Review of Educational Research
February 2020, Vol. 90, No. 1, pp. 24–46
DOI: 10.3102/0034654319877153
Article reuse guidelines: sagepub.com/journals-permissions
© 2019 AERA. http://rer.aera.net

Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review

Terri D. Pigott
Loyola University Chicago

Joshua R. Polanin
American Institutes for Research

This methodological guidance article discusses the elements of a high-quality meta-analysis that is conducted within the context of a systematic review. Meta-analysis, a set of statistical techniques for synthesizing the results of multiple studies, is used when the guiding research question focuses on a quantitative summary of study results. In this guidance article, we discuss the systematic review methods that support high-quality meta-analyses and outline best practice meta-analysis methods for describing the distribution of effect sizes in a set of eligible studies. We also provide suggestions for transparently reporting the methods and results of meta-analyses to influence practice and policy. Given the increasing use of meta-analysis for important policy decisions, the methods and results of meta-analysis should be both transparent and reproducible.

Keywords: meta-analysis, systematic review

This methodological guidance article is focused on the use of meta-analysis in a systematic review. A prior article in this series, Alexander (in press), discusses the art and science of all systematic reviews with an emphasis on the importance of the literature search, coding, and results interpretation. Systematic reviews analyze and synthesize a body of literature in a logical, transparent, and analytical manner. We use the term systematic review to refer to any effort to synthesize a body of literature using transparent and comprehensive methods, whether that literature includes studies that use quantitative or qualitative methods (Gough, Oliver, & Thomas, 2017).

Meta-analysis, a set of statistical techniques for synthesizing the results of multiple studies (Borenstein, Hedges, Higgins, & Rothstein, 2009; Higgins & Green, 2011), is used in a systematic review when the guiding research question focuses on a quantitative summary of study results. For example, Dietrichson, Bøg, Filges,

and Jørgensen (2017) conducted a systematic review to understand the effectiveness of interventions for increasing academic achievement for low-income children. They used meta-analysis to estimate the average treatment effect across included studies, how much this effect varied across studies, and what characteristics of the study and intervention were related to this treatment variation.

The term meta-analysis was first used by Gene Glass (1976) in his presidential address at the AERA (American Educational Research Association) annual meeting, though Pearson (1904) used methods to combine results from studies on the relationship between enteric fever and mortality in 1904. The 1980s was a period of rapid development of statistical methods (Cooper & Hedges, 2009) leading to the use of meta-analysis in many fields to synthesize the results of primary studies. The 1990s and 2000s brought about systematization of the systematic review, including formal guidelines and standards (Appelbaum et al., 2018; Moher, Liberati, Tetzlaff, & Altman, 2009). Recent decades have seen an increase in the use of meta-analysis (Williams, 2012), availability of approachable meta-analytic software (Polanin, Hennessey, & Tanner-Smith, 2017[1]), and multiple methodological developments (Pigott, 2012).

The reason for the raised profile of meta-analysis is its usefulness to decision makers. Meta-analysis summarizes the results of several studies, allowing researchers and policymakers to understand both the average effect across studies and its variability, thus leading to more informed decisions about important policy issues. Attention to study reproducibility in psychology (Open Science Collaboration, 2015) and medicine (Ioannidis, 2005) highlights the danger in making policy decisions based on a single study. International organizations such as the Cochrane Collaboration (www.cochrane.org) for medicine and the Campbell Collaboration (www.campbellcollaboration.org) for social interventions support the conduct and publishing of systematic reviews with meta-analysis to promote evidence-based policy decisions. The Cochrane Library alone includes thousands of systematic reviews of health-related interventions that are used to guide future research and current practice.

In this article, we provide guidance for conducting and reporting results from a high-quality meta-analysis that is part of a systematic review. For simplicity, we use the term meta-analysis in the remainder of the article. Three themes run throughout our guidance. First, like primary research studies synthesized in a meta-analysis, methods used in a meta-analysis should be fully transparent and reproducible. Transparency in methods is critical in meta-analysis given its use in policy decisions (Cordray & Morphy, 2009). Conducting a meta-analysis is a complex research task requiring the organization and analysis of a large number of studies. As we discuss in this guidance, many decisions made throughout a meta-analysis have serious consequences for the quality and validity of the results. Meta-analyses should also strive for reproducibility, given their use in important policy decisions and their influence in guiding practice and future research. Meta-analysts should provide enough detail so that readers can assess whether the methods used will lead to valid results. Since meta-analyses use aggregated data without personally identifiable information, all plans for and products from a meta-analysis should also be openly available to increase transparency and reproducibility (Stewart, Moher, & Shekelle, 2012).

A second theme is that a meta-analysis aims to summarize quantitative data from a set of studies to make claims about what we do or do not know in a given area. Researchers using meta-analysis are synthesizing the quantitative results of a sample of studies and are implicitly generalizing to all relevant studies conducted on an issue. When policymakers use meta-analysis findings to support a decision, they are assuming that the results summarize all the eligible and relevant studies on that issue. The methods used in a meta-analysis should provide evidence (as in a primary research study) that claims derived from a review are warranted by the methods and analytic results. Our recommended guidance is aimed at ensuring meta-analyses provide evidence to support claims about the distribution of effect sizes in all relevant studies on an issue. In addition, our guidance also emphasizes the importance of clearly discussing the limitations of the external validity of the meta-analysis results (Wood & Eagly, 2009). The external validity of a meta-analysis depends on the methods, participants, and other characteristics of the primary studies conducted in an area. The ability of a meta-analysis to generalize depends on how well the eligible studies themselves generalize to the contexts important to the review question.

A final theme concerns the gap between emerging methods for meta-analysis and methods used currently in the field (Tipton, Pustejovsky, & Ahmadi, 2019b). Methodological advances are occurring rapidly, particularly around the development of multivariate, multilevel methods that better reflect the true nature of meta-analytic data. We see the guidance provided in this article as the minimal requirements for best practice in meta-analysis. We anticipate that our guidance may change as new meta-analytic techniques are developed, validated, and disseminated.

This guidance article is organized into three sections. The first section discusses the systematic review methods needed to support a meta-analysis including the literature search and the screening and coding of eligible studies. The second section discusses best practice meta-analysis including the choice of effect size models, the description of the effect size distribution across studies, and the exploration of variability in effect sizes, also termed heterogeneity, through effect size modeling. The final section discusses the presentation and interpretation of results from a meta-analysis within a systematic review.

Elements of a Systematic Review Required for Meta-Analysis

This section expands on Alexander (in press) to provide specific guidance for the beginning stages of a systematic review that includes a meta-analysis. While not all systematic reviews include a meta-analysis, the guidance we provide applies to a meta-analysis conducted in the context of a systematic review. A high-quality systematic review can still be of high quality without a meta-analysis—but a high-quality meta-analysis often relies on a systematic review. In rare cases, a meta-analysis is conducted without the benefit of a thorough systematic review; for example, in situations where the goal is to synthesize effects across multiple school sites (i.e., a multisite randomized controlled trial) or where the goal is to synthesize studies from a specific set of interventions found in one lab. In this guidance article, however, we assume a thorough, comprehensive, and well-designed systematic review precedes a meta-analysis. As discussed above, the aim of a meta-analysis is to make claims about the distribution of effect sizes in a set of studies and to

make claims about effect sizes in a literature. Thus, the systematic review methods, including the methods for developing a research question, searching the literature, and screening and coding eligible studies, should support claims about the relevant studies on a topic.

Research Questions for a Meta-Analysis Focus on Summarizing Effect Sizes

A meta-analysis addresses questions of aggregation of results from a set of studies (Gough et al., 2017). A common use of meta-analysis is to estimate a treatment effect in a set of experimental studies. Meta-analysis will provide an estimate of the average treatment effect and the variation of that effect across studies. A recent example is a review of academic interventions for low-income students in elementary and middle school (Dietrichson et al., 2017). The Dietrichson et al. (2017) meta-analysis examined the average effect of interventions for low-income students, and the variation across studies in the treatment effect. Meta-analysis is also used in systematic reviews that explore the differences in groups defined by characteristics such as gender or English-language learner status. For example, Duong, Badaly, Liu, Schwartz, and McCarty (2016) were interested in the average generational differences in academic achievement among students who were first-, second-, or third-generation immigrants and how those differences might relate to country of origin and socioeconomic status.

Other meta-analyses may focus on estimating the magnitude and direction of an association, such as the correlation between exposure to cyber-bullying and academic achievement (Gardella, Fisher, & Teurbe-Tolon, 2017). The Gardella et al. (2017) study examined how the correlation between exposure to cyber-bullying and difficulties in achievement and attendance varied by gender, race/ethnicity, and age. Meta-analysis techniques also exist for the prevalence of a phenomenon across studies such as in estimating the prevalence of autism spectrum disorder in preterm infants (Agrawal, Rao, Bulsara, & Patole, 2018) and for diagnostic and prognostic test accuracy.

Quality of a Meta-Analysis Depends on a Comprehensive Search and Unbiased Screening and Coding Procedures

Systematic reviews follow three basic steps: searching the literature, screening abstracts and full-text documents, and coding included studies. When a meta-analysis is included in a systematic review, researchers need to use methods that will support generalizations to the eligible studies on a topic. Each of the three basic steps must demonstrate that the researcher has attempted to identify, screen, and code all eligible studies on a given topic.

Searching for All Eligible Studies

A meta-analysis aims to make claims about the distribution of effect sizes in a set of eligible studies. To support arguments about the treatment effect present in the literature, researchers using meta-analysis must conduct a systematic and comprehensive search that identifies all eligible studies in an area (Kugley et al., 2017). The search should be systematic in the sense that the search uses terms, strings, databases, limiters, and tools that are sensitive enough to capture all relevant studies.

The second search descriptor, comprehensive, refers to the breadth of the search. Education research is multidisciplinary, and eligible studies will be conducted in several disciplines such as economics, sociology, psychology, and social work. A comprehensive search in a meta-analysis will require terms unique to several disciplines. Searches in a meta-analysis include both online databases that index published literature as well as sources such as Google Scholar and Web of Science. Meta-analysis also includes strategies such as retrospective reference harvesting, prospective forward citation searching, and contacting prominent or active authors in the field. Finally, publication bias, the tendency of published studies to report larger, statistically significant effects, is a well-known problem (Polanin, Tanner-Smith, & Hennessy, 2016), and thus, meta-analysis searches should attempt to identify unpublished literature such as dissertations and reports from independent research firms. We also strongly suggest that reviewers document each of these multiple searches—and track them during the search process so that others can reproduce the search.

Unbiased Screening of Eligible Studies for a Meta-Analysis

Once the first step is completed, and the duplicated citations removed, the next step is to screen the collected citations, and eventually, the full-text PDF. Polanin, Pigott, Espelage, and Grotpeter (2019) provide a set of best practices that help guide the meta-analyst in conducting these processes reliably and efficiently. As they suggest, screening typically begins by creating a short screening tool used on study abstracts and titles that eliminates clearly ineligible articles such as essays or non-empirical studies. The tool should be used in conjunction, preferably, with text-mining software like Abstrackr (Wallace, Small, Brodley, Lau, & Trikalinos, 2012) that helps organize and sort abstracts based on their probability of inclusion. We strongly recommend that screening is conducted with two independent screeners to avoid the loss of eligible studies. Major guidelines for systematic review and meta-analysis all require the practice of double-screening (Centre for Reviews and Dissemination, 2009; Higgins & Deeks, 2011; Institute of Medicine, 2011; Methods Group of the Campbell Collaboration, 2016). While screening study titles and abstracts, screeners should ideally meet two to four times per month particularly for large-scale meta-analyses that identify thousands of potentially eligible studies.

Once title and abstract screening ends, reviewers will collect all the included full-text PDFs, a process often referred to as retrieval. The full-text screening process follows closely the abstract screening process: create a screening tool, screen each article, and then make a determination about whether it should be included in the coding phase. Unfortunately, no reliable text-mining application yet exists to help reviewers conduct the full-text screening process. Full-text documents should also ideally be screened by two independent reviewers for the same reasons it is recommended at the title and abstract screening stage. If this is too costly at the full-text stage, then one person should screen the article and a second person should validate the decision making and sign off on the eligibility decision.

Coding Important Moderators of Effect Size Variability

The final step prior to conducting a meta-analysis is to code each included study. As Alexander (in press) explains, coding in a systematic review allows

researchers to understand the contexts and methods used in a set of studies, and thus to understand the limitations of the external validity of the review. In a meta-analysis, coding studies serves two purposes. As in any systematic review, coding the studies highlights the contexts, participants, and methods used in relevant studies so that the reviewer understands the limits of the external validity of the review (Wood & Eagly, 2009). If eligible studies are conducted with only elementary school children, then any conclusions from the review will not apply to high school students. The second purpose of coding in a meta-analysis is for examining how effect size varies as a function of the methods, contexts, participants, and other characteristics of studies. Meta-analysts code aspects of studies to use as moderators in models of effect sizes.

In the meta-analyses typically published in *Review of Educational Research* (*RER*), researchers expect effect sizes to vary across studies, and the coding completed in a meta-analysis will focus on capturing the most likely correlates of effect size variation. A high-quality meta-analysis will provide a rationale for the coding schema in the background section of the systematic review. Dietrichson et al. (2017), for example, found that tutoring and feedback interventions were more effective than interventions without them. Duong et al.'s (2016) review revealed that the academic achievement advantage of second-generation students over recent immigrants varied with race/ethnicity and socioeconomic status. In both instances, the authors coded studies to capture items such as intervention components, and race/ethnicity and socioeconomic status of participants, codes that reflected a priori hypotheses about potential reasons for effect size variation across studies. Meta-analysts should clearly state their a priori hypotheses about why effect sizes vary across studies and provide a clear analysis plan that includes these moderators. We discuss pre-analysis plans later in this guidance.

Reliable Coding of Studies in a Meta-Analysis

We consider study coding the most time consuming and tedious aspect of the review process, typically requiring 60% of the total review time. Coding requires painstaking precision and attention to detail and must be reliable to ensure the validity of effect size models. Study coders must read lengthy reports, decipher difficult to understand jargon, and determine what choice is most reasonable. Sometimes the correct choice is not obvious; often the decision is a thoroughly discussed deduction. A good study coding process therefore requires three elements: (a) an easy to follow, concise codebook containing information pertaining to each decision; (b) a simple-to-use coding spreadsheet or database that does not require a coder to make difficult decisions about where and how to input the data; and (c) an effective teacher or leader who can guide the study coder at the beginning phase and support the coder in decision making through the coding process. Short any of these elements and study coding will persist long after it is scheduled, or worse, result in inaccurate or unreliable extracted data. A thorough codebook review is outside the scope of this article, but at a minimum, a high-quality codebook includes characteristics of the sample, intervention and comparison groups (if applicable), outcome measurements, setting, research design including

methodological quality, and effect size information. We note here that coding of study quality is essential; all existing guidance on meta-analysis requires coding of the risk of bias for experimental studies or the methodological quality of other research designs in a meta-analysis (e.g., Appelbaum et al., 2018; Moher et al., 2009). The Campbell Collaboration publishes protocols and their associated completed reviews and are a source of examples of codebooks (www.campbellcollaboration.org). Polanin (2018) also includes a codebook for a meta-analysis on the consequences of school violence.

High-Quality Meta-Analyses Should Publish a Protocol

Transparency and reproducibility are key quality indicators of a meta-analysis. Many meta-analyses focus on questions that have direct policy implications. To assess the quality of a meta-analysis, we need to understand and assess the methods used. For this guidance, we are influenced by our long-standing work with the Campbell Collaboration (www.campbellcollaboration.org). All systematic reviews published in the Campbell library include a peer-reviewed protocol, where the review team provides the plan for the review including the rationale for the review, the guiding research questions, the literature search plan with sample search terms, the screening strategy, the draft coding protocol, and strategy and the analysis plan. These steps parallel those described for the preregistration of randomized trials in the new REES, the Registry for Education Effectiveness Studies, as described by Anderson, Spybrook, and Maynard (2019).

Experienced meta-analysts know that a plan is critical given the complexity of the steps involved, and the number of small decisions made that could influence the validity of the results of a meta-analysis. A published protocol for a meta-analysis allows readers to assess the procedures for searching and screening, the coding process and its documentation, and the preanalysis plan. The goal is to be able to understand what the meta-analyst considered a priori so that a future reader can determine if any protocol deviations potentially bias the results. We emphasize the publication of a preanalysis plan given our concerns about conducting too many analyses in the search for statistical significance (Polanin & Pigott, 2015). As described above, meta-analyses include a large number of codes that can be used indiscriminately to search for statistically significant relationships. We concur with Tipton, Pustejovsky, and Ahmadi's (2019a) recommendation that all meta-analyses distinguish between planned and exploratory analyses, and that protocols include a preanalysis plan.

Many options exist for the publication of a meta-analysis protocol in addition to partnering with the Campbell Collaboration. The Open Science Framework is a free, online registry that includes various templates for publishing research. Polanin (2018) is an example of a current review on the consequences of school violence (https://osf.io/6hak7/) where the protocol and analysis plan are published. The PROSPERO (https://www.crd.york.ac.uk/prospero/) registry is hosted by the University of York and publishes protocols related to health and medicine. The open-access journal *Systematic Reviews* also publishes protocols for systematic reviews. In the future, REES may allow the registration of systematic review protocols (Maynard, personal communication, 2019).

Best Practice Meta-Analysis

After a meta-analyst has conducted a comprehensive literature search, screened for eligible studies and coded the studies, the next stage summarizes the distribution of effect sizes in eligible studies using meta-analysis techniques. Meta-analysis includes the process of extracting quantitative data from each study, accounting for missing data, synthesizing the study's effects, assessing and analyzing heterogeneity among those effects, explaining the heterogeneity, and interpreting the results. The field of meta-analysis has made great strides in the past 40 years. Below, we outline these methods while acknowledging the areas where new research to improve meta-analysis techniques is ongoing.

Compute and Report All Effect Sizes in a Meta-Analysis

Researchers conducting a meta-analysis use effect sizes as the outcome of each study's findings. As detailed in other references (Borenstein et al., 2009; Cooper, 2017; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001), the most common effect size metrics are the standardized mean difference (for continuous measures, such as in experimental or group difference studies), the correlation (for associations between two or more variables), and the odds ratio (for dichotomous outcomes). The past several years has seen the development of effect sizes and corrections for effect sizes for more complex statistical analyses. Hedges (2007, 2011) developed procedures to correct the standard errors of effect sizes computed from two-level and three-level cluster randomized trials. Aloe and Becker (2012) and Aloe and Thompson (2013) discuss effect sizes computed from the results of regressions for use in meta-analyses that are interested in estimating a treatment effect or a correlation between constructs.

We advocate for meta-analysts to compute effect sizes for all relevant outcomes measured in the study, and to use all available evidence including querying the primary study author to obtain information needed for effect size computation. Tools such as Wilson's Practical Meta-Analysis Effect Size Calculator (Wilson, n.d.) exist to assist researchers in using all available information to compute an effect size. The R program *metafor* includes a function, *escalc*, that computes a wide range of effect sizes for meta-analysis (Viechtbauer, 2010). Conversions between effect size metrics are also possible (Polanin & Snilstveit, 2016), but readers are cautioned to make a strong theoretical argument and document such decisions in the published report.

We advocate for computing effect sizes for all relevant outcomes measured for two reasons. Given the effort to conduct a large-scale meta-analysis, extracting and computing the effect sizes for all outcomes and publishing those outcomes will facilitate the update of meta-analyses when new research is conducted in an area. We also urge researchers to code all outcomes relevant to a review given the development of techniques to address dependent effect sizes within studies as we discuss in more detail below.

Meta-Analyses Should Address Missing Data and Its Consequences

Missing data can and will occur when conducting a meta-analysis (Pigott, in press). Even the best codebook and data extraction tools cannot prevent truly

missing information, whether from absent characteristics, outcomes, or entire studies. Studies differ in how they report information about a study's methods and participants, and these reporting differences contribute to missing data across studies. Therefore, the meta-analyst must deal with the missingness, and our process can be distilled into three steps: infer, initiate, and impute.

The first step is to make an inference based on what the authors stated in the article. The goal of this step is to make a truly educated assumption—not a guess but an inference. An example is the typical problem of capturing information about the sample's age. The example codebook may instruct coders to record the average age of the sample within each study, and but average age can only be extracted in 80% of the studies. One option is simply to mark the remaining 20% as missing, but additional information may be available that allows the reviewer to code the sample's age—or at least make an educated inference. Knowing that all students are in the seventh grade, for example, allows the meta-analyst to make an educated inference about the sample's age; the meta-analyst can convert the seventh-grade value into an approximate average age of 12.5 years old. A primary study author may also state that all students were recruited from a junior high school in a particular city. A search for the grade configuration in that city may reveal that the typical junior high school includes only seventh and eighth grades. This information could again be converted to an average age. We often suggest that meta-analysts capture the age of the sample in various forms like grade level or school level so it can easily be inferred.

In the case where little information exists to make an informed inference, the meta-analyst should turn to the second step in the process: initiate. By initiate, we mean initiate contact with the primary study author to ask directly for the missing information. We expect about 40% of authors to respond, and about 20% to 25% of authors provide the requested information (Polanin & Terzian, 2019). Whenever an author query is sent via email, we also suggest sending a data sharing agreement (Polanin & Williams, 2016). Finally, before sending the request, we suggest reviewing other published evidence syntheses (e.g., the What Works Clearinghouse website or other meta-analyses) that make study information publicly available. This will reduce the burden on the primary study author as well as reducing the burden on the project team of sending and tracking an author query.

Failing both the inference and initiation steps, the meta-analyst is forced to make some difficult, albeit all too common, decisions on how to handle the missing data. The options range from the simple, complete case analysis (i.e., listwise deletion), to the complex, such as using a method for missing data such as maximum likelihood or multiple imputation (Pigott, in press). Methods for missing data are in rapid development in the statistical literature (Audigier et al., 2018; Enders, Mistler, & Keller, 2016; Grund, Lüdtke, & Robitzsch, 2018), but have not been studied extensively in the context of meta-analysis. The exception is the application of full maximum likelihood methods in the context of meta-analysis structural equation models (Jak & Cheung, 2018) although this method has not been applied to other types of meta-analyses.

How the meta-analyst should handle truly missing data, either through a simple or complex analytic technique, can be answered primarily by answering two questions: (a) What information is missing? (b) How much data are available?

The answer to the latter question provides a straightforward solution: to conduct more sophisticated missing data analyses the meta-analyst must have a sufficiently large data set and be willing to make the assumption of missing-at-random data. We expect at least 20 to 40 studies and perhaps many more effect sizes be available, for example. Meta-analysts with fewer studies or effect sizes available may be forced to use less sophisticated methods like complete case analysis.

The former question, what information is missing, asks whether the review is missing effect size or moderator data. We typically do not suggest imputing effect sizes, regardless of the size of the review. This is akin to the imputation of missing outcome data in primary studies; given the debate about this imputation practice, we understand opinions may vary on whether this is an appropriate practice. Therefore, we note that it is statistically possible to impute missing effect size data, but we urge meta-analysts to consider carefully the implications of this choice and the impact imputed effect size data has on the overall results. Missing moderator data, however, is typically where we advocate for multiple imputation (Pigott, in press).

Should one use a multiple imputation technique, the steps advocated by primary researchers should be followed. The meta-analyst should run a complete case analysis and compare the results with the multiple imputation analysis. Comparing a complete case analysis with the results from multiple imputation allows the assessment of the robustness of the results given the presence of missing data. Any meta-analysis of sufficient size should be using some version of a modern missing data technique—the options for conducting the analyses are being developed rapidly though more research is needed on the application of missing data methods to meta-analysis.

Meta-Analyses Should Report the Mean Effect Size and Its Variability

One key goal of a meta-analysis is to estimate the average effect size and its variability across studies (whether a treatment effect, a correlation, or an odds ratio). In educational research, we expect that studies will vary in their effect size. What makes educational research both challenging and exciting is the important ways teaching and learning vary depending on the context and the students and teachers in that setting. Thus, all high-quality meta-analyses focus on both the average effect size and the variation across studies in that effect. When possible, a meta-analysis should plan for the exploration of potential reasons for the variation in effect size through the use of effect size models to be discussed below.

Providing a Rationale for the Use of Fixed- or Random-Effects Models

A high-quality meta-analysis should clearly state whether a fixed effects or a random effects model will be used. Borenstein et al. (2009) provided a discussion of these models. Briefly, a fixed-effect analysis assumes that all studies are estimating a common effect size, a situation that may occur when all the studies in a meta-analysis are close replications of one another. For example, a meta-analysis narrowly focused on a single type of intervention where the studies' participants are highly similar might be a case for using a fixed-effect model. In general, the systematic reviews with meta-analyses published in *RER* tend to focus on broader questions where we expect variation across studies. Random-effects models in

meta-analysis are used when eligible studies use a range of methods, samples, and settings. High-quality meta-analyses state a priori the model that will be used based on the assumptions researchers make about the studies in the meta-analysis. For most systematic reviews that include a meta-analysis, we recommend the use of random effects models for the effect size analysis.

Describing the Distribution of Effect Sizes

We recommend that a high-quality meta-analysis describe the included studies and the distribution of their effect sizes. As Alexander (in press) describes, the systematic review should provide an overall description of the methods, samples, and other important characteristics of the eligible studies. The description of the "landscape" of the included studies allows readers to understand the evidence base for a given research question, and the gaps that might exist in that evidence. In a meta-analysis, the next important step is to describe the distribution of the effect sizes from the included studies. One useful graph is a forest plot that presents the effect sizes and their 95% confidence interval from each study. This plot allows readers to visualize the overall pattern of results including the overall mean and variation around that mean (see, e.g., De La Rue, Polanin, Espelage, & Pigott, 2017[2]). If there are too many effect sizes for a forest plot (more than thirty effect sizes), meta-analysts should provide the overall mean effect size, the confidence interval for that mean, and the appropriate measure of heterogeneity for the chosen model for each outcome in the meta-analysis. For random-effects models, the overall mean effect size and its 95% prediction interval should be reported (Borenstein et al., 2009). For transparency, reviewers should state the software used for the analysis, and the method used to estimate the random effects variance. Our recommendations are consistent with guidelines for reporting meta-analysis from many fields (Appelbaum et al., 2018; Centre for Reviews and Dissemination, 2009; Higgins & Green, 2011; Institute of Medicine, 2011; Moher et al., 2009; Shea et al., 2017).

Exploring the Impact of Publication Bias

The current discussions about quality of research and reproducibility has heightened researchers' awareness of publication bias. Many researchers have documented the existence of publication bias in medicine and psychology (Dickersin, 1990; Ferguson & Brannick, 2012) as well as in education (Polanin et al., 2016). High-quality meta-analyses should discuss the strategies used to identify unpublished studies for inclusion and should explore whether results are sensitive to publication bias. The most commonly used methods are visual inspection of a funnel plot and Egger's test of funnel plot symmetry (Sterne, Egger, & Moher, 2011). These methods, however, do not perform well particularly when there is heterogeneity among effect sizes (Macaskill, Walter, & Irwig, 2001; Pustejovsky & Rodgers, 2019). It is increasingly common for meta-analysts to use selection modeling strategies to explore the robustness of meta-analysis to publication bias (Citkowicz & Vevea, 2017; Vevea & Hedges, 1995). Selection models provide meta-analysts with an estimate of the presence of selective reporting bias by explicitly modeling the process by which studies are chosen for inclusion in a publication. Researchers are actively exploring methods for examining

publication bias for meta-analyses with multiple effect sizes, a situation not currently addressed by current methods for assessing publication bias.

Meta-Analyses Should Use Methods Appropriate for Dependent Effect Sizes

Meta-analysts must also decide how multiple effect sizes within studies will be addressed in the analysis. Primary studies report on a range of outcomes resulting in multiple effect sizes measured on the same study sample. Even in cases with independent samples within studies, researchers may argue that effect sizes computed within any study are related or highly correlated due to their occurrence in the same research project. In the past, meta-analysts used a shifting-units-of-analysis approach to deal with multiple effect sizes within studies (Patall, Cooper, & Robinson, 2008). This approach entails conducting a separate analysis for each construct in the meta-analysis. For example, if a meta-analyst was interested in all academic outcomes for a given intervention, they would conduct one analysis for all reading outcomes and another for mathematics outcomes.

Computing separate analyses by outcome increases the probability that a meta-analysis will be subject to multiplicity problems, or issues with conducting too many statistical tests (Polanin & Pigott, 2015). In recent years, standard best practice for meta-analysis is the use of robust variance estimation for handling multiple effect sizes within studies (Hedges, Tipton, & Johnson, 2010; Tanner-Smith, Tipton, & Polanin, 2016). The technique provides robust variance estimates for the mean effect size and for parameters of effect size models under conditions where model misspecification may occur, such as in the case of dependent effect sizes. Instead of conducting separate analyses, for example, for reading and mathematics outcomes, the meta-analysis will conduct an analysis that includes all academic outcomes within studies.

Programs for implementing robust variance estimation exist in STATA and R (Tanner-Smith et al., 2016). The program *robumeta* (Fisher, Tipton, & Zhipeng, 2017) implements robust variance estimation for meta-analysis. For more general uses, the R package *clubSandwich* (Pustejovsky, 2019) will provide the robust variance estimates for an effect size model from the R package *metafor* (Viechtbauer, 2010). Polanin et al. (2017) provide an overview of many of the available R packages that conduct meta-analyses.

We currently recommend that researchers use random-effects models with robust variance estimation for estimating the mean effect size and its confidence interval. We urge researchers to limit the use of shifting-units-of-analysis when studies report multiple effect sizes per study. In some cases, a researcher may argue that some outcomes are conceptually distinct, such as reading and mathematics outcomes, and these outcomes require separate analyses. When studies report multiple effect sizes for a single construct such as multiple mathematics measures or assessments of anxiety, the researcher should use robust variance estimation in a single analysis of an overarching construct.

Methods are also currently in development for using multilevel models for dependent effect sizes (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2015). Meta-analysis of single-case studies (Moeyaert et al., 2017) and meta-analysis of structural equation models (Cheung, 2015) also account for the dependent nature of the meta-analytic data using multilevel

models of effect sizes. We expect that multilevel models for effect size data will become more commonly used in meta-analysis (Tipton et al., 2019a).

Meta-Analysis Should Explore Heterogeneity Among Effect Sizes

As discussed above, meta-analysis reports should describe the eligible studies in tables (Alexander, in press), present the mean effect size and its confidence or prediction interval and when possible, include a forest plot of the effect sizes in the review. A second critical goal of the meta-analysis is the examination of the heterogeneity of the effect sizes in the review. Researchers use effect size moderator models to explore how study characteristics such as the sample, methods, quality of study design, and/or settings, for example, are associated with variation in effect sizes across studies. As discussed in a prior section, meta-analyses in *RER* typically anticipate that the effect sizes from a set of studies will vary, and that the variability is associated with the differences among studies in their interventions, their participants, their contexts, and their methods. High-quality meta-analyses will include models of effect size that examine how characteristics of studies are associated with effect size variability.

At the advent of meta-analysis, researchers often explored these associations using analogues to one-way analysis of variance models, comparing the average effect size within groups defined by a categorical model such as community context (urban, rural, suburban) to test for significant differences. Researchers would conduct a series of these one-way analysis of variance models, examining the effect size differences for each potential moderator separately (Tipton et al., 2019b).

Current best practice in all statistical modeling is to use multiple moderators in a single model to reduce difficulties caused by confounding moderators. If there is an association between two moderators, such as if all the high schools in the study are located in urban settings, we cannot be sure whether the effect size difference we observe is associated with grade level or setting. Thus, we recommend researchers conducting moderator analyses should use meta-regression with robust variance estimation.

Use of Meta-Regression to Examine Effect Size Variability

Meta-regression can accommodate both continuous and categorical moderators with appropriate dummy coding schemas. As in any application of regression, meta-regression can limit problems with confounding and should include moderators that are considered control variables. Meta-analytic results, for example, sometimes demonstrate that research design quality is associated with variation in effect sizes. As a result, a moderator related to study quality should be included in a meta-regression model (Tipton et al., 2019a). When a meta-analysis includes multiple outcomes that measure different constructs, the meta-regression could include a fixed effect for each construct to test whether effect size moderators have differing associations depending on the construct of the effect size.

Meta-analyses published in *RER* tend to focus on broad research questions and thus typically include a sufficient number of studies for the use of meta-regression. Though Hedges and Pigott (2004) discuss power analysis for effect size models, research has not been conducted on power for moderator models with

multiple effect sizes. To guard against conducting too many meta-regressions in a search for significant results, we recommend the preregistration of analysis plans for moderator models in a meta-analysis. Tipton et al. (2019a) discuss the need for meta-analysts to distinguish between confirmatory and exploratory analyses. Reviewers have a priori assumptions about important moderators of effect size, and these analyses should be planned in advance. Other analyses suggested by the data collected should be clearly labeled as exploratory. Analysis plans might be included in a preregistered protocol or as a separate preanalysis plan (Anderson et al., 2019). Polanin (2018) provides an example of a preanalysis plan for a meta-analysis.

In summary, current best practice in meta-analysis is the use of meta-regression using robust variance estimation with multiple moderators and dependent effect sizes. However, research on the most appropriate models for exploring effect size heterogeneity is moving toward models that more accurately reflect the underlying structure of meta-analytic data. Tipton et al. (2019a) strongly advocate for the use of multivariate effect size models that include estimates of the covariance among effect sizes within each study. These models reflect the multilevel and multivariate nature of the data from a meta-analysis where effect sizes are nested within studies. Our current recommendation for the use of meta-regression with robust variance estimation should be considered a minimal requirement for best practice; in our own work, we are moving toward the use of multivariate models that reflect the true structure of meta-analytic data.

Meta-Analysis Should Provide a Clear Description of Methods and Interpretation of Results for Multiple Audiences

Education practitioners, policymakers, and decision makers rely on the results of meta-analyses to inform programmatic and policy decisions. Researchers rely on meta-analyses to inform current understandings of critical research questions and identify gaps in their extant knowledge. All audiences are looking for clear, actionable recommendations from the findings of a meta-analysis. It is therefore the responsibility of the meta-analyst to (a) make responsible recommendations and (b) provide enough information to support arguments about such recommendations and actions. We suggest several steps that should drive the interpretation and dissemination of meta-analytic findings. As in any primary study, a meta-analysis should include a clear description of the methods used. A meta-analysis should transparently report any potential limitations of the results such as conflicts of interest and any gaps in the evidence base that affect the applicability of the results. A meta-analysis should also provide enough information to reproduce the results, including publishing the data collected and the code used for the analysis.

The interpretation of results for multiple audiences is also a critical part of a high-quality meta-analysis. The meta-analyst should discuss how the meta-analysis results relate to the current understanding of the literature as outlined in the background section. In addition, the meta-analysis should include an interpretation of the results suitable for multiple audiences. These points are discussed in more detail below.

Reporting the Meta-Analytic Methods Used

Describing the Meta-Analysis Methods

The meta-analyst should aim to transparently report the processes conducted in the review. The MARS (Appelbaum et al., 2018) and PRISMA (Moher et al., 2009) guidelines each provide a helpful framework for meta-analysts to follow when reporting the results of the study. The PRISMA guidelines provide a flow chart of the review process that should also be included at the end of each review. The meta-analyst should report the major decisions made by the research team, including decisions related to: inclusion criteria; search terms, string, databases, and returned results; screening processes including how many screeners participated at each stage; coding processes, including the codebook and how many people participated; analysis decisions about effect sizes, transformations, and data cleaning; and synthesis methods. When the researchers have preregistered their protocol, the completed review should provide details about areas that deviated from the original plan. The goal is to allow future researchers and meta-analyses to understand, and potentially reproduce, what was done so that the results may be interpreted appropriately.

Reporting Conflicts of Interest

The meta-analyst should be clear and upfront about potential personal or financial conflicts of interest (COIs). COIs are easier to spot in primary research, relative to meta-analysis, because of the study's funding source or intervention development. A financial COI in primary research occurs, for example, when an education program developer evaluates the effectiveness of her own program. COIs are more difficult to identify in meta-analyses because it can be hard to pinpoint which studies the meta-analyst, or someone from the team, have potential COIs. A financial COI might occur in a meta-analysis when the members of the meta-analytic team were involved in one or more of the studies included.

To combat the bias that can occur from potential COIs, we strongly suggest that the meta-analytic team clearly identify their potential COIs in a protocol and the steps taken to abate the potential bias. If any leader or member of the meta-analysis team has any involvement in any primary study included, financial or otherwise, the meta-analyst should describe the steps taken to ensure that the members of the team who worked on the primary study included did not code or bias the results of the coding process. In addition, these team members and studies should be identified in the limitations or conflict of interest sections.

We are concerned with COIs because it is surprisingly easy to bias the results through involvement of primary study team members. The meta-analytic team makes numerous decisions that may be clouded by COIs at every stage of the review. Search terms can be added to ensure that certain studies are found (or not). Selection criteria can be altered to enable studies remain eligible for inclusion. Coding items can be changed to reflect information viewed as more favorable to the study. And, of course, researchers could alter the computation of an effect size, its magnitude, and its weight in the model. It is therefore the responsibility of the meta-analyst to provide a reasonable plan for combating COIs and limit the potential for bias.

Publishing Meta-Analysis Data and Analysis Code

An additional step the meta-analyst should take for transparent and reproducible results is the dissemination of all collected information as well as the statistical code used to clean and analyze the meta-analytic data. We recognize that conducting a meta-analysis is costly and time-consuming, and some meta-analysts may be hesitant to publish their collected data. However, publishing the data from a meta-analysis allows others to reproduce the analysis, and facilitates future efforts to update the results when new research is conducted. The same can be said for publishing the statistical code, especially given the rise of open-source software options like R or Python. The meta-analyst should strive to publish statistical code that, with a few clicks, cleans the original dataset, conducts the overall analyses, runs moderator or sensitivity analyses, and reports the final tables. This information will help inform peer-reviewers of the steps taken as well as future meta-analysts interested in using the results. We therefore believe strongly that meta-analysts have a responsibility to the field to document their collected information not only for the transparency of results, but the reproducibility and *reuse* of the data at a future time point. These data sets could be stored in a depository such as Interuniversity Consortium for Political and Social Research or with the preregistered protocol and final report on Open Science Framework or PROSPERO.

Interpreting the Meta-Analysis Results and Their Limitations

Discussing the Applicability of Meta-Analysis Results to Other Contexts

A meta-analysis should follow Alexander's (in press) advice to provide a summary table describing the included studies' characteristics to highlight the variation in participants, constructs, research designs, and intervention components used in a research area. The meta-analysis should describe any gaps in the evidence base that may limit the applicability of results to important contexts. For example, a meta-analyst may observe gaps in the types of individuals studied. We can imagine a meta-analysis where the samples have been limited to schools in neighborhoods with high socioeconomic status or where students primarily identify as white. Understanding the sample of studies in this manner is different from attempting to explain heterogeneity of treatment effects—here the goal is simply to understand how the characteristics of the studies limit the external validity of the results (Wood & Eagly, 2009).

Researchers are also using evidence and gap maps to more formally examine the limitations in external validity of studies in a given area. Evidence and gap maps (EGMs) illustrate specific areas where evidence exists and where it is lacking. The International Initiative for Impact Evaluation (3ie) has developed systems for creating EGMs (Snilstveit, Vojtkova, Bhavsar, Stevenson, & Gaarder, 2016). One example is a map of studies conducted in low- and middle-income countries on interventions to develop noncognitive, life skills in youth (http://gapmaps.3ieimpact.org/evidence-maps/youth-transferable-skills-evidence-gap-map). The Campbell Collaboration has recently published guidelines for the conduct and reporting of EGMs (White et al., 2018). The goal of an EGM is to visualize all relevant studies in one table, where the rows and columns of the table represent key characteristics of the studies included in the review. A typical EGM includes interventions or intervention components as the rows and outcomes of

interest as columns for a given topic of interest. Each cell of the table reports the number of studies that have been conducted using a particular intervention or intervention component and measuring a specific outcome. Blank cells in the table represent "gaps" in the literature. Additional characteristics of the studies can simultaneously be represented in the table as well. The meta-analyst, using the EGM's results, can provide a visualization of the studies conducted in an area, highlighting where there is evidence available, and where more research is needed.

Interpreting Results

The last objective for the meta-analysis is perhaps the most difficult and requires expert consultation: interpretation of the results. Several considerations can and should be given to interpreting the results, including (a) effect size transformation, (b) moderator or covariate description, and (c) audience translation.

Transforming effect sizes to interpretable metrics. An effect size, while widely used among researchers and the primary tool of the meta-analyst, is not a particularly intuitive measure. Even for researchers and meta-analysts who use the effect sizes regularly, the average effect size may be difficult to interpret. We therefore suggest placing the effect size in greater context and transforming it, if possible, to a natural metric. Placing the effect size into an understandable context may involve "benchmarking," where the magnitude of the effect size estimated by the meta-analysis is discussed comparatively as large or small relative to estimates from other primary studies. This is especially useful for researchers concerned with comparing particular samples, relationships, or treatments.

We also suggest transforming effect sizes to a natural metric, particularly when the underlying metric is test scores, final exams, or grade point averages. In these cases, a meta-analyst may simply multiply the effect size by the average standard deviation (or a standard deviation taken from a large sample) to "back-transform" the effect into an interpretable metric. For example, if a test has a standard deviation of 10 points, an effect size of 1.0 indicates that the experimental group would, on average, score 10 points higher than the control group. Should the underlying metric of interest be dichotomous, for example, graduation or dropout, the meta-analyst should consider transforming it back to a proportion from the odds ratio. One other option is to place the effect size in the scale of the WWC's Improvement Index, which is a version of Cohen's U3, and indicates how well a member of the comparison group, in percentile rank, would perform had they participated in the intervention. Using any of the suggested applications will provide readers with a better understanding of the effect size. Baird and Pane (2019)[3] discuss in more detail issues with transforming effect sizes to interpretable metrics.

Interpreting the effect size moderator results. Similar to interpreting the effect size, meta-analysts should also describe the covariate or moderator analysis results. The caveat here, of course, is that the results of these analyses always constitute correlations among study characteristics and the effect size, so the description of the results should avoid causal language. As described earlier, we

urge researchers to use meta-regression to examine effect size heterogeneity. When describing the results of a meta-regression, the meta-analyst must describe the results in more detail than simply stating a "statistically significant relationship" exists. What does the relationship mean? How does a one standard deviation increase in the covariate, say in the percentage of males included in the sample, relate to the treatment effect? What makes one intervention, developed and then implemented by the same research team, differ from an intervention studied independently by a research organization? Similar to primary research, meta-analysts should use modern analytic principles such as stating primary and secondary hypotheses, controlling for confounding variables, and adjusting for Type 1 error rates *before* making claims about the findings. High-quality meta-analyses will follow the same practices as primary researchers when interpreting the statistical analyses in a quantitative study.

Interpreting the results for multiple audiences. The final aspect of interpretation is to describe the results in plain language for various audiences other than researchers. A meta-analysis could, for example, include an "Implications for Practitioners" section or write a plain language summary that accompanies the published manuscript as an online appendix. The Campbell Collaboration includes plain language summaries for its published reviews. A meta-analyst should translate results for individuals making decisions about policy and practice. Why does it matter that the average treatment effect was nonstatistically significant? What does the large or small heterogeneity of treatment effects mean for administrators trying to decide which intervention to purchase and use for their school district? How can the gaps in the evidence inform what decisions policymakers make in terms of what research programming to fund? These questions, and many more, should be included in the results or discussion sections of any published meta-analysis.

Summary

This article provides methodological guidance for high-quality meta-analysis. One key aim of a meta-analysis is to make inferences about the distribution of effect sizes across a set of studies, whether those effect sizes represent treatment effects, group differences, or correlations. A high-quality meta-analysis, like any high-quality primary study, must provide a strong argument that the methods and analytic strategy can support claims about the distribution of effect sizes across studies and thus about the quantitative results in a given literature base. Meta-analysts should provide a preregistered protocol and analysis plan, provide the code used to clean and analyze the effect size data, and publish the data collected from studies for others to reanalyze or to use to update the review. These steps will provide readers with evidence to assess the potential bias in the results of the review.

We also recognize the rapid development of methods for meta-analysis that more accurately reflect the multivariate and multilevel nature of effect size data. The guidance we provide in this article represents what we consider as the minimal requirements for a high-quality meta-analysis. We encourage researchers

interested in keeping up to date on meta-analysis methods to seek out workshops at the AERA annual conference, the Campbell Collaboration, or from one of the many meta-analysis methodologists who conduct trainings regularly.

We also believe that the role of researchers using systematic review and meta-analysis is to produce both high-quality analyses and to interpret those results in ways accessible to a wide audience. A high-quality systematic review and meta-analysis is difficult and time-consuming to produce; it is worth the effort to ensure that the results inform future research and policymaking through clear discussion of the results. Researchers should consider preparing different summaries of their review tailored to their audience of researchers, policymakers, and practitioners.

ORCID iD

Terri D. Pigott https://orcid.org/0000-0002-5976-246X

Notes

This project was supported by Award No. R305B170019, awarded by the Institute of Education Sciences, National Center for Education Research, US Department of Education.

[1]Polanin, Hennessey, and Tanner-Smith (2017) is freely accessible through https://journals.sagepub.com/stoken/default+domain/EXK7IDYSYAS4CBSXQS9G/full.

[2]De La Rue, Polanin, Espelage, and Pigott (2017) is freely accessible through https://journals.sagepub.com/stoken/default+domain/5NUIKZYWJM9W9FBGGPUJ/full.

[3]Baird and Pane (2019) is freely accessible through https://journals.sagepub.com/stoken/default+domain/CWQNUQMNG8JP9YZJQTVK/full.

References

Agrawal, S., Rao, S. C., Bulsara, M. K., & Patole, S. K. (2018). Prevalence of autism spectrum disorder in preterm infants: A meta-analysis. *Pediatrics*, *142*(3), e20180134. doi:10.1542/peds.2018-0134

Alexander, P. (in press). The art and science of quality systematic reviews. *Review of Educational Research*.

Aloe, A. M., & Becker, B. J. (2012). An effect size for regression predictors in meta-analysis. *Journal of Educational and Behavioral Statistics*, *37*, 278–297. doi:10.3102/1076998610396901

Aloe, A. M., & Thompson, C. G. (2013). The synthesis of partial effect sizes. *Journal of the Society for Social Work and Research*, *4*, 390–405. doi:10.5243/jsswr.2013.24

Anderson, D., Spybrook, J., & Maynard, R. (2019). REES: A registry of efficacy and effectiveness studies in education. *Educational Researcher*, *48*, 45–50. doi:10.3102/0013189X18810513

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*, 3–25. doi:10.1037/amp0000191

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., . . . Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, *33*, 160–183. doi:10.1214/18-STS646

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher*, *48*, 217–228. doi:10.3102/0013189X19848729

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

Centre for Reviews and Dissemination. (2009). *Systematic reviews: CRD's guidance for undertaking reviews in health care*. York, England: University of York.

Cheung, M. W. L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*, 1521.

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods*, *22*(1), 28–41. doi:10.1037/met0000119

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3–18). New York, NY: Russell Sage Foundation.

Cooper, L. V., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). 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* (2nd ed., pp. 473–496). New York, NY: Russell Sage Foundation.

De La Rue, L., Polanin, J. R., Espelage, D. L., & Pigott, T. D. (2017). School-based interventions to reduce dating and sexual violence: A systematic review. *Review of Educational Research*, *87*, 7–34. doi:10.3102/0034654316632061

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association*, *263*, 1385–1389.

Dietrichson, J., Bøg, M., Filges, T., & Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, *87*, 243–282. doi:10.3102/0034654316687036

Duong, M. T., Badaly, D., Liu, F. F., Schwartz, D., & McCarty, C. A. (2016). Generational differences in academic achievement among immigrant youths: A meta-analytic review. *Review of Educational Research*, *86*, 3–41. doi:10.3102/0034654315577680

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*, 222–240. doi:10.1037/met0000063

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*, 120–128. doi:10.1037/a0024445

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). robumeta (Version 2.0) [Computer software]. Retrieved from https://cran.r-project.org/package=robumeta

Gardella, J. H., Fisher, B. W., & Teurbe-Tolon, A. R. (2017). A systematic review and meta-analysis of cyber-victimization and educational outcomes for adolescents. *Review of Educational Research*, *87*, 283–308. doi:10.3102/0034654316689136

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, *5*(10), 3–8. doi:10.2307/1174772

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). London, England: Sage.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*, 111–149. doi:10.1177/1094428117703686

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370. doi:10.3102/1076998606298043

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*, 346–380. doi:10.3102/1076998610376617

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*, 426–445. doi:10.1037/1082-989X.9.4.426

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. doi:10.1002/jrsm.5

Higgins, J. P. T., & Deeks, J. J. (2011). Selecting studies and collecting data. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions Version 5.1.0* (pp. 151–187). Hoboken, NJ: Wiley.

Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). Hoboken, NJ: Wiley.

Institute of Medicine. (2011). *Finding what works in health care: Standards for systematic reviews*. Washington, DC: National Academies Press. doi:10.17226/13059

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Jak, S., & Cheung, M. W.-L. (2018). Accounting for missing correlation coefficients in fixed-effects MASEM. *Multivariate Behavioral Research*, *53*(1), 1–14. doi:10.1080/00273171.2017.1375886

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jorgensen, A.-M. K., & Thune, K. (2017). *Searching for studies: A guide to informational retrieval for Campbell systematic reviews—Campbell Methods Guide 1*. Retrieved from http://www.campbellcollaboration.org/images/Campbell_Methods_Guides_Information_Retrieval.pdf

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*, 641–654. doi:10.1002/sim.698

Methods Group of the Campbell Collaboration. (2016). *Methodological expectations of Campbell Collaboration intervention reviews: Conduct standards*. Retrieved from https://www.campbellcollaboration.org/library/campbell-methods-conduct-standards.html

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, *20*, 559–572.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. doi:10.1371/journal.pmed.1000097

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716–aac4716. doi:10.1126/science.aac4716

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). Parent involvement in homework: A research synthesis. *Review of Educational Research*, *78*, 1039–1101.

Pearson, K. (1904). Report on certain enteric fever inoculation studies. *British Medical Journal*, *3*, 1243–1246.

Pigott, T. D. (2012). *Advances in meta-analysis*. New York, NY: Springer.

Pigott, T. D. (in press). Missing data in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed.). New York, NY: Russell Sage Foundation.

Polanin, J. R. (2018). The consequences of school violence: A systematic review and meta-analysis, *Open Science Framework*, Retrieved from https://osf.io/6hak7/

Polanin, J. R., Hennessy, E. A., & Tanner-Smith, E. E. (2017). A review of meta-analysis packages in R. *Journal of Educational and Behavioral Statistics*, *42*, 206–242. doi:10.3102/1076998616674315

Polanin, J. R., & Pigott, T. D. (2015). The use of meta-analytic statistical significance testing. *Research Synthesis Methods*, *6*, 63–73. doi:10.1002/jrsm.1124

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, *10*, 330–342. doi:10.1002/jrsm.1354

Polanin, J. R., & Snilstveit, B. (2016). *Campbell methods policy note on converting between effect sizes* (Version 1.1). Oslo, Norway: Campbell Collaboration. doi:10.4073/cmpn.2016.3

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*, 207–236. doi:10.3102/0034654315582067

Polanin, J. R., & Terzian, M. (2019). A data-sharing agreement helps to increase researchers' willingness to share primary data: Results from a randomized controlled trial. *Journal of Clinical Epidemiology*, *106*, 60–69. doi:10.1016/j.jclinepi.2018.10.006

Polanin, J. R., & Williams, R. T. (2016). Overcoming obstacles in obtaining individual participant data for meta-analysis. *Research Synthesis Methods*, *7*, 333–341. doi:10.1002/jrsm.1208

Pustejovsky, J. (2019). clubSandwich (Version 0.3.3) [Computer software]. Retrieved from https://cran.r-project.org/package=clubSandwich

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*, 57–71. doi:10.1002/jrsm.1332

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., . . . Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *British Medical Journal*, *358*, j4008. doi:10.1136/bmj.j4008

Snilstveit, B., Vojtkova, M., Bhavsar, A., Stevenson, J., & Gaarder, M. (2016). Evidence & gap maps: A tool for promoting evidence informed policy and strategic research agendas. *Journal of Clinical Epidemiology*, *79*, 120–129.

Sterne, J. A., Egger, M., & Moher, D. (2011). Addressing reporting biases. In *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). Retrieved from www.handbook.cochrane.org

Stewart, L., Moher, D., & Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. *Systematic Reviews*, *1*(1), 7. doi:10.1186/2046-4053-1-7

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*, 85–112. doi:10.1007/s40865-016-0026-5

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019a). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*, 161–179. doi:10.1002/jrsm.1338

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019b). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*, 180–194. doi:10.1002/jrsm.1339

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

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*, 419–435. doi:10.1007/BF02294384

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the ACM International Health Informatics Symposium* (pp. 819–824). New York, NY: ACM.

White, H., Welch, V., Pigott, T., Marshall, Z., Snilstveit, B., Mathew, C., & Littell, J. (2018). *Campbell Collaboration checklist for evidence and gap maps: Conduct standards*. Retrieved from https://www.campbellcollaboration.org/library/campbell-collaboration-checklist-for-evidence-and-gap-maps-conduct-standards.htm

Williams, R. T. (2012). *Using robust standard errors to combine multiple regression estimates with meta-analysis* (Doctoral dissertation; Order No. 3526372). Available from ProQuest Dissertations & Theses Global. (1081489864)

Wilson, D. B. (n.d.). *Practical meta-analysis effect size calculator* [Online calculator]. Retrieved from https://www.campbellcollaboration.org/research-resources/research-for-resources/effect-size-calculator.html

Wood, W., & Eagly, A. H. (2009). Advantages of certainty and uncertainty. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 455–472). New York, NY: Russell Sage Foundation.

Authors

TERRI D. PIGOTT is a professor in the School of Public Health and College of Education and Human Development, Georgia State University, PO Box 3995, Atlanta, GA 30302-3995, USA; email: *tpigott@gsu.edu*. Her research focuses on methods for meta-analysis.

JOSHUA R. POLANIN is principal research in the Research and Evaluation program at American Institutes for Research, 1000 Thomas Jefferson St. NW, Room 3271, Washington, DC 20007, USA; email: *jpolanin@air.org*. He has extensive experience in quantitative methodology, and has led or co-led more than 15 peer-reviewed, published meta-analyses.