and an adjusted pooled effect is calculated using this supplemented dataset in order to provide an estimate of the number of missing studies. A problem with the Trim & Fill method is that it is based on the assumption that there should be a symmetric funnel plot and any asymmetry is due solely to publication bias rather than other variables (Terrin, Schmid, Lau, & Olkin, 2003). However, in a series of experiments with simulated data sets, Moreno et al. (2009) concluded the Trim & Fill method is inferior to regression-based methods because of misleading adjustments and poor coverage probabilities, especially when between-study variance is present.

P-Hacking: A Corollary to the File Drawer Problem

Journals tend to publish studies that are statistically significant and reject (i.e., place in the file-drawer) studies that do not show significant results. Researchers are well aware of this phenomenon and, consequently, sometimes statistically manipulate nonexistent effects in order to obtain "significance." This practice (i.e., "chasing the significant") may lead to the publication of false-positives that result in a body of work being unrepresentative of reality (Ioannidis, 2008). More specifically, researchers have various decisions to make while collecting and analyzing data including, but not limited to, whether to collect additional data, which outliers to omit, which dependent variables to analyze, which covariates, and what type, if any, of post hoc tests to run. Researchers who do not make these decisions in advance but instead make them during data analysis—especially changing or adding statistical procedures to improve the odds of publishing—adds to the unrepresentativeness of a body of research (Kunda, 1990). Therefore, rather than placing entire studies in the file-drawer, researchers may simply discard the subsets of analyses that produce non-significant results. Simonsohn, Nelson, and Simmons (2013) referred to this practice as p-hacking to describe researchers finding statistically significant support for nonexistent effects.

Simonsohn et al. (2013) developed the p-curve as a way to combat p-hacking and the file drawer problem. The p-curve represents the distribution of statistically significant p values for the findings from a collection of independent studies. Right skewed p-curves consist of many low probabilities (e.g., .01s) where as left-skewed p-curves have higher p values (e.g., .05s). The p-curves that are right-skewed are indicative of evidential value whereas left-skewed p-curves suggest the presence of intense p-hacking. Simonsohn and his colleagues indicated that three aspects of p values must be addressed in order for the p-curve to be valid: (a) associated with the hypothesis, (b) statistically independent from other selected p values, and (c) distributed uniformly under the null hypothesis. They also indicated that the p-curve is counter-indicated for assessing the validity of a set of non-significant findings because it is exclusively derived from statistically significant findings. Therefore, the p-curve adds another statistical method for addressing the file drawer problem but like all other statistical methods it has limitations. Perhaps the biggest question is this: Does the file drawer really pose a problem to the analysis of a systematic review?

Validity of the File Drawer Problem

Perhaps a larger issue than evaluating methods for addressing the file drawer is whether it poses a problem at all to the analysis of systematic reviews. Throughout the history of published research, there has always been censorship of studies for a variety of reasons (Rothstein et al., 2005). However, the importance of the file drawer problem has received considerable attention with the widespread use of meta-analytic approaches and concerns regarding the ability to accurately interpret results culled from them. After all, if meta-analyses do not include a random sample, then their use to identify practices as evidence-based becomes problematic. The reason is because a meta-analysis is theoretically analogous to conducting a randomized controlled trial study except the participants in the former are at the study level rather than individual level (Borenstein et al., 2009).

The statistical methods described previously attempted to account for the absence of non-significant studies but there nevertheless is no way to determine what percentage of research is not published. The very premise of the file drawer problem makes it impossible to know the number of null results. The irony is that it is much easier to correct publication bias than it is to detect it with any level of certainty (Cooper & Hedges, 2009).

In an attempt to experimentally determine the extent to which the file drawer actually posses a problem, Dalton, Aguinis, Dalton, Bosco and Pierce (2012) conducted five experiments examining a total of 80,710 correlations included in 787 matrices from non-experimental research. In study 1, they examined correlations from studies published in three journals (Academy of Management Journal, Journal of Applied Psychology, and Personnel Psychology) between 1985 and 2009 and found that 46.81% were insignificant. Study 2 examined correlations from 51 meta-analyses published in the aforementioned three journals and found that 44.31% were insignificant. Study 3 evaluated correlations from non-published manuscripts by identifying faculty members of 30 schools of business and psychology, randomly contacted one half (n = 361), and requesting correlation matrices for papers over the years they decided not to submit for publication. Dalton and his colleagues found that 45.45% of the correlations were statistically insignificant. Study 4 examined 20,860 correlations from doctoral dissertations and found that 50.78% were insignificant. In study 5, they compared the average magnitude of a sample of 1,002 correlations from Study 1 (published articles) against 1,224 from Study 4 (dissertations) and found them virtually identical (.2270 and .2279, respectively). From the five studies, they concluded that the file drawer problem does not create an inflation bias and does not present a threat to the validity of conclusions derived from meta-analyses. However, the results of Dalton et al. do not prove unequivocally the absence of the file drawer problem for two reasons. First, they did not examine randomly controlled trials nor research using single case methodology. Second, the research they examined was restricted to the journals sampled. Therefore, the results for other constructs in the social sciences with other experimental designs may yield different results and, as such, represents a fruitful area for replication.

A different and novel approach to test the file drawer problem was proposed by Howard et al. (2009) using the psychotherapy efficacy literature. Rather than correcting for bias statistically, they suggested performing a mini-literature meta-analysis of new and, as of yet, unpublished studies and determining if the results approached the value of a meta-analysis obtained from the entire literature (presumably biased because of the file drawer problem) or whether it was closer to the null value (d = .00). For example, if the authors' first three unpublished psychotherapy outcome studies yielded effect sizes (i.e., percentage of treatment participants

who were superior to the average control participant) of 77%, 69%, and 73%, the mean of 73% would be closer to the literature (i.e., 75%) than it would be to the null of 50%. Therefore, whatever bias the file drawer exerts on the psychotherapy outcome literature would be small and relatively unimportant. Conversely, if the first three unpublished outcome studies yielded d-values that convert to percentages of 54%, 48%, and 50%, it would be reasonable to expect that the file drawer effect might be enormous and may abrogate the existing literature. Howard and his colleagues concluded their discourse by suggesting that effect sizes from non-published studies (i.e., free of the file drawer effect) that are substantially lower than those obtained from the entire literature does not categorically indicate a body of literature is specious. Rather, it simply indicates that additional sets of studies conducted by different researchers need to be amassed before assessing the final disposition of an intervention.

Solutions to the File Drawer Effect

Solutions to the thorny and intransigent problems of the file drawer effect will most likely be unpalatable and prickly. It requires a change in the scientific culture that non-significant findings are, in fact, worthy of publication—truly a daunting task. Three approaches have been advocated: result-blind reviews, abandoning traditional statistical procedures, and the universal implementation of study registries.

RESULTS-BLIND REVIEWS

Greve, Bröder, and Erdfelder (2013) suggested that editors of journals engage in result-blind peer reviews. They also described an "ideal" editorial policy in which a journal nurtured a scientific culture of submitting well-designed and technically sound empirical research regardless of the results obtained. To support this principle, they contended that certain results obtained from a null hypothesis may be more "significant" than alternative hypothesis generated from studies controlling for Type-1 and Type-2 error probabilities. This orientation was first advocated by Walster and Cleary (1970) over 40 years ago, and to this day is still being endorsed (e.g., Howard et al., 2009). The persistent difficulty turning this long-standing recommendation into reality has not been easy for two reasons.

First, it is not just the mentality of editors and publishers of journals that must change but also that of authors who chose to either submit or "shelf" manuscripts. For example, about 20 years ago, Rotton, Foos, Van Meek, and Levitt (1995) surveyed 740 authors of empirical articles that had appeared in 75 journals. The most frequently reason they gave for not publishing a manuscript was non-significant results. More recently Reysen (2006) surveyed 236 psychologists' opinions about publication of non-significant results. Tenured faculty were significantly less likely to write manuscripts for studies with non-significant results than non-tenured faculty. Reasons for not writing up results from a study were perceived inability to publish non-significant results, inability to interpret results, and the belief that the results were unimportant.

Second, there are some disciplines that straddle both social and biological sciences in which methods sections of manuscripts are virtually identical due to commonly accepted protocols such as in many areas of audiology, and the clinical research on vestibular physiology that examines inner ear functioning in cases of blast or sports-related head trauma. Closer to psychology and special education, the Journal of Applied Behavior Analysis publishes primarily single-case experimental design studies in which certain methodologies are quite similar (e.g., Iwata's traditional functional analysis, methods for implementing self-monitoring).

Conversely, other research in these fields examines complex multifaceted interventions in which describing detailed and empirically valid procedures are essential. It then becomes a subjective endeavor for journal editors to determine which manuscripts to exclude results, such as those examining complex interventions, versus manuscripts with standard methodology protocols, in which case critiquing only methods sections would provide little differential information. Parenthetically, the acceptance of a manuscript for publication does not just depend on the rigor of the methods section but also includes the level of sophistication with which researchers integrate results with the discussion of previous research, limitations, and areas requiring further study.

Abandoning Traditional Statistics

Another, but extremely prickly, solution would be to abandon traditional statistics in favor of Bayesian statistics (Howard et al., 2009). In this approach, constraints of a given population are considered to be random and consisting of explicit probability distributions. These probabilities measure "degree of belief." The rules of conditional probability are used to express a subjective degree of belief to account for observed data. Bayes' theorem has been used to assist in the diagnosis of both physical and psychological conditions—especially when information from two or more sources need to be combined. For example, Kemp et al. (1998) used Bayes' theorem to calculate the probability that a diagnosis of physical abuse would be correct by combining information from the number of bruises with the prior likelihood of abuse. The problem is that the Bayesian approach is antithetical to the way many social scientists were trained.

Universal Implementation of Study Registries

In order to minimize the file drawer problem, Chan, Hrobjartsson, Haahr, Gotzsche, and Altman (2004) recommended the universal implementation of study registries—a policy also endorsed by Howard et al. (2009). Studies would be registered and protocols published online prior to conducting a study. This approach would also help to minimize p-hacking and the practice of certain results being selectively reported while others totally omitted. For example, the International Committee of Medical Journal Editors (ICMJE; 2004) adopted a resolution that required registration of all clinical trials in a public registry before any consideration for publication was made. Researchers can register at the ICMJE website (http://www.icmje.org) or register at six other ICMJE accepted registries. Currently, there are 14 journals that are members of ICMJE and literally hundreds of other journals world-wide that follow their recommendations. Unfortunately, no similar organizations currently exist in many social science disciplines such as psychology and education.

Regardless, the adoption of each recommendation would be very costly and require disciplines in the social sciences to rethink how results are presented and published. Howard et al. (2009) proposed two alternatives. First, a given discipline could occasionally invite a researcher to construct a new mini-literature to test confidence in the results and conclusions of the entire literature. Second, common methodologies could be periodically tested by researchers and possibly improved. Unfortunately, replication for the purposes of enhancing methodology is rare compared to replication for extending findings.

CONCLUSION

The adoption of the three recommended solutions to publication bias would indeed be very prickly because of the cost and requiring disciplines in the social sciences to rethink how results are presented and published. Howard et al. (2009) suggested two ways for

circumventing these problems: (a) a given discipline could occasionally invite a researcher to construct a new mini-literature to test confidence in the results and conclusions of the entire literature and (b) common methodologies could be periodically tested by researchers and possibly improved. Unfortunately, replication for the purposes of enhancing methodology is rare compared to replication for extending findings.

Hence, it is likely that the types of studies that get published and the lack of journals publishing studies reflecting null results will continue to create bias in the way meta-analytic reviews are evaluated and the conclusions researched that, in turn, impact the decisions of policymakers and practitioners who rely on these reviews. Furthermore, without a thorough unbiased critique of a body of literature, a particular intervention may not be as evidence-base as the reviewers believed it to be from the extant research analyzed. In order to ensure exhaustiveness for systematic review, Potvin, Sepehry, and Stip (2007) recommended including studies that appear in Dissertation Abstracts and book chapters. Parenthetically. Shadish, Doherty and Montgomery (1989) concluded, based on evidence from their studies and others, that a conservative assumption is that population effect sizes are only 70–90% as large as those computed from published studies.

The range to which social scientists perceive publication bias to impact the results of systematic reviews varies. However, there is a growing emphasis and discussion on the importance of addressing and accounting for bias (e.g., Banks et al., 2012; Dalton et al., 2012; Howard et al., 2009; Simonsohn et al., 2013). Ferguson and Heene (2012) have been more urgent and forceful as the title of their article indicates: "A Vast Graveyard of Undead Theories." They specifically argued that the practice in social sciences to avoid publishing null results limits accurate replication—the cornerstone for determining which practices are evidence-based. They also cautioned that, because science relies on the process of falsification, without the acknowledgement of failed results certain ideologically popular theories may be perpetuated in the absence of any factual basis. A classic example described by Cook (2014) was the widespread fear that the measles-mumps-rubella (MMR) vaccine was associated with autism even though contrasting evidence existed from large, high-quality studies and meta-analyses.

The issues presented in this article focused primarily on meta-analytic reviews because they may be more robust than narrative, and certainly descriptive, reviews. Yet where are all the null results in these reviews? asked Ferguson and Heene (2012). To answer this question, they described the phenomenon of researchers "chasing the significant" by increasing sample sizes until they obtain statistical significance regardless of the inconsequentiality of the ensuing findings. Banks et al. (2012) suggested that all systematic reviews be required to address publication bias as a condition for acceptance in any journal. It is probably unrealistic to institute procedures to eradicate all publication bias, yet to not try devalues the scientific method that has yielded amazing advancements in so many social science fields, and the power of meta-analytic reviews to indicate the degree to which an intervention is evidence-based.

References

Atkins, D., Best, D., Briss, P. A., Eccles, M., Falck-Ytter, Y., Flottorp, S., . . . GRADE Working Group. (2004). Grading quality of evidence and strength of recommendations. British Medical Journal, 328, 1490.

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. Educational Evaluation and Policy Analysis, 34, 259-277.

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

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

Bradley, M. T., & Gupta, R. D. (1997). Estimating the effect of the file drawer problem in meta-analysis. Perceptional & Motor Skills, 85, 719-722.

Chan, A., Hrobjartsson, A., Haahr, M. T., Gotzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols for published articles. Journal of the American Medical Association, 291, 2457-2465.

Charles, C. M. (2008). Building classroom discipline (9th ed.). Boston, MA: Pearson Education.

Cook, B. G. (2014). A call for examining replication and bias in special education research. Remedial and Special Education, 35, 233-246.

Cook, B. G., & Odom, S. (2013). Evidence-based practices and implementation science in special education. Exceptional Children, 79(2), 135-144.

Cooper, H., & Hedges, L. V. (2009). Potentials and limitations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (pp. 561-572). New York, NY: Russell Sage Foundation.

Cordray, D. S., & Morphy, P. (2009). Research synthesis and public policy. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (pp. 473-497). New York, NY: Russell Sage Foundation.

Council for Exceptional Children. (2014). Council for Exceptional Children standards for evidence-based practices in special education. Retrieved from http://www.cec.sped.org/~/media/Files/Standards/Evidence%20based%20Practices%20and%20Practice/CECs%20EBP%20Standards.pdf

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. Personnel Psychology, 65, 221-249.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 11-34). Chichester, UK: Wiley.

Drotar, D. (2010). Editorial: A call for replications of research in pediatric psychology and guidance for authors. Journal of Pediatric Psychology, 35, 801-805.

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

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

Elementary and Secondary Education Act [ESEA], P.L. 107-110 (2002).

Evangelou, E., Trikalinos, T. A., & Ioannidis, J. P. (2005). Unavailability of online supplementary scientific information from articles published in major journals. The FASEB Journal, 19, 1943-1944.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. Perspectives on Psychological Science, 7, 555-561.

Greve, W., Bröder, A., & Erdfelder, E. (2013). Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture. European Psychologist, 18, 286-294.

Heward, W. L. (2009). Exceptional children: An introduction to special education (9th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Howard, G. S., Lau, M. Y., Maxwell, S. E., Venter, A., Lundy, R., & Sweeny, R. M. (2009). Do research literatures give correct answers? Review of General Psychology, 13, 116-121.

International Committee of Medical Journal Editors. (2004). Recommendations for the conduct, reporting, editing, and publication of scholar work in medical journals. Retrieved from http://www.icmje.org/

Individuals with Disabilities Education Improvement Act [IDEA], 34 C.F.R. §300.7 (2004).

Ioannidis, J. P. A. (2008). Why most published research findings are false [Editorial material]. PLoS Medicine, 2, 696-701. doi: 10.1371/journal.pmed.0020124

Järvholm, B., & Bohlin, I. (2014). Evidence-based evaluation of information: The centrality and limitations of systematic reviews. Scandinavian Journal of Public Health, 42, 3-10.

Kemp, A. M., Kemp, K. W., Evans, R., Murray, L., Guildea, Z. E. S., Dunstan, F. D. J., & Sibert, J. R. (1998). Diagnosing physical abuse using Bayes' theorem: A preliminary study. Child Abuse Review, 7, 178-188.

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108, 480-498.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. A., . . . Moher, D. (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), e1000100. doi:10.1371/journal.pmed.1000100

Maag, J. W., Losinski, M., & Katsiyannis, A. (2014). Meta-analysis of psychopharmacologic treatment of child and adolescent depression: Deconstructing previous reviews, moving forward. Journal of Psychology and Psychotherapy, 4, 158-167.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. Personnel Psychology, 59, 927-953. doi:910.1111/j.1744-6570.2006.00059.x

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. BMC Medical Research Methodology, 9(2). doi: 10.1186/1471-2288-9-2

Mundy, K. M., & Stein, K. F. (2008). Meta-analysis as a basis for evidence-based practice: The question is, why not? Journal of the American Psychiatric Nurses Association, 14, 326-328.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. Journal of Social Behavior and Personality, 5, 85-90.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. Exceptional Children, 71, 137-148.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. Journal of Educational Statistics, 8, 157-159.

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

Potvin, S., Sepehry, A. A., & Stip, E. (2007). Co-morbid substance-use in schizophrenia: The file drawer effect. Schizophrenia Research, 90, 351-352.

Render, G. F., Nell, J. E., Padilla, M., & Krank, H. M. (1989). What research really shows about assertive discipline. Educational Leadership, 46(6), 72-75.

Reysen, S. (2006). Publication of nonsignificant results: A survey of psychologists' opinions. Psychological Reports, 98, 169-175.

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. Evolution, 59, 464-468.

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

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analyses. West Sussex, England: John Wiley & Sons.

Rotton, J., Foos, P., Van Meek, L., & Levitt, M. (1995). Publication practices and the file drawer problem: A survey of published authors. Journal of Social Behavior and Personality, 10, 1-13.

Schünemann, H. J., Oxman, A. D., & Fretheim, A. (2006). Improving the use of research evidence in guideline development: 6. Determining which outcomes are important. Health Research Policy and Systems, 4, 18.

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from the family/marital psychotherapy literature. Clinical Psychology Review, 9, 589-603.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-curve: A key to the file-drawer. Journal of Experimental Psychology: General, 143, 534-547.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., . . . Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. Health Technology Assessment, 14, 1-220. doi: 10.3310/hta14080

Spring, B., Neville, K., & Russell, S. W. (2012). Evidence-based practice. In V. S. Ramachandran (Ed.), Encyclopedia of human behavior (pp. 86-93). New York, NY: Academic Press.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias I meta-analysis. In H. R. Rotherstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 99-110). Chichester, UK: John Wiley & Sons, Ltd.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., . . . Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. Journal of the American Medical Association, 283, 2008-2012.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (pp. 435-455). New York, NY: Russell Sage Foundation.

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

U.S. Department of Education, Institute of Education Sciences. (2003). What Works Clearinghouse. Retrieved from http://ies.ed.gov/ncee/wwc/default.aspx

Walster, G. W., & Cleary, T. A. (1970). A proposal for a new editorial policy in the social sciences. American Statistician, 24(2), 16-19.

