Meta-analytic Reviews in the Organizational Sciences: Two Meta-analytic Schools on the Way to MARS (the Meta-analytic Reporting Standards)

Sven Kepes · Michael A. McDaniel ·
Michael T. Brannick · George C. Banks

Abstract

Purpose The purpose of this study was to review the Meta-Analysis Reporting Standards (MARS) of the American Psychological Association (APA) and highlight opportunities for improvement of meta-analytic reviews in the organizational sciences.

Design/Methodology/Approach The paper reviews MARS, describes "best" meta-analytic practices across two schools of meta-analysis, and shows how implementing such practices helps achieve the aims set forth in MARS. Examples of best practices are provided to aid readers in finding models for their own research.

Implications/Value Meta-analytic reviews are a primary avenue for the accumulation of knowledge in the organizational sciences as well as many other areas of science. Unfortunately, many meta-analytic reviews in the organizational sciences do not fully follow professional guidelines and standards as closely as they should. Such deviations from best practice undermine the transparency and replicability of the reviews and thus their usefulness for the generation of cumulative knowledge and evidence-based practice. This study shows how implementing "best" meta-analytic practices helps to achieve the aims set forth in MARS. Although the paper is written primarily for organizational scientists, the paper's recommendations are not limited to any particular scientific domain.

Keywords Systematic review · Psychometric meta-analysis · Hedges and Olkin tradition of meta-analysis · Organizational sciences · Medical sciences

S. Kepes (✉) · M. A. McDaniel
School of Business, Virginia Commonwealth University, 301 West Main Street, P.O. Box 844000, Richmond, VA 23284, USA
e-mail: skepes@vcu.edu

M. T. Brannick
University of South Florida, Tampa, FL, USA

G. C. Banks
Longwood University, Farmville, VA, USA

Systematic reviews using meta-analysis are a primary avenue for the development of cumulative knowledge in the organizational sciences. They affect new theoretical developments (Viswesvaran and Sanchez 1998), direct research agendas (Cooper and Hedges 2009; Hunter and Schmidt 2004), and provide organizations with evidence regarding the effectiveness of interventions and people management practices (Briner and Rousseau 2011; Le et al. 2007). Thus, such reviews guide the organizational sciences toward evidence-based practice (Briner and Denyer 2012; Briner and Rousseau 2011).

However, meta-analytic reviews in the organizational sciences often fall short of their potential, which can undermine their usefulness (e.g., Aytug et al. 2011; Briner and Denyer 2012; Geyskens et al. 2009). Recently, the *American Psychological Association* (APA 2008, 2010) issued their Meta-Analysis Reporting Standards (MARS). These standards facilitate two objectives. First, MARS is a vehicle by which psychology-related disciplines, including the organizational sciences, can share common meta-analytic practices across disciplines. Second, MARS allows for discipline-specific priorities. Some methodological aspects of a meta-analytic review may be more critical to one discipline than another. Thus, MARS calls for a common structure while allowing for some flexibility.

In this article, we describe how to achieve several aims detailed in MARS through the integration of two schools of meta-analysis. Those schools are (a) the approach typically used in the organizational sciences (i.e., the traditions, procedures, and methods associated with psychometric meta-analysis; Hunter and Schmidt 2004) and (b) the approach used in other areas in the social and medical sciences (e.g., Borenstein et al. 2009; Hedges and Olkin 1985; Hedges and Vevea 1998; Raudenbush 1994). When appropriate, we integrate the two approaches and illustrate how meta-analytic researchers in the organizational sciences can use these to comply with MARS. We also provide recommendations that go beyond MARS to further advance the transparency, replicability, and accuracy of meta-analytic reviews. Before reviewing MARS, describing "best practices," and making recommendations, we briefly describe some terms.

Meta-analysis is a quantitative method used to combine the quantitative outcomes (effect sizes) of primary research studies. Meta-analysis is the statistical or data analytic part of a *systematic review* of a research topic. A systematic review follows a specific, replicable protocol to collect and evaluate scientific evidence with the primary objective of producing answers to research questions that cannot be addressed adequately by single studies (Cooper 1998; Cooper and Hedges 2009). However, meta-analysis, as a statistical technique, can be used to analyze data without the systematic review process (Cooper and Hedges 2009). For example, a medical school might use meta-analysis to examine the mean sex differences of the school's students on national standardized medical achievement test scores cumulated over a series of years. Thus, the meta-analysis per se does not require an understanding of the literature and can be performed without regard for the systematic review process. As with any statistical analysis, the quality of the result depends upon the quality of the data.

Because the quality of the systematic review depends upon the data and researcher skills, it is the responsibility of the meta-analyst to conduct a sensible review in terms of relevance and thoroughness, and to be transparent about the process of data extraction and analysis (Cooper 1998; Cooper and Hedges 2009; Ioannidis 2010). As a result, meta-analytic reviews should be guided by explicit decision rules regarding the literature search, the extraction of effect sizes, the data analysis, and other consequential details (Cooper and Hedges 2009; Egger et al. 2001b). Only then are the reviews systematic, transparent, and replicable.

MARS and the Meta-analytic Approaches

Our review and integration of the two primary quantitative review approaches used in the organizational sciences and other scientific areas follows the structure used in MARS. Due to space considerations, we do not discuss every topic in the systematic review process, but focus on issues that are of utmost importance to MARS and the quality of the meta-analytic review (e.g., APA 2008, 2010; Cooper 1998; Cooper and Hedges 2009; Egger et al. 2001b). In particular, we address issues that need clarification, or have been previously identified as being problematic, particularly in the organizational sciences (e.g., Aytug et al. 2011; Geyskens et al. 2009). We also provide model examples that illustrate the use of proper procedures and techniques so that meta-analysts, reviewers, and editors have models to help comply with MARS and improve the transparency, replicability, and accuracy of published meta-analytic reviews. Editors and reviewers are the gatekeepers of our sciences (Crane 1967) and should ensure that the published meta-analytic reviews follow our standards and "best practices." Our recommendations are summarized in Table 1.[1]

Title, Abstract, and Introduction

Systematic or meta-analytic reviews should be easy to identify. Thus, MARS recommends that the word 'meta-analysis' should appear in the title of any quantitative review. However, MARS makes little distinction between *systematic review* and *meta-analysis*. A meta-analysis is a statistical method that can be applied in contexts other than that of a literature review. We suggest that, at a minimum, the title conveys that the paper reports a review, either a systematic review or a meta-analytic review. If one chooses 'systematic' for the title, then the abstract should mention the meta-analysis.

Meta-analytic abstracts often omit quantitative information, particularly numerical results. Even studies for which the main aim was to find an overall effect size may fail to report it. A good abstract will at least include the number of effect sizes and the magnitude of effect for the main research question (see, e.g., Park and Shaw, in press; Van Iddekinge et al. 2012). Such an effect could be the mean effect size or the variance accounted for by a hypothesized moderator.

Any meta-analytic review should also include a clear statement of the goal for the review as well as the question(s) or relation(s) under investigation, including the historic background, theoretical review, etc. (APA 2008, 2010). This requirement mirrors the problem formulation stage of the systematic review (Cooper 1998) and includes the development of theoretical questions that need to be answered and the operationalization of the constructs of interest. In the organizational sciences, 96 % of all the meta-analytic reviews include this information (Aytug et al. 2011). In addition, descriptions and rationales for the selection and the coding of potential moderators, particularly methodological moderators

---

[1] We note that the table includes some exemplar models that do not fully comply with a specific recommendation. However, even the ones that do not fully comply with a specific recommendation provide more information than the typical meta-analytic review in the organizational sciences and, as such, can serve as an exemplar model.

Table 1 Recommendations for the procedural, methodological, and statistical considerations (in accordance with MARS)

Section and topic	Recommendations	Exemplar study or review[a]
(1) Title and abstract	Use of the term systematic review or meta-analytic review to clearly identify the paper	Halbert et al. (2006); Miller et al. (2008)
	Inclusion of quantitative information in the abstract	Park and Shaw (in press); Van Iddekinge et al. (2012)
(2) Introduction	Discussion of potential moderators hypothesized to be consequential	Else-Quest et al. (2010); Judge et al. (2001); Miller et al. (2008); Van Iddekinge et al. (2012)
(3) Method		
Design	Prospective meta-analyses are desirable	Cholesterol Treatment Trialists' Collaborators (2005)
Inclusion and exclusion criteria and moderator analyses	Detailed discussion and definition of inclusion and exclusion criteria, for example:	Banks et al. (2010); Carey et al. (2007); Else-Quest et al. (2010); Gurusamy et al. (2008); Hermelin et al. (2007); Judge et al. (2001); Park and Shaw (in press)
	The operational characteristics of variable(s), the eligible participant population(s), eligible research design features, the required time period in which the primary studies have been conducted, etc.	
	Detailed discussion and definition of coding categories to test for moderators	McDaniel et al. (1994); Miller et al. (2008); Kepes et al. (2012); Van Iddekinge et al. (2012)
	A priori considerations and coding of objective methodological factors such as study setting (e.g., lab study or field study), study design (e.g., concurrent or predictive design), sample type (e.g., randomized trial or convenience sample), measure (e.g., validated or non-validated measures), measure type (e.g., PCI Big Five or IPIP & NEO Big Five)	
	Related recommendation for journals:	
	Reporting of all relevant methodological and data information in primary studies (the leading journals could take a principal role establishing this practice)	
Search strategies	Detailed description of the literature search process	Carey et al. (2007); Gurusamy et al. (2008); Halbert et al. (2006); McDaniel et al. (1994); Terrizzi et al. (2013)
	Searches of electronic databases, manual searches (e.g., of journal articles, book chapters, and technical reports), personal communication (e.g., on list servers), etc.	
	Detailed description regarding the inclusion of grey literature (e.g., articles in languages other than English, samples, unpublished reports)	Halbert et al. (2006); Kuncel et al. (2001); McDaniel et al. (1994)
	Related recommendations for journals and organizations in the organizational sciences:	
	Establishment of supplemental article content (e.g., additional results) online (the leading journals could take a principal role establishing this practice)	
	Establishment of research registries (SIOP and AOM could take a pivotal role creating such registries)	
Coding procedures	Detailed discussion of the coding procedures to determine sample eligibility, including the number of coders, the inter-rater agreement, etc.	Carey et al. (2007); Gurusamy et al. (2008); Halbert et al. (2006); Kuncel et al. (2001); McDaniel et al. (1994); Park and Shaw (in press); Van Iddekinge et al. (2012)
	Description of winnowing at each step (e.g., of X studies retrieved, Y had insufficient quantitative information to compute an effect size)	

Table 1 continued

Section and topic	Recommendations	Exemplar study or review[a]
(4) Statistical methods	Disclosure of general information, such as formulae (or citations) used to convert effect size estimates and meta-analytic software used	Miller et al. (2008)
Estimation model	Random-effects models should be estimated	Costanza et al. (2012); Puts et al. (2008)
Weighting method and estimating the mean (and the confidence interval)		
Weighting method	More research, especially Monte Carlo simulations, to provide clear practical recommendations. Until then:	
	The weighting method should be appropriate to the research goals, which need to be clearly stated in the meta-analytic review	
Estimating the mean (random-effects)	The research question and goals should be clearly stated (e.g., estimating observed, operational, or latent effect sizes). Corrections for measurement error and range restriction should be consistent with the research question	Judge et al. (2001); Van Iddekinge et al. (2012)
	The confidence interval should be reported	Judge et al. (2001); Van Iddekinge et al. (2012)
Estimating the REVC (and the credibility interval)		
Estimating the REVC	The REVC should be estimated and reported	Kamdi et al. (2011); Kisamore and Brannick (2008)
	As with the mean, the researcher's goals should be made explicit, and any appropriate corrections incorporated into the calculations	
	Some statistic or index of uncertainty of the value of the REVC should be provided	
Estimating the credibility interval	The credibility or prediction intervals need to be reported (the interval should correspond to stated goals of the analysis)	Judge et al. (2001); Kamdi et al. (2011); Kisamore and Brannick (2008); Viechtbauer (2007)
	Desirable to report the prediction interval	
Estimating the heterogeneity of effect sizes	A confidence interval around the REVC should be estimated and reported	
	In addition to the prediction interval (noted above), it is desirable to report either the percent of variance attributable to moderators or I^2	Kamdi et al. (2011)
Explaining the heterogeneity of effect sizes	If possible, meta-regression or the meta-analytic analog to analysis of variance should be used to test for hypothesized moderators, and the magnitude of variance explained should be reported	Baltes et al. (1999); Conn et al. (2011); Kamdi et al. (2011); Kepes et al. (2012); Park and Shaw (in press)
	Subgroup analyses should be used:	
	After meta-regression to inspect the means of interesting subgroups	
	If missing data problems prevent the use of meta-regression	
	Moderators that were not hypothesized prior to the analysis should be reported under sensitivity analysis rather than as substantive variables of interest	
Sensitivity analyses	Use of detailed sensitivity analyses to investigate the impact of coded variables not hypothesized to be influential (e.g., language of journal or validation design) as well as characteristics of the data (e.g., outliers and missing data)	Banks et al. (2010); Conn et al. (2011); Halbert et al. (2006); Miller et al. (2008)

Table 1 continued

Section and topic	Recommendations	Exemplar study or review[a]
	Publication bias analysis:	Banks et al. (2012a); Kepes et al. (2012); Kepes et al. (in press)
	The possibility of publication bias should be addressed. The failsafe *N* method should be avoided, but other methods such as funnel plots, Egger's regression, and trim-and-fill should be considered as appropriate to the data	
	Research (e.g., Monte Carlo simulations) is needed to assess the appropriateness and robustness of different publication bias analyses techniques	
	Extrapolation:	Banks et al. (2010); McDaniel (2005); Miller et al. (2008)
	Sensitivity analyses should be conducted to illustrate the effect of corrections (e.g., for measurement error and range restriction) and to determine whether the conclusions of the meta-analysis would change depending on whether or not the corrections are applied	
(5) Results and discussion	Reporting of all relevant information from primary studies included in the meta-analytic review (e.g., sample size, effect size information, including mean and standard deviations, reliability estimates)	Judge et al. (2001); Kamdi et al. (2011); Miller et al. (2008)
	Reporting of the results from all analyses, including sensitivity (e.g., outlier and publication bias analyses) and moderator and mediator analyses. For example:	Dulebohn et al. (2012); McDaniel et al. (2007); Kamdi et al. (2011); Kepes et al. (2012); Kepes et al. (in press); Puts et al. (2008)
	Reporting of an effect size range:	
	Reporting of results from various supplementary analyses, including results from sensitivity analyses, publication bias analyses, moderator analyses (e.g., for methodological factors)	
	Use of tables, graphical displays (e.g., funnel plots, box-plots, forest plots, etc.) to communicate results	

[a] Exemplars are intended as models for the particular recommendation; they need not be outstanding or even adequate in other areas. Furthermore, the table includes some exemplar studies or reviews that do not fully comply with our recommendation. For instance, for sensitivity analyses, we could not locate a meta-analytic review that assessed the influence of language (e.g., English vs. non-English journal) on the meta-analytic results. It is possible that such meta-analyses exist in some area of science, but we were unable to locate such meta-analysis. We suggest that our lack of success in locating meta-analyses for some recommendations is indicative that the recommended practice is rarely or never incorporated into existing meta-analyses

(see, e.g., Else-Quest et al. 2010; Judge et al. 2001; Miller et al. 2008), should be provided to motivate the study (APA 2008, 2010). Unfortunately, such information is not regularly reported in the organizational sciences (Aytug et al. 2011). We thus recommend more detailed discussions of the moderators hypothesized to be consequential.

Method

Design

An issue that is often overlooked in the organizational sciences is the design of the meta-analytic study. In virtually all sciences, *retrospective* studies, which evaluate cumulative data and provide guidance for future research (Berlin and Ghersi 2005), are used to answer the research question(s). In addition to retrospective studies, some scientific areas conduct *prospective* meta-analyses where researchers collaborate a priori to develop a series of primary research studies, and it is decided prior to the completion of the primary studies that the findings will be meta-analyzed (Berlin and Ghersi 2005). This a priori collaboration allows for the standardization of research methods, the ability to pre-specify subgroups, and the avoidance of unnecessary duplication of research efforts. For instance, prospective studies are conducted by medical researchers to evaluate the effectiveness of a particular

drug (e.g., Cholesterol Treatment Trialists' Collaborators 2005). Prospective meta-analytic studies have several advantages, including reducing bias in problem formulation, allowing for the standardization of measures, and minimizing the possibility of publication bias (e.g., Berlin and Ghersi 2005; Higgins and Green 2009). In order to generate cumulative knowledge more effectively, we recommend the use of prospective meta-analyses in the organizational sciences.

Inclusion Criteria and Moderators

To ensure that the internal and external validity of the meta-analytic review can be assessed, it is imperative to define and explain the inclusion and exclusion criteria for the selection of primary samples (Cooper 1998). The specification of inclusion and exclusion criteria is also important to judge the possibility of biases in the meta-analysis (Kepes et al. 2012; Rothstein 2012). MARS explicitly mentions these issues, including descriptions of the operational characteristics of variables, the participant population(s), research design features, the required time period in which the primary studies should have been conducted, and any other variables of consequence for inclusion in the study. Also, the coding categories to test for potential moderating effects need to be defined (APA 2008, 2010). These definitions and explanations should be aligned with the problem formulation in the introduction (Cooper 1998). Thus, we recommend the detailed discussion of exclusion and inclusion criteria as well as descriptions of operational characteristics (see, e.g., Banks et al. 2010; Carey et al. 2007; Else-Quest et al. 2010; Hermelin et al. 2007; Judge et al. 2001; Van Iddekinge et al. 2012).

Search Strategies

Once the operational characteristics of the variables of interest are defined, the literature search can begin (Cooper 1998). Search strategies, including the review of the literature and the data collection, are among the most important aspects of a meta-analytic review. Without a detailed and rigorous search, the meta-analytic effect size data can be biased, leading to potentially erroneous results (Kepes et al. 2012; Rothstein 2012; Rothstein et al. 2005b). Furthermore, without the detailed reporting of the systematic search process, the transparency and replicability of the meta-analytic review is compromised (Cooper and Hedges 2009; Egger et al. 2001b). We thus recommend a very detailed systematic search process and the thorough description of it (see Rothstein 2012, for additional details).

MARS includes several topics of relevance for the literature search, including the identification of searched databases, registries, manual search efforts, and the process

of determining sample eligibility.[2] In the organizational sciences, general electronic databases (e.g., PsycInfo, Web of Science, etc.) are typically searched, and journals, conference proceedings, and reference sections of articles may be examined. In addition, personal correspondence with other researchers is sometimes used to gather unpublished and otherwise unavailable samples.

MARS also explicitly mentions the access of articles and reports in languages other than English and a description of how unpublished samples were treated. Unfortunately, these issues are rarely addressed in meta-analytic reviews in the organizational sciences. As a result, data from sources such as government or company reports, unpublished studies, and data from articles in languages other than English are infrequently included in meta-analytic reviews in the organizational sciences (Aytug et al. 2011), which can lead to biased meta-analytic results (Kepes et al. 2012; Rothstein 2012). In contrast to the organizational sciences, other scientific areas have research registries (Berlin and Ghersi 2005; White 2009), such as the *Campbell Collaboration* in social work and social psychology, the *What Works Clearinghouse* in education, and the *Cochrane Collaboration* and *ClinicalTrials.gov* in the medical sciences.[3] Such registries allow systematic searches to identify relevant unpublished and prospective research studies (Kepes et al. 2012). Such registries are explicitly mentioned in MARS but do not exist in the organizational sciences. Similarly, journals in the medical and related sciences often include supplementary analyses and results on their websites (Evangelou et al. 2005). This supplementary information can then be searched and relevant results can be included in the meta-analysis. Unfortunately, such practices are rare in the organizational sciences (Kepes et al. 2012).

Even though there is detailed guidance on the reporting practices for the organizational sciences (e.g., APA 2008, 2010; Cooper 1998; Hunter and Schmidt 2004), meta-analytic reviews tend to provide substantially less information than is recommended. For example, a 2007 *Journal of Applied Psychology* article summarized its literature review in one sentence: "We conducted a search of the OCB literature by using a number of online databases (e.g., Web of Science, PsycINFO) as well as by examining the reference lists of previous reviews" (Hoffman et al. 2007, p. 577). Unfortunately, limited descriptions of the systematic search process are relatively common in organizational science journals. By contrast, the standardization

---

[2] In meta-analytic studies, the terms studies and samples are often used interchangeably. We use the term samples throughout this manuscript because a single study can contain multiple samples.

[3] In the medical sciences, many top-tier journals preclude the publication of clinical trials unless they were registered prior to completion (Berlin and Ghersi 2005; Laine et al. 2007).

and reporting standards in other scientific areas tend to be more stringent (Berman and Parker 2002; Higgins and Green 2009); the reporting of the systematic search process is typically very detailed, enhancing the transparency and replicability of the meta-analytic review (Cooper and Hedges 2009; Egger et al. 2001b).

In the interest of better generation and development of cumulative knowledge and improved compliance with MARS, we recommend the implementation of some practices used in other sciences. For instance, there is still considerable room for further improvement in describing the literature search process (Aytug et al. 2011). As an example, the time period covered by the search is seldom described in the organizational sciences, and the keywords used in the search are also inconsistently disclosed (Aytug et al. 2011). However, MARS recommends the reporting of both of these items. Also, more explicit descriptions of how studies in non-English languages and non-published samples have been treated should become common practice (see, e.g., Halbert et al. 2006; Kuncel et al. 2001; Terrizzi et al. 2013). Finally, more rigorous reporting standards regarding the literature search should be implemented in our journals (see Rothstein 2012). Journals should also start to provide supplemental information online, such as results of subgroup analyses and related descriptive statistics (Evangelou et al. 2005). Finally, the implementation of research registries, which permit a more thorough search and the identification and potential inclusion of unpublished research studies (Berlin and Ghersi 2005; White 2009), would clearly be advantageous. Organizations within the fields of management and I/O psychology (e.g., the *Society for Industrial and Organizational Psychology* [SIOP] and the *Academy of Management* [AOM]) could take a pivotal role in creating such registries (Banks and McDaniel 2011; Kepes et al. 2012).

Coding Procedures

Coding procedures are part of the data evaluation stage of a meta-analytic review (Cooper 1998). Several considerations are used in deciding whether to include a sample in the meta-analysis, including the methodological adequacy of a sample (e.g., evaluating whether a study's sample was collected using an appropriate research design), the sample's relevance to the review, and whether an effect size can be computed from information presented in the study. The use of transparent coding procedures and experienced coders is vital as is a description of the filtering process (e.g., how many samples were excluded for lack of effect size information; see, e.g., Carey et al. 2007; Gurusamy et al. 2008; Halbert et al. 2006; McDaniel et al. 1994; Van Iddekinge et al. 2012). A flow chart (see Fig. SM1 in the supplemental materials) can be used to display the process

of winnowing each step of the search and coding process (see also, e.g., Carey et al. 2007; Gurusamy et al. 2008; Halbert et al. 2006). Unfortunately, only slightly more than half of the meta-analytic reviews in the organizational sciences indicate the number of coders used and how potential disagreements were resolved (Aytug et al. 2011), which are integral parts of the process for determining sample eligibility (APA 2008, 2010).

In the organizational sciences, there is rarely screening of studies based on quality. Conversely, other scientific areas may use a priori quality assessments, presumably to eliminate studies of poor methodological quality from a review in order to conduct a "best-evidence synthesis" (Slavin 1986). Often, this evaluation is used to select only samples from randomized controlled trials (Egger et al. 2001a). Although rarely used in the social sciences, such assessments have been used (e.g., Richardson and Rothstein 2008).

Outside of the organizational sciences, as part of a quality assessment, some research assigns each sample a "quality score," which can be used as a selection criterion for inclusion or exclusion in a systematic review, as a weighting value for each sample, or as a tool to categorize samples into subgroups for comparisons (Berman and Parker 2002). Yet, quality scores can bias findings if researchers make decisions regarding sample inclusion that are not empirically tested (Hunter and Schmidt 2004). Furthermore, there is evidence to demonstrate that inter-rater agreement among even experienced evaluators for research quality is relatively low (inter-judge correlations of around .5; Cooper 1998). Thus, the use of quality scores is problematic.

The coding of study design features is also relevant to quality assessment. Unfortunately, meta-analytic reviews in the organizational sciences rarely report whether design features were coded and what they were (Aytug et al. 2011). Table SM1 in the supplemental materials provides an exemplar template for such coding procedures. For instance, in only around 25 % of all meta-analytic reviews is it clear whether different study designs are separated (Aytug et al. 2011). This is undesirable because study design features and types of measures can have a substantial effect on meta-analytic summary statistics and on the interpretation of those summaries (e.g., Kepes et al. 2012; Lipsey and Wilson 2001). In accordance with MARS, we recommend the coding of objective study design and methodological factors for the formation of subgroups and sensitivity analyses based on these subgroups (Lipsey and Wilson 2001).

Statistical Methods

Thus far, we discussed procedural aspects of the systematic review process. Next is the data analysis stage (i.e., the

meta-analysis itself; Cooper 1998), which is addressed under the topic statistical methods in MARS. Before we address specific issues, we briefly describe some general matters regarding meta-analytic statistical approaches.

Although the organizational sciences predominantly use psychometric meta-analysis (i.e., the H&S approach; e.g., Hunter and Schmidt 2004; see Aytug et al. 2011), the statistical meta-analytic approach used in other areas in the social (e.g., education and some disciplines in psychology) and medical sciences is usually based on what we call the Hedges and Olkin (H&O) statistical approach to meta-analysis (Borenstein et al. 2009; Hedges and Olkin 1985; Hedges and Vevea 1998). Although other statistical meta-analytic approaches can also be used in the data analysis stage, they are not described here because the H&S and H&O approaches are the two most widely used approaches. Also, related approaches in the psychometric realm (e.g., Raju et al. 1991) are similar to the H&S approach (Hunter and Schmidt 2004). Finally, we do not discuss Bayesian meta-analytic models (e.g., Brannick 2001; Steel and Kammeyer-Mueller 2008) or the varying coefficient model of Bonett (2008) due to space considerations.

MARS requires that meta-analytic reviews address the effect size metric used, including formulae for effect size transformations and possible corrections. We focus primarily on correlation coefficients in this paper.[4] The reason for this is twofold. First, correlations are the most common and analyzed effect size in the organizational sciences (Aytug et al. 2011; Geyskens et al. 2009). Second, both statistical approaches allow the user to analyze correlation coefficients. However, we note that the H&O approach permits analyses of many more effect size indices than the psychometric approach, ranging from correlations and correlation ratios (e.g., r, η^2, ω^2), unstandardized and standardized mean differences (e.g., D, d, g), to effect sizes for binary data such as risk and odds ratios (Borenstein et al. 2005; Borenstein et al. 2009). However, as the vast majority of the effect size measures in the organizational sciences tend to be correlation coefficients or standardized mean differences, the ability to handle additional effect sizes may be of limited interest to many organizational researchers. Furthermore, formulae to convert some effect sizes are readily available (e.g., Borenstein et al., in press; Lipsey and Wilson 2001). According to MARS, the formulae to calculate or transform effect sizes as well as the meta-analytic software used should be disclosed. Unfortunately, such disclosure appears in less than half of the

meta-analytic reviews in the organizational sciences (Aytug et al. 2011).

Regarding the available meta-analytic software packages, organizational researchers may be familiar with software designed to follow the H&S approach to meta-analysis (Roth 2008). However, there are now several stand-alone packages capable of computing the H&O analyses and creating various graphical displays for the visual communication of results. A recent review suggested that two programs, Comprehensive Meta-Analysis (Borenstein et al. 2005) and MIX (Bax et al. 2006), are easy to use and offer many different capabilities for the analysis as well as the graphical presentation of results (Bax et al. 2007).

Estimation Model

There are two generic models, random- and fixed-effects models[5] (Borenstein et al. 2009; Hedges and Vevea 1998; Hunter and Schmidt 2004; Schmidt et al. 2009). Because both estimation models have different assumptions and can yield differing results, the model used in a meta-analytic review should be identified (APA 2008, 2010). Meta-analytic estimation models in the H&O approach can be fixed- or random-effects (Borenstein et al. 2009; Hedges and Olkin 1985; Hedges and Vevea 1998). Meta-analyses following the H&S approach are random-effects models (Hunter and Schmidt 2004) because they include a term for residual or unexplained variability in the infinite-sample effect sizes after accounting for the other terms in the model (sampling error and statistical artifacts). Due to the assumptions underlying both estimation models, the random-effects model is the more general (i.e., it incorporates the fixed-effects model) and appears more consistent with the research questions posed in systematic reviews in the organizational, social, and medical sciences (Borenstein et al. 2009; Hunter and Schmidt 2004; Sutton 2005). Because the amount of error in the parameters computed by fixed-effects meta-analytic models is underestimated when the true variance between studies is greater than zero, statistical problems (e.g., inflated Type I error rates) and inaccurate meta-analytic results (e.g., overly optimistic confidence intervals) may occur (Field 2005; Hedges and Vevea 1998; Hunter and Schmidt 2000; Schulze 2004). Thus, the use of the random-effects model is recommended for most applications, and our discussion henceforth is confined largely to this model.

Both the H&S and H&O statistical approaches to meta-analyses have similar aims. They are typically used to

---

[4] Most statistical considerations we discuss in the context of correlation coefficients also apply to other effect size statistics such as standardized mean differences. Formulae for other effect sizes are available in the respective literatures (e.g., Borenstein et al. 2009; Hedges and Olkin 1985; Hedges and Vevea 1998; Hunter and Schmidt 2004).

[5] We note that there are also models described as 'mixed-effects' (e.g., Raudenbush and Bryk 1985). Because these models contain terms for the residual variance in underlying effect sizes, we classify them as random-effects models for the purposes of this paper.

estimate an overall, or mean effect size. They both estimate the amount of sampling error variance and the residual, or random-effects variance in correlations that remain when moderators and sampling error are removed. The two approaches also address moderator analyses. Although both approaches consider random sampling error, the H&S statistical approach also addresses other artifactual sources of error such as differences across samples in reliability of measurement and range restriction. As with sampling error, these additional artifactual errors increase the observed variability in effect sizes. Some artifacts, such as measurement error and range restriction, also have a biasing or moderating effect on the effect sizes (Hunter and Schmidt 2004).

The authors from the two meta-analytic traditions (e.g., the H&S approach and the H&O approach) tend to use different notations (e.g., Borenstein et al. 2009; Hedges and Olkin 1985; Hedges and Vevea 1998; Hunter and Schmidt 2004) for the random-effects overall mean and the random-effects variance component (REVC; the variance of rho $[\sigma_\rho^2]$ in the H&S approach; tau-squared $[\tau^2]$ in the H&O approach). For purposes of consistency, we use the term REVC for our descriptions of both statistical meta-analytic approaches.

Estimating the Mean and Confidence Interval

To ensure the transparency and replicability of a meta-analytic review, the meta-analytic review needs to provide information about the process in which the mean and the confidence interval were estimated. Most meta-analytic approaches calculate a weighted mean because some samples contain more or better information than do others. The choice of weights is different in the two approaches (Borenstein et al. 2009; Hedges and Vevea 1998; Hunter and Schmidt 2004). MARS recommends describing the weighting procedure used to compute the weighted mean (i.e., overall effect size estimate) and the calculation of the confidence interval for it, both of which we address next.

The H&S Approach

In the simplest, or 'bare bones' version of psychometric meta-analysis, the effect sizes are weighted by the sample size (N_i) (Hunter and Schmidt 2004). Unlike the 'bare bones' version, the full version of psychometric meta-analysis includes correction for artifacts, which may include reliability (in either X, Y, or both), range restriction, and potentially additional attenuation factors, such as those resulting from dichotomizing a continuous variable (Hunter and Schmidt 2004). Under the assumption that the artifacts are independent, the compound attenuation factor (A, the ratio of unadjusted to adjusted effect, less than or equal to 1.0) is computed for each sample (Hunter and Schmidt 2004, p. 121; see also Hunter et al. 2006). The effect of the compound attenuation factor is to reduce the weight given to samples that have a large amount of error due to artifacts (the weights for the full correction model are $w = NA^2$).

The *confidence interval* for the mean follows the conventional method for setting a confidence interval for raw data except that the variance is weighted and the elements may already be corrected. Hunter and Schmidt (2004, p. 206) provide the formulae, including the necessary adjustments when using correlations that have been corrected for statistical artifacts.

The H&O Approach

In the H&O statistical approach, for the fixed-effects model, effect sizes are weighted by the inverse of their sampling error variance (the square of the effect size's standard error) (Borenstein et al. 2009; Hedges and Olkin 1985). Before correlations are analyzed, they are converted to Fisher's z. For an individual correlation coefficient converted to Fisher's z, the sampling variance is approximately $1/(N_i - 3)$. Thus, in the fixed-effects case, the weights for the H&O method are $N_i - 3$, which is nearly identical to N_i, the H&S approach 'bare bones' meta-analysis weight.

In the H&O random-effects model, which is of primary interest for the organizational sciences, the weights are slightly more complicated because they incorporate uncertainty both from the individual sample's sampling error and from between-sample variance in the underlying parameter, the REVC (Borenstein et al. 2009; Hedges and Olkin 1985). Given the z-transformation in the H&O approach, the calculation of the 95 % *confidence interval* around the estimate is a bit more complicated but formulae are readily available (Borenstein et al. 2009; Field 2005; Hedges and Olkin 1985; Hedges and Vevea 1998). The overall mean and its upper and lower boundary estimates are still expressed in *z*, however. To return them to their original metric and to make them comparable to the H&S method, they must be converted back from *z* to *r*.

Comparison of the Two Approaches and Recommendations

The rationale for the weighting schemes in both meta-analytic approaches is conceptually similar with respect to sampling error variance and differs with respect to non-sampling error variance. Both models consider correlations from samples with less sampling error to be better estimators of the population effect. In other words, both statistical approaches attempt to give greater weight to more precise estimates. In the H&S approach, this is

implemented with sample-size weighting (Hunter and Schmidt 2004), and in the H&O approach with inverse sampling error variance weighting (Borenstein et al. 2009; Hedges and Olkin 1985). With respect to non-sampling error, the key difference in weighting between the statistical approaches is that the H&O approach incorporates the non-sampling between-sample variance (REVC, τ^2) in the weights for random-effects model, but the H&S statistical meta-analytic approach does not. One effect of incorporating the random-effects variance in the weights is to move them closer to unit weights (i.e., the effect size weights become more similar to each other) as the random-effects variance becomes large. When the random-effects variance is small, there will usually be very little difference between the H&S (bare bones) and the H&O weights and overall means. Further, as the number of effect sizes (i.e., samples) increases, the effect of the different weighting schemes diminishes (e.g., Einhorn and Hogarth 1975).

Using the inverse of the effect's sampling variance to weight effect sizes has statistical advantages (e.g., Böhning 2000; Hedges and Olkin 1985). However, these advantages may not always manifest themselves in practical applications (Brannick et al. 2011; Field 2005; Schulze 2004). Thus, the debate over which weighting procedures are optimal for practical applications of meta-analytic models is likely to continue,[6] and we recommend additional research on this matter.

Although the H&S 'bare bones' meta-analytic mean and its confidence interval will usually be quite similar to that computed by the H&O fixed-effects model, the H&S model with corrections for statistical artifacts (e.g., measurement error and/or range restriction) may lead to a meaningfully larger difference in the magnitude of the mean effect size estimate and the associated confidence interval (Hunter and Schmidt 2004; Hunter et al. 2006). When corrections for statistical artifacts are made, the mean estimate will tend to become larger in absolute value, and the width of the confidence interval around the mean will also expand in proportion to the increase in absolute value.

Corrections for measurement error and range restriction have been used for decades prior to the beginning of the H&S approach and were commonly presented in psychometric texts. However, there is an ongoing debate regarding the appropriateness of these corrections (e.g., Geyskens et al. 2009; Hunter and Schmidt 2004; Rosenthal 1991). In our view, the appropriate course of action depends upon the researcher's objective. For example, if the objective of the meta-analytic review is to summarize the existing data to determine "what is" rather than "what might be," no corrections should be applied (Rosenthal 1991). By contrast, when estimating the validity coefficient of selection tests, it is customary to correct correlations for unreliability in the criterion but not the predictor when computing the overall mean. The rationale is that in practice, the fallible test scores will be used as predictors, but the actual benefit, not our imperfect measure of it, will accrue to the company using the test (Hunter and Schmidt 2004).

If corrections for range restrictions are made, one also needs to consider whether the restriction is direct or indirect (Hunter et al. 2006). Typically, such corrections are only performed when using the H&S approach. However, corrections are possible in the H&O approach (e.g., Aguinis and Pierce 1998; Borenstein et al. 2009; Hall and Brannick 2002; Hedges and Olkin 1985; Lipsey and Wilson 2001). Currently, only around half of the meta-analytic reviews in the organizational sciences provide information regarding such corrections (Geyskens et al. 2009). In considering range restriction, one must also consider the population to which the results are to be generalized. For example, suppose one were to meta-analyze the relations between cognitive ability and success in engineering jobs. The population variance for cognitive ability will be much larger in the general population than in a population of engineering graduates applying for jobs in engineering firms. Applying the correction for indirect range restriction for the general population would not be appropriate if one were interested in the selection of engineers from the population of engineering graduates. In accordance with MARS, we encourage authors of meta-analytic reviews to communicate clearly their objectives and, accordingly, make corrections (or refrain from making corrections) for measurement error and/or range restriction (i.e., direct or indirect) (see, e.g., Judge et al. 2001; Van Iddekinge et al. 2012).

Estimating the REVC and Credibility Interval

As discussed previously, one primary purpose of meta-analytic reviews is to estimate the REVC and the extent to which the mean estimate is affected by sources of heterogeneity, such as moderating effects. Such a between-sample variance, the REVC, is estimated by finding a residual variance that essentially contains variance left over after estimating variance due to sampling error and possibly other artifacts. The two approaches use different computations to estimate the REVC. The estimation of the REVC is critical because it affects the calculation of the credibility interval (called prediction interval in the H&O approach), which estimates the dispersion around a mean estimate (Borenstein et al. 2009; Hunter and Schmidt 2004; Whitener 1990).

---

[6] We note that there is little difference in estimates of the mean between both weighting schemes for the majority of effect size statistics in the organizational sciences (e.g., correlations and standardized mean differences). Conceptually, given the determinants of the variance in binary data, for effect size indices such as the odds ratio, the differences can be noticeable, favoring the inverse variance weights (Borenstein et al. 2009; Indrayan 2008; Sutton et al. 2000).

A credibility interval is a posterior probability interval (Edwards et al. 1963) that reflects the underlying, population effect sizes. Compared to a confidence interval, which addresses the precision of the estimated mean, the credibility interval contains a percentage of the distribution of a random variable. If the underlying or 'true' effect size varies from sample to sample, we would like to know the likely range of those 'true' effect sizes. Thus, the credibility interval provides an estimated range and can be used to judge the degree to which the mean estimate is affected by moderating effects and other potentially unobserved influences (see Whitener 1990, for an expanded discussion of the difference between confidence and credibility intervals). Due to its importance, MARS specifically mentions the credibility interval on their list of information that should be included in any meta-analysis. Unfortunately, only around 22 % of all meta-analytic reviews report such intervals (Geyskens et al. 2009).

The H&S Approach

The H&S approach considers three sources of observed variance: variance due to moderators, sampling error, and other statistical artifacts, such as variance due to measurement error and range restriction (Hunter and Schmidt 2000, 2004). The variance of sampling error (in the 'bare bones' case) or the variance of sampling error plus the variance of artifacts (in the full corrections case) is subtracted from the variance of the effect sizes (uncorrected or corrected, respectively). Thus, the REVC is computed by finding a residual. If necessary, corrections need to be made individually for each sample before estimating the REVC (Hunter and Schmidt 2004, pp. 121–126).

Hunter and Schmidt's (2004) psychometric meta-analysis approach made the *credibility interval* popular for estimating the dispersion around an estimated mean population correlation (an estimate after sampling error variance and variance due to statistical artifacts have been removed; Hunter and Schmidt 2004). No assumption about the underlying distribution of ρ is necessary to compute the REVC using the H&S approach. However, it is conventional to assume that the underlying distribution is normal, which allows setting conventional boundaries using tabled values.

The H&O Approach

The meta-analytic approach advocated by H&O allows for several different estimators of the REVC. The one most commonly used was developed by DerSimonian and Laird (1986; see also, e.g., Borenstein et al. 2009; Hedges and Vevea 1998). If the effect sizes vary only by sampling error, the weighted sum of squared deviations from the

mean (Q) will be distributed as χ^2 with $k - 1$ degrees of freedom (where k is the number of samples). This fact forms the statistical basis for using Q as a significance test for the homogeneity of effect sizes and for estimating the REVC.

Borenstein et al. (2009) described the quantity analogous to the H&S credibility interval as the *prediction interval*. In contrast to the credibility interval, the prediction interval incorporates uncertainty both in the value of the mean and the value of the REVC. With primary studies, the t distribution is used in place of z when the population variance is unknown (as is the case of the typical application of the t test). The t distribution is used in the H&O approach to acknowledge that the REVC is estimated, and that the estimate becomes more precise as the number of effect sizes (and thus degrees of freedom) becomes larger.

Comparison of the Two Approaches and Recommendations

Monte Carlo simulations suggested that there is little difference in the credibility intervals computed by both methods provided that sampling error is the only source of error (Hall and Brannick 2002). Thus, one would expect there to be little difference in the size of the REVC between the H&O 'bare bones' method and the typical application of the H&O approach. However, the H&S approach considers differences in effect sizes attributable to statistical artifacts other than sampling error, such as differences in measurement error and range restriction. To the degree that such artifacts are producing variance in the observed correlations, the REVC will be overestimated by the H&O approach as it is typically employed. Therefore, because applications of the H&O statistical approach typically do not correct for these artifacts,[7] we expect that when there are large differences between samples in measurement error and range restriction, the H&S approach should produce more accurate estimates of the REVC when such differences are accounted for properly. Additional simulation work suggests that the H&S method has particular advantages for the meta-analysis of correlations, but not necessarily for other effect sizes, such as d (Brannick et al. 2011; Marín-Martínez and Sánchez-Meca 2010).

On the other hand, the prediction interval appears to do a better job accounting for statistical uncertainties regarding the mean and variance in the meta-analysis. In the typical application of meta-analysis, the mean effect size may have a small error associated with it, particularly if the REVC is

---

[7] However, as mentioned previously, corrections for statistical artifacts are possible in the H&O meta-analytic approach (e.g., Aguinis and Pierce 1998; Borenstein et al. 2009; Hall and Brannick 2002; Hedges and Olkin 1985; Lipsey and Wilson 2001).

small. This is because the error variance associated with the estimate of the overall mean will be largely a function of the total sample size, and the uncertainty about the mean will be negligible. However, the uncertainty about the REVC is likely to be considerable. Although Hunter and Schmidt (2004, p. 207) noted the uncertainty regarding this estimate, they did not provide any means of dealing with it. Borenstein et al. (2009, pp. 122–124) provided confidence intervals for the REVC for the H&O approach (see Viechtbauer 2007).

In summary, the H&O prediction interval would appear to overestimate the width of the interval when there are artifactual sources of variance beyond sampling error, and the H&S credibility interval would appear to underestimate the width of the interval when the individual sample sizes are small, and, especially, when the number of samples in the meta-analysis is small. However, the computation of the credibility and/or prediction interval is not dependent upon whether researchers use the H&S or H&O approach to estimate the REVC. One could use the prediction interval reported in Borenstein et al. (2009) with the H&S approach. Such a procedure would allow for uncertainty in the estimates of the mean and REVC. Alternatively, one could compute the REVC using the H&O approach with statistical artifacts (unreliability and range restriction) treated as covariates (continuous moderators; see Borenstein et al. 2009, pp. 348–349) in a meta-regression (addressed in the next section). Such a procedure would allow the computation of the REVC controlling for statistical artifacts, and is conceptually similar to the estimation of the REVC by the H&S approach when correcting for such artifacts.

Finally, both statistical approaches to meta-analysis assume that the underlying distribution of effect sizes is normal. Hall and Brannick (2002; see also Field 2001) showed that as the mean and variance of the underlying normal distribution of correlations increased above zero, the H&O prediction values became increasingly inaccurate. Such a finding is due to the r to z transformation, which stretches the upper tail of the distribution, resulting in an overestimate of the mean and variance of r when converted back into its original metric (see Schulze 2004). Other authors have objected to the assumption of the normal distribution of 'true' correlations (Bobko and Roth 2008; Kemery et al. 1989; Thomas 1988). However, Kisamore (2008) showed that the shape of the underlying distribution has a rather modest effect on the accuracy of the estimate of the lower bound of the credibility interval, and that the size of the REVC and the desired level of confidence (e.g., 95 % vs. 80 %) have a greater effect on accuracy. In addition, Hafdahl and Williams (2009) showed that the bias in the r to z transformation could be largely controlled by using an additional transformation based on the assumption

of a normal distribution. However, their study is based on knowing the distribution of 'true' correlations, and the effect of violations of the assumption is unknown.

It is not a coincidence that MARS explicitly mentions the credibility interval; it is vital to assess the variability around the meta-analytic mean estimate. Thus, the fact that less than one fourth of all meta-analytic reviews in the organizational sciences report this information (Geyskens et al. 2009) is troublesome. We recommend that the REVC be estimated and reported. We also recommend the routine calculation and reporting of prediction intervals to indicate the likely range of effect sizes because this interval allows for uncertainty in both the mean and the REVC. Assuming that the REVC is estimated accurately, a small range suggests that further primary research is unlikely to show much difference from the current meta-analysis. By contrast, a large prediction interval suggests the presence of either a large number of nuisance variables (e.g., small magnitude moderators) or a smaller number of large magnitude moderators, and thus the need for additional research in the area. Likely sources of artifactual variance should be acknowledged, and their potential impact on the estimate of the REVC should be considered.

Estimating the Heterogeneity of Effect Sizes

As mentioned previously, one primary objective of a meta-analytic review is to assess the dispersion around the mean estimate. Thus, one is interested in estimating the stability or precision of the meta-analytic mean estimate. To do this, one needs to estimate the heterogeneity of the effect sizes. Not surprisingly, MARS recommends a description of how the heterogeneity in effect sizes was assessed or estimated. Unfortunately, almost two-thirds of all meta-analyses do not describe how this issue was addressed and approximately half fail to identify whether heterogeneity exists (Aytug et al. 2011). We already addressed confidence as well as credibility and prediction intervals, which provide information regarding the stability of the meta-analytic mean estimate.[8] However, there are additional statistics that can be used.

The H&S Approach

Schmidt and Hunter (1977) proposed the 75 % rule to assess the degree of homogeneity of effect sizes. In the H&S approach, the REVC is considered to be zero for practical purposes if 75 % or more of an effect size's

---

[8] We recommended the reporting confidence intervals for the meta-analytic mean effect size as well as the computation of the REVC and its confidence interval (Viechtbauer 2007). We prefer the reporting of the prediction interval over the credibility interval.

variance can be explained by sampling error and other statistical artifacts (Hunter and Schmidt 2004). If this threshold is met, Hunter and Schmidt (2004) suggested that moderators and unaccounted for statistical artifacts do not substantially influence the effect size. The logic behind this recommendation is that some artifacts, such as clerical errors (e.g., transcription errors), contribute to the estimate of the REVC, but cannot be taken into account by the calculations.

The 75 % rule has been criticized on several grounds (e.g., Sackett et al. 1986; Spector and Levine 1987). Of particular concern, when the sample sizes are small, 75 % of the observed variability may be explained by sampling error even when the REVC is large, resulting in a meaningfully large range of 'true' correlations (Borenstein et al. 2009). That is, expressing the REVC as a percentage rather than as a range of likely 'true' values fails to convey the most important information concerning the variability of effect sizes. The prediction interval is thus more informative for describing the heterogeneity of effect sizes.

The H&O Approach

In the H&O approach, there are two statistics, the homogeneity test Q and the I^2 index, that assess the homogeneity of the 'true' effect sizes. Q estimates the variation between effect sizes across samples. A statistical test to estimate whether the samples share a common effect size (i.e., the null hypothesis) can also be computed (Borenstein et al. 2009). The H&S approach notes the Q statistic as well, but its use is discouraged because the interpretation of Q relies on a statistical significance test, and because it almost always indicates that there is some variance that is not attributable to sampling error (Hunter and Schmidt 2004).

The I^2 index complements the Q statistic as the latter only indicates whether homogeneity is present or absent, but not its magnitude. The I^2 index quantifies the magnitude of the heterogeneity and is easily interpretable (i.e., as I^2 approaches 100 %, all of the observed variance can be attributable to 'true' variance), and, contrary to the Q statistic, I^2 does not depend on the degrees of freedom (Higgins and Thompson 2002; Higgins et al. 2003). Confidence intervals around the I^2 index can also be estimated (Borenstein et al. 2009; Higgins and Thompson 2002; Higgins et al. 2003). Finally, as with the percentage of variance accounted for (i.e., the 75 % rule), the index can be used to assess the fit of alternative models by comparing the magnitude of separate I^2 indices for moderator variables (Huedo-Medina et al. 2006).

Comparison of the Two Approaches and Recommendations

None of these rules of thumb, indices, or statistics are comparable across meta-analyses because they are dependent on the sample sizes of the studies included.[9] Although they are computed differently, conceptually, the I^2 index and the percentage of variance accounted for are complementary, such that $(1 - I^2)$ equals the percentage of variance due to artifacts (e.g., 75 % variance accounted for by artifacts corresponds conceptually to an I^2 of 25 %).

The value of Q can be tested for statistical significance by relating the value of Q to the χ^2 distribution (with $k - 1$ degrees of freedom; Borenstein et al. 2009). The percentage of variance accounted for and I^2 provide descriptive information related to the practical significance. Hunter and Schmidt (2004) advocated against the Q test because it is a statistical significance test, and because it will have low power with small k, but high power with large k. This high power would typically result in a statistically significant Q. However, the two types of statistics are complementary. The Q test provides information about the observed data, and the percentage of variance and I^2 provide information about the relative magnitude of observed and random-effects variances. Thus, each statistic or index provides a unique piece of information.

Although the significance test and the I^2 or percentage of variance accounted for may be interesting by themselves, they are not as interpretable and, therefore, typically not as meaningful as the prediction interval, which shows the likely range of 'true' effect sizes in the original metric. We thus recommend the routine estimation and reporting of the REVC and prediction interval along with either the percentage of variance accounted for or I^2 (see, e.g., Kamdi et al. 2011; Kisamore and Brannick 2008).

Explaining the Heterogeneity of Effect Sizes

In addition to the estimation of heterogeneity, meta-analytic reviews should explain potential reasons for the unexplained variance. Whether a priori or post hoc, such investigations are tests of, or searches for, moderators. MARS specifically recommends moderator analyses to explain the heterogeneity of effect sizes. We next describe several approaches to do this.

Although the statistical tests make no distinction between hypothesized and post hoc tests of moderators, there is an important difference in the meaning of the tests.

---

[9] For instance, conceptually and computationally, meta-analyses with smaller sample sizes per sample will have smaller I^2 indices, on average, due to greater sampling error variance than will meta-analyses with large sample sizes per sample due to smaller sampling error variance, even if their between-sample variability (e.g., moderator variance) is identical. Hunter and Schmidt's 75 % rule shares this problem because a given 'true' variance of the meta-analytically derived effect size will be a larger percentage of the total variance as the sample size of the primary samples increases.

The strongest use of meta-analysis is as a tool for the evaluation of scientific propositions using previous research outcomes as data. In such a case, meta-analysis is employed for data analysis just as regression or analysis of variance would be used in primary research studies in which data are gathered to test specific hypotheses. On the other hand, meta-analysis can be used to account for observed variance by computing a myriad of post hoc tests for coded potential moderators, which will lead to numerous Type I and Type II errors (Hunter and Schmidt 2004).

The H&S Approach

According to Hunter and Schmidt (2004), if less than 75 % of variance can be attributed to artifacts, subgroup analysis should be conducted. In such an analysis, effect sizes are grouped into meaningful categories, and each subgroup is separately meta-analyzed. Hunter and Schmidt (2004) suggested that the means and the REVC of the subgroups are examined. If there is a large difference in group means and a corresponding reduction in the REVC from the analysis in which both groups are combined, a moderator is present. The meaning of 'large' may vary depending on the research question.

Such an approach has been criticized on several grounds. First, subgroup analyses may require the subdivision of continuous moderator variables, which limits variation, resulting in lower statistical power to detect moderating effects (Stone-Romero and Anderson 1994). Also, if continuous moderators are converted into categories, variance is restricted, and moderator effects tend to be underestimated (Steel and Kammeyer-Mueller 2002; Stone-Romero and Anderson 1994). Second, different subgroups are often statistically dependent, so that the results of one test can be predicted from the results of another (Hunter and Schmidt 2004, pp. 424–426). Consequently, a result may indicate the existence of a moderating effect when none is present because the result is actually due to another, correlated moderator.

Another technique to explain heterogeneity is study characteristics correlations (Hunter and Schmidt 2004; also called vector correlations, Jensen 1998), which are correlations between effect sizes and moderators. Such correlations describe the covariance between two vectors; one vector is the effect size (e.g., observed or corrected correlation), the second vector is the moderator variable. A vector correlation analysis has the limitation that only one moderator can be analyzed at any given time. Furthermore, when examining various moderator variables separately, multicollinearity cannot be readily addressed, which can lead to erroneous and misleading interpretations of results (Steel and Kammeyer-Mueller 2002).

The H&O Approach

In the H&O statistical approach, the terms study characteristics or vector correlation are not used but the analogous analysis method is meta-regression (Borenstein et al. 2009; Thompson and Higgins 2002). Meta-regression applies the concept of multiple correlation and regression analysis to the sample level. Meta-regression can simultaneously assess the effects of multiple study characteristics (for an illustrative example see Baltes et al. 1999). Thus, meta-regression may be used to test multiple moderators and the incremental value of individual predictors (e.g., the incremental validity of independent variables). As with traditional regression analysis, multiple regression methods are available, including ordinary least squares (OLS) and weighted least squares (WLS). Theoretical arguments and empirical evidence indicate that the WLS[10] method with the inverse sampling variance as weights is the preferred moderator estimation technique (Hedges and Olkin 1985; Steel and Kammeyer-Mueller 2002). It has also been suggested that the preliminary test of heterogeneity (either statistical or the 75 % variety) be abandoned if the moderator in question is hypothesized a priori (Overton 1998).

Comparison of the Two Approaches and Recommendations

Both statistical meta-analytic approaches allow for the estimation of subgroup means and respective residual REVCs. Both also allow for the computation of the correlation between study characteristics and effect sizes (although Hunter and Schmidt recommend creating subgroups rather than computing correlations). Hunter and Schmidt's subgroup analysis is, by far, the most often used approach in the organizational sciences (Geyskens et al. 2009). However, the weighted meta-regression approach has two advantages. First, multiple variables can be considered simultaneously, and both categorical and continuous moderators can be handled. Second, this method includes statistical significance tests for each moderator. On the other hand, special problems arise in the application of such models in the H&O approach. While the weights for means in the random-effects case depend upon the REVC, calculating the means across levels of the moderator involves finding a residual variance (i.e., the REVC

---

[10] We note that the WLS regression procedures in the H&O approach use the inverse variance weight, and the estimation procedure for the standard errors differs from that of regular WLS regression estimation procedures. Thus, standard WLS techniques in software packages such as SPSS and SAS cannot be used to accurately estimate the regression model (Hedges and Olkin 1985), even when appropriate weighting is used. D. Wilson provides computationally correct macros for meta-regression (http://mason.gmu.edu/~dwilsonb/ma.html).

after accounting for the moderator). The REVC may be pooled across levels of the moderator, or estimated separately for each level of the moderator. The choice is up to the meta-analyst, and it may not be obvious which is best. The weights of the H&S approach are always N_i, so no such difficulty arises.

Given the advantages of meta-regression, we recommend its use over subgroup analyses unless the moderating variable is categorical (e.g., gender or race) to explain heterogeneity (see, e.g., Baltes et al. 1999; Conn et al. 2011; Kamdi et al. 2011; Kepes et al. 2012). However, although the ability to model multiple moderators at once is a conceptual advantage, its use in practice could be limited by missing data problems. If one eliminates all samples with missing values on any moderator, one may be left with few samples, low power, and an inability to estimate any parameters of interest with accuracy. Also, unlike the power for the null hypothesis of the overall mean effect size, the power for the tests of moderators is often quite poor (Cafri et al. 2010). Finally, our recommendation is not meant to prevent researchers from investigating means of subgroups created from continuous variables where such groupings have important theoretical or practical interest.

Sensitivity Analyses

Because meta-analytic results can have a substantial impact on a scientific literature stream (Borenstein et al. 2009; Geyskens et al. 2009; Hunter and Schmidt 2004; Schmidt and Hunter 2003), they need to be as accurate as possible. It is thus crucial to assess their validity with sensitivity analyses (Greenhouse and Iyengar 2009; Kepes et al. 2012; Kepes and McDaniel, in press; Rothstein et al. 2005a). The role of such an analysis is to determine whether different assumptions or decisions made during the systematic review process have a substantial effect on the obtained results. Thus, sensitivity analyses address the question, "what happens if aspects of the data or analyses are changed?" (Greenhouse and Iyengar 2009, p. 418). Such analyses involve comparing the results of two or more meta-analyses of the same effect size distribution using different assumptions and/or decisions. To the degree that the implications of the analysis are unchanged by sensitivity analyses, one gains confidence in the meta-analytic results and conclusions. Because of the importance of such analyses, MARS explicitly recommends their use, including the reporting of how missing data were handled in the analysis as well as an assessment of potential causes of non-robustness (e.g., outliers, missing data, publication bias). Regrettably, Aguinis et al. (2011) estimated that only 16 % of meta-analytic reviews conduct sensitivity analyses. Next, we describe four types of decisions about data that are commonly made and should be accompanied by sensitivity

analysis in all meta-analytic reviews: outliers, missing data and imputations, publication bias, and extrapolations.

Outliers

The results and conclusion of a meta-analysis may be heavily influenced by a single, large sample or by one or more effect sizes of deviant magnitude. Unfortunately, only between 3 (Aguinis et al. 2011) and 9 % (Geyskens et al. 2009) of meta-analytic reviews in the organizational sciences examine the influence of outliers on meta-analytic results. Hunter and Schmidt advocate the specific-sample removed analysis. In this analysis, specific samples based on a theoretical or methodological rationale are excluded from the meta-analysis, and the meta-analytic results with and without the excluded samples are compared to assess the robustness of the meta-analytically derived statistics (Hunter and Schmidt 2004). Exploratory data analysis using stem-and-leaf, box, and funnel plots can be used to assess and identify potential problematic samples (e.g., outliers) before excluding them (Greenhouse and Iyengar 2009). Figure SM2 in the supplemental materials provides an example of a contour-enhanced funnel plot, and how it can be used to identify potential outliers. Also, statistics such as the sample-adjusted meta-analytic deviancy statistic (Beal et al. 2002) can be used to identify outliers. Thus, comparisons of meta-analytic results with all samples and results from analyses excluding certain samples serve as sensitivity analyses.

In addition to specific sample removed analyses, a one sample removed analysis is also recommended (Borenstein et al. 2009; Greenhouse and Iyengar 2009). Instead of assessing the sensitivity of the meta-analytic result by removing specific samples (e.g., potentially problematic samples based on a theoretical or methodological rationale), this analysis is performed by removing each individual sample from the meta-analysis, one at a time, and retaining the remaining samples and re-computing the meta-analytic mean effect size. Thus, the influence of each individual sample on the meta-analytically derived effect size is evaluated. This method is appealing if one would like to examine the possible range of results if any individual sample is removed from the analysis.

Missing Data and Imputations

Missing data may be imputed by several approaches. The simplest way is to substitute the mean of the non-missing observations for the missing observation. More sophisticated imputation methods include using regression to estimate missing values, and the use of values from studies not included in the current meta-analysis.

Effect sizes and related statistics, such as reliability coefficients and estimates of range restriction, can be imputed individually. Assuming that sample statistics and statistical artifacts are independent across samples and from the 'true' population effect size, one can calculate the averages of the required statistics using the available information from the samples that provide them, or from prior meta-analytic reviews. Missing information of individual samples is thus replaced by the calculated mean or by previously estimated meta-analytic means. Then, if desired, effect sizes can be corrected and a meta-analysis using individually corrected effect sizes can be performed. Evidence indicates that both approaches typically yield similar results (Hunter and Schmidt 1994, 2004; Law et al. 1994a, b).

Whenever imputations are used, the analysis with and without imputations should be compared. A comparison of the samples without missing information with the ones that have missing information can provide information on the robustness of the meta-analytically derived statistics as well as guidance on the appropriateness of imputations (McDaniel 2005). Similarly, alternative imputation approaches or artifact distributions could be used to assess the robustness of the meta-analytic results. Finally, because coding decisions can lead to different results, it can be particularly useful to conduct sensitivity analyses to show the extent to which different coding decisions affect the effect size estimates (Lipsey and Wilson 2001). Unfortunately, such sensitivity analyses are rarely done in the organizational sciences.

Publication Bias

Publication bias describes a situation in which "the research that appears in the published literature is systematically unrepresentative of the population of completed studies" on a relation of interest (Rothstein et al. 2005a, p. 1). This bias stems from the tendency to submit and/or publish studies with samples based on the direction or statistical significance of the results, rather than on the study design and data collected. As a result, small sample studies with statistically insignificant results could be missing (i.e., "suppressed") from the published literature (Greenwald 1975; Kepes et al. 2012; McDaniel et al. 2006; Rothstein et al. 2005b). Because this could be one of the greatest threats to the validity of meta-analytic results (Rothstein et al. 2005a), MARS addresses publication bias under the topic of data censoring. Despite this, only a minority of meta-analytic reviews in the organizational sciences address the issue of publication bias, with estimations indicating that only between two (Aguinis et al. 2011) and 31 % (Banks et al. 2012b; Kepes et al. 2012) test for this bias. Furthermore, even if assessed, this is typically

done with questionable methods (Banks et al. 2012b; Kepes et al. 2012).

In the organizational sciences, Rosenthal's (1979) or Orwin's (1983) failsafe N analyses are typically used to detect publication bias and thus to assess the stability of meta-analytic results. Despite its frequency of use (Aytug et al. 2011; Banks et al. 2012b; Kepes et al. 2012), the failsafe N cannot be recommended to assess publication bias and therefore the robustness of meta-analytic results (Becker 2005; Higgins and Green 2009; McDaniel et al. 2006). Unfortunately, Dalton et al.'s (2012) recent paper that concluded that publication bias does not pose a threat to the organizational sciences relied on a very similar approach. In addition, the authors did not differentiate between effects sizes on a specific relation of interest and ancillary effect sizes. Thus, one may argue that Dalton et al. (2012) did not, in fact, assess the potential presence of publication bias in the organizational sciences as this bias is concerned with the availability of effect sizes on a particular relation of interest (e.g., Dickerson 2005; Kepes et al. 2012; Rothstein et al. 2005b) and not with the availability of all possible effect sizes in an entire scientific field such as the organizational sciences. Therefore, given the evidence for the existence of publication bias in the organizational and related sciences based on appropriate methods (e.g., Banks et al. 2012a; Banks et al. 2012b; Kepes et al. 2012; Kepes et al., in press; McDaniel et al. 2006; Renkewitz et al. 2011), we recommend a rigorous assessment of publication bias with appropriate methods (Banks et al. 2012b; Kepes et al. 2012; Rothstein et al. 2005b).

Although there are no perfect methods of detecting and correcting for publication bias, multiple graphical (e.g., contour-enhanced funnel plots), regression-based (e.g., Begg and Mazumdar's rank correlation test and Egger's test of the intercept), and other (e.g., trim-and-fill analysis, selection models, and cumulative meta-analysis) methods can be used to examine publication bias (for reviews and recommendations of these and additional methods see, e.g., Kepes et al. 2012; Rothstein et al. 2005a). Figures SM3 and SM4 in the supplemental materials provide illustrative examples of a contour-enhanced funnel plot and two cumulative meta-analyses. Methods, such as trim-and-fill and selection models, also allow for the calculation of a publication bias adjusted mean effect size to examine robustness of the meta-analytic estimate.

Extrapolation

The H&S approach typically involves extrapolation from observed data to estimate what would be observed with perfect reliability and a distribution of scores that was not

observed (to account for range restriction). Because reliability and range restrictions statistics are often imputed for some samples (or sometimes all samples), it is recommended that one conduct sensitivity analyses using varying estimates of artifact (i.e., reliability and range restriction) values. One would want to see the extent of results and conclusions from results that are robust to varying magnitudes of artifact values and varying approaches for estimating artifact values. At a minimum, if the analysis of corrected values is presented, then uncorrected values should also be presented.

Results and Discussion

The results and discussion sections of a meta-analytic review should contain sufficient detail to allow the reader to assess the accuracy of the results and conclusions. Although basic guidelines for the reporting of results are provided for all sciences (e.g., Borenstein et al. 2009; Cooper 1998; Cooper et al. 2009; Egger et al. 2001a; Higgins and Green 2009; Hunter and Schmidt 2004), meta-analytic reviews in the organizational sciences rarely provide enough information to assess the accuracy of the results. Furthermore, the transparency and replicability of the review is often questionable. For instance, detailed information regarding coding decisions and descriptive statistics for each primary sample (e.g., sample size, effect size, reliability coefficients, and standard deviations) is typically missing. Meta-analytic reviews in the medical and related sciences typically provide such information. As mentioned previously, reporting practices in other sciences also place an additional emphasis on reporting results from sensitivity analyses, including outlier and publication bias analyses (Borenstein et al. 2009; Higgins and Green 2009; Kepes et al. 2012).

Graphical displays (Cooper 1998; Rothstein 2003), such as box, funnel, and forest plots, are recommended by MARS, but meta-analytic reviews published in the organizational sciences almost never implement this guidance (Aytug et al. 2011). Specific methods such as the contour-enhanced funnel plot can also be used to visually display the results of publication bias analyses (Kepes et al. 2012). Forest plots are also an excellent way to communicate information about individual effect sizes and the overall meta-analytic results (see Figs. SM4 and SM5 in the supplemental materials).

In addition, MARS mentions the discussion of the results for a sample quality assessment, if used. As described previously, we do not recommend such quality assessments but, instead, the coding of methodological factors (study design, scale type, etc.; Lipsey and Wilson 2001). Such a recommendation is consistent with MARS.

Based on the a priori coding of such methodological features, moderator analyses should be conducted to assess the robustness of the meta-analytic results (Kepes et al. 2012; McDaniel et al. 2007). Currently, examinations of the impact of methodological features on meta-analytic results are infrequently conducted (Aytug et al. 2011). Similarly, the study design (e.g., concurrent or predictive design) is rarely disclosed or even examined as a moderating variable (Aytug et al. 2011) although it can have a substantial effect on the obtained results (Kepes et al. 2012). Consistent with MARS, we also recommend the more comprehensive communication of meta-analytic input data (study and sample characteristics) and results (see, e.g., Judge et al. 2001; Kamdi et al. 2011; Miller et al. 2008). Table SM1 in the supplemental materials provides a template for the comprehensive and transparent communication of meta-analytic input data.

Conclusions

Meta-analytic reviews in the organizational sciences currently fall short of following well established meta-analytic guidelines (e.g., Borenstein et al. 2009; Cooper 1998; Cooper et al. 2009; Egger et al. 2001a; Higgins and Green 2009; Hunter and Schmidt 2004) as well as MARS (APA 2008, 2010). As a result, many meta-analytic reviews in our sciences are not adequately transparent and replicable, which is one of the hallmarks of meta-analytic reviews (Cooper and Hedges 2009; Egger et al. 2001b). Furthermore, current meta-analytic practices may hurt the accuracy of meta-analytic results and conclusions. In this paper, we focused on issues that are of importance to improving the accuracy, transparency, and replicability of the meta-analytic review and/or have been previously identified as being problematic in the organizational sciences (e.g., Aytug et al. 2011; Cooper 1998; Cooper and Hedges 2009; Egger et al. 2001b; Geyskens et al. 2009). We illustrated how the integration of "best practices" from two schools of meta-analysis can be applied to improve the quality of quantitative reviews in the organizational sciences and tied our recommendations to MARS. Table 1 presents a summary of our recommendations following the structure provided by MARS. In addition, the supplemental materials include several illustrative templates that are aligned with these recommendations.

We note that several of our recommendations go beyond MARS. Recommendations such as the more widespread use of prospective meta-analytic reviews in the organizational sciences, a detailed description of the treatment of non-published samples, the provision of supplemental materials on journal websites, the development of research registries, the estimation and reporting of the REVC and

prediction intervals, and the requiring of comprehensive sensitivity analyses with appropriate methods are not necessarily part of MARS. However, these and other practices are used in other scientific disciplines and, if adopted, would improve the accuracy, transparency, and replicability of our meta-analytic reviews.

Table 1 also includes references to systematic reviews that provide examples for the implementation of our recommendations. The individual papers are intended as models only for the particular aspects of the review noted in the table. Just as with primary research, meta-analytic reviews may be outstanding in some places and deficient in others. We also note that not all examples provided in Table 1 fully conform to the specific recommendation for which they appear. Still, the examples provided in Table 1 come at least close to having implemented the particular recommendation for which they are referenced. Overall, the adoption of the recommended practices should provide a firmer foundation of cumulative knowledge upon which to build better theory and practice. The topics covered in this paper are those we consider to be of utmost immediate benefit to organizational researchers. However, we note that MARS and other guidelines for meta-analytic reviews (e.g., Borenstein et al. 2009; Cooper 1998; Cooper et al. 2009; Egger et al. 2001a; Higgins and Green 2009; Hunter and Schmidt 2004) include a variety of recommendations beyond the scope of this paper. Meta-analysts as well as editors and reviewers of meta-analytic reviews should consult these and other sources as well when writing or evaluating a meta-analytic review.

Best practice in meta-analysis and the systematic review continues to evolve. We described and integrated several practices, ideas, concepts, and techniques across scientific areas and statistical approaches to help organizational researchers keep abreast of current developments. We integrated two main approaches to systematic and meta-analytic reviews in order to improve the quality of such reviews as a whole. Better meta-analytic reviews will result in better inferences, and ultimately may provide the best opportunity to close the often lamented gap between research and practice (Le et al. 2007).

References

Aguinis, H., & Pierce, C. A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management, 24*, 577–592. doi:10.1016/s0149-2063(99)80074-9.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management, 37*, 5–38. doi:10.1177/0149206310377113.

American Psychological Association. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*, 839–851. doi:810.1037/0003-1066X.1063.1039.1839.

American Psychological Association. (2010). *Publication manual of the American psychological association* (6th ed.). Washington, DC: American Psychological Association.

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2011). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods, 15*, 103–133. doi:10.1177/1094428111403495.

Baltes, B. B., Briggs, T. E., Huff, J. W., Wright, J. A., & Neuman, G. A. (1999). Flexible and compressed workweek schedules: A meta-analysis of their effects on work-related criteria. *Journal of Applied Psychology, 84*, 496–513. doi:10.1037/0021-9010.84.4.496.

Banks, G. C., & McDaniel, M. A. (2011). The kryptonite of evidence-based I–O psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 40–44. doi:10.1111/j.1754-9434.2010.01292.x.

Banks, G. C., Batchelor, J. H., & McDaniel, M. A. (2010). Smarter people are (a bit) more symmetrical: A meta-analysis of the relationship between intelligence and fluctuating asymmetry. *Intelligence, 38*, 393–401. doi:10.1016/j.intell.2010.04.003.

Banks, G. C., Kepes, S., & Banks, K. P. (2012a). Publication bias: The antagonist of meta-analytic reviews and effective policy making. *Educational Evaluation and Policy Analysis, 34*, 259–277. doi:10.3102/0162373712446144.

Banks, G. C., Kepes, S., & McDaniel, M. A. (2012b). Publication bias: A call for improved meta-analytic practice in the organizational sciences. *International Journal of Selection and Assessment, 20*, 182–196. doi:10.1111/j.1468-2389.2012.00591.x.

Bax, L., Yu, L.-M., Ikeda, N., Tsuruta, N., & Moons, K. G. (2006). Development and validation of MIX: Comprehensive free software for meta-analysis of causal research data. *BMC Medical Research Methodology, 6*, 50. doi:10.1186/1471-2288-6-50.

Bax, L., Yu, L.-M., Ikeda, N., & Moons, K. G. (2007). A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology, 7*, 40–48. doi:10.1186/1471-2288-7-40.

Beal, D. J., Corey, D. M., & Dunlap, W. P. (2002). On the bias of Huffcutt and Arthur's (1995) procedure for identifying outliers in the meta-analysis of correlations. *Journal of Applied Psychology, 87*, 583–589. doi:10.1037/0021-9010.87.3.583.

Becker, B. J. (2005). The failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment, and adjustments* (pp. 111–126). West Sussex: Wiley.

Berlin, J. A., & Ghersi, D. (2005). Preventing publication bias: Registeries and prospective metaanalysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment, and adjustments*. West Sussex: Wiley.

Berman, N. G., & Parker, R. A. (2002). Meta-analysis: Neither quick nor easy. *BMC Medical Research Methodology, 2*, 10. doi:10.1186/1471-2288-2-12.

Bobko, P., & Roth, P. L. (2008). Psychometric accuracy and (the continuing need for) quality thinking in meta-analysis. *Organizational Research Methods, 11*, 114–126. doi:10.1177/1094428107303155.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others*. Boca Ratton: Chapman and Hall/CRC.

Bonett, D. G. (2008). Meta-analytic interval estimation for Pearson correlations. *Psychological Methods, 13*, 173–189. doi:10.1037/a0012868.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2005). *Comprehensive meta-analysis (Version 2)*. Englewood: Biostat.

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (in press). *Computing effect sizes for meta-analysis*. West Sussex: Wiley.

Brannick, M. T. (2001). Implications of empirical Bayes meta-analysis for test validation. *Journal of Applied Psychology, 86*, 468–480. doi:10.1037/0021-9010.86.3.468.

Brannick, M. T., Yang, L.-Q., & Cafri, G. (2011). Comparison of weights for meta-analysis of r and d under realistic conditions. *Organizational Research Methods, 14*, 587–607. doi:10.1177/1094428110368725.

Briner, R. B., & Denyer, D. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. In D. M. Rousseau (Ed.), *The Oxford handbook of evidence-based management*. New York: Oxford University Press.

Briner, R. B., & Rousseau, D. M. (2011). Evidence-based I–O psychology: Not there yet. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 3–22. doi:10.1111/j.1754-9434.2010.01287.x.

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45*, 239–270. doi:10.1080/00273171003680187.

Carey, K. B., Scott-Sheldon, L. A. J., Carey, M. P., & DeMartini, K. S. (2007). Individual-level interventions to reduce college student drinking: A meta-analytic review. *Addictive Behaviors, 32*, 2469–2494. doi:10.1016/j.addbeh.2007.05.004.

Cholesterol Treatment Trialists' Collaborators. (2005). Efficacy and safety of cholesterol-lowering treatment: Prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *The Lancet, 366*, 1267–1278. doi:10.1016/s0140-6736(05)67394-1.

Conn, V. S., Hafdahl, A. R., & Mehr, D. R. (2011). Interventions to increase physical activity among healthy adults: Meta-analysis of outcomes. *American Journal of Public Health, 101*, 751–758. doi:10.2105/ajph.2010.194381.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks: 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. 4–15). New York: Russell Sage Foundation.

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

Costanza, D. P., Badger, J. M., Fraser, R. L., Severt, J. B., & Gade, P. A. (2012). Generational differences in work-related attitudes: A meta-analysis. *Journal of Business and Psychology, 27*, 375–394. doi:10.1007/s10869-012-9259-4.

Crane, D. (1967). The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist, 2*, 195–201. doi:10.2307/27701277.

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 non-published correlation matrices. *Personnel Psychology, 65*, 221–249. doi:10.1111/j.1744-6570.2012.01243.x.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177–188. doi:10.1016/0197-2456(86)90046-2.

Dickerson, K. (2005). Publication bias: Recognizing the problem, understandings 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). West Sussex: Wiley.

Dulebohn, J. H., Bommer, W. H., Liden, R. C., Brouer, R. L., & Ferris, G. R. (2012). A meta-analysis of antecedents and consequences of leader-member exchange: Integrating the past with an eye toward the future. *Journal of Management, 38*, 1715–1759. doi:10.1177/0149206311415280.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193–242. doi:10.1037/h0044139.

Egger, M., Smith, G. D., & Altman, D. (2001a). *Systematic reviews in health care: Meta-analysis in context* (2nd ed.). London: BMJ Books.

Egger, M., Smith, G. D., & O'Rourke, K. (2001b). Rationale, potentials, and promise of systematic reviews. In M. Egger, G. D. Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 3–22). London: BMJ Books.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior & Human Performance, 13*, 171–192. doi:10.1016/0030-5073(75)90044-6.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103–127. doi:10.1037/a0018053.

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. doi:10.1096/fj.05-4784lsf.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*, 161–180. doi:10.1037/1082-989X.6.2.161.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods, 10*, 444–467. doi:10.1037/1082-989x.10.4.444.

Geyskens, I., Krishnan, R., Steenkamp, J.-B. E. M., & Cunha, P. V. (2009). A review and evaluation of meta-analysis practices in management research. *Journal of Management, 35*, 393–419. doi:10.1177/0149206308328501.

Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 417–433). New York: Russell Sage Foundation.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20. doi:10.1037/h0076157.

Gurusamy, K., Aggarwal, R., Palanivelu, L., & Davidson, B. R. (2008). Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *British Journal of Surgery, 95*, 1088–1097. doi:10.1002/bjs.6344.

Hafdahl, A. R., & Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods, 14*, 24–42. doi:10.1037/a0014697.

Halbert, R. J., Natoli, J. L., Gano, A., Badamgarav, E., Buist, A. S., & Mannino, D. M. (2006). Global burden of COPD: Systematic review and meta-analysis. *European Respiratory Journal, 28*, 523–532. doi:10.1183/09031936.06.00124605.

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology, 87*, 377–389. doi:10.1037/0021-9010.87.2.377.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486–504. doi:10.1037/1082-989x.3.4.486.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*, 405–411. doi:10.1111/j.1468-2389.2007.00399.x.

Higgins, J. P. T., & Green, S. (Eds.). (2009). *Cochrane handbook for systematic reviews of interventions* (Version 5.0.2, updated September 2009). The Cochrane Collaboration. Retrieved from www.cochrane-handbook.org.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539–1558. doi:10.1002/sim.1186.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560. doi:10.1136/bmj.327.7414.557.

Hoffman, B. J., Blair, C. A., Meriac, J. P., & Woehr, D. J. (2007). Expanding the criterion domain? A quantitative review of the OCB literature. *Journal of Applied Psychology, 92*, 555–566. doi:10.1037/0021-9010.92.2.555.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*, 193–206. doi:10.1037/1082-989x.11.2.193.

Hunter, J. E., & Schmidt, F. L. (1994). Estimation of sampling error variance in the meta-analysis of correlations: Use of average correlation in the homogeneous case. *Journal of Applied Psychology, 79*, 171–177. doi:10.1037/0021-9010.79.2.171.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275–292. doi:10.1111/1468-2389.00156.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park: Sage.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594–612. doi:10.1037/0021-9010.91.3.594.

Indrayan, A. (2008). *Medical biostatistics* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Ioannidis, J. P. A. (2010). Meta-research: The art of getting it wrong. *Research Synthesis Methods, 1*, 169–184. doi:10.1002/jrsm.19.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport: Praeger.

Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin, 127*, 376–407. doi:10.1037/0033-2909.127.3.376.

Kamdi, A. S., Kandavalli, N. B., Emusu, D., Jain, N., Mamtani, M., & Porterfield, J. R. (2011). Weak or absent evidence for the association of HLA-DR antigens with risk of thyroid carcinoma: A meta-analysis of observational studies. *Tissue Antigens, 78*, 382–389. doi:10.1111/j.1399-0039.2011.01754.x.

Kemery, E. R., Mossholder, K. W., & Dunlap, W. P. (1989). Meta-analysis and moderator variables: A cautionary note on transportability. *Journal of Applied Psychology, 74*, 168–170. doi:10.1037/0021-9010.74.1.168.

Kepes, S., & McDaniel, M. A. (in press). How trustworthy is the scientific literature in I-O psychology? *Industrial and Organizational Psychology: Perspectives on Science and Practice*.

Kepes, S., Banks, G. C., McDaniel, M. A., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15*, 624–662. doi:10.1177/1094428112452760.

Kepes, S., Banks, G. C., & Oh, I.-S. (in press). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology*. doi:10.1007/s10869-012-9279-0.

Kisamore, J. L. (2008). Distributional shapes and validity transport: A comparison of lower bounds. *International Journal of Selection and Assessment, 16*, 27–29. doi:10.1111/j.1468-2389.2008.00406.x.

Kisamore, J. L., & Brannick, M. T. (2008). An illustration of the consequences of meta-Analysis model choice. *Organizational Research Methods, 11*, 35–53. doi:10.1177/1094428106287393.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162–181. doi:10.1037/0033-2909.127.1.162.

Laine, C., et al. (2007). Clinical trial registration: Looking back and moving ahead. *New England Journal of Medicine, 356*, 2734–2736. doi:10.1056/NEJMe078110.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994a). Nonlinearity of range corrections in meta-analysis: Test of an improved procedure. *Journal of Applied Psychology, 79*, 425–438. doi:10.1037/0021-9010.79.3.425.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994b). A test of two refinements in procedures for meta-analysis. *Journal of Applied Psychology, 79*, 978–986. doi:10.1037/0021-9010.79.6.978.

Le, H., Oh, I.-S., Shaffer, J., & Schmidt, F. L. (2007). Implications of methodological advances for the practice of personnel selection: How practitioners benefit from meta-analysis. *Academy of Management Perspectives, 21*, 6–15.

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

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement, 70*, 56–73. doi:10.1177/0013164409344534.

McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence, 33*, 337–346. doi:10.1016/j.intell.2004.11.005.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599–616. doi:10.1037/0021-9010.79.4.599.

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:10.1111/j.1744-6570.2006.00059.x.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91. doi:10.1111/j.1744-6570.2007.00065.x.

Miller, A. J., Worthington, E. L., & McDaniel, M. A. (2008). Gender and forgiveness: A meta-analytic review and research agenda. *Journal of Social and Clinical Psychology, 27*, 843–876. doi:10.1521/jscp.2008.27.8.843.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157–159. doi:10.2307/1164923.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*, 354–379. doi:10.1037/1082-989x.3.3.354.

Park, T.-Y., & Shaw, J. D. (in press). Turnover rates and organizational performance: A meta-analysis. *Journal of Applied Psychology*. doi:10.1037/a0030723.

Puts, D. A., McDaniel, M. A., Jordan, C. L., & Breedlove, S. M. (2008). Spatial ability and prenatal androgens: Meta-analyses of congenital adrenal hyperplasia and digit ratio (2D:4D) studies. *Archives of Sexual Behavior, 37*, 100–111. doi:10.1007/s10508-007-9271-3.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology, 76*, 432–446. doi:10.1037/0021-9010.76.3.432.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*, 75–98. doi:10.2307/1164836.

Renkewitz, F., Fuchs, H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and Decision Making, 6*, 870–881.

Richardson, K. M., & Rothstein, H. R. (2008). Effects of occupational stress management intervention programs: A meta-analysis. *Journal of Occupational Health Psychology, 13*, 69–93. doi:10.1037/1076-8998.13.1.69.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641. doi:10.1037/0033-2909.86.3.638.

Rosenthal R. (1991). *Meta-analytic procedures for social research* (Revised ed.). Newbury Park: Sage.

Roth, P. L. (2008). Software review: Hunter-Schmidt meta-analysis programs 1.1. *Organizational Research Methods, 11*, 192–196. doi:10.1177/1094428106298972.

Rothstein, H. R. (2003). Progress is our most important product: Contributions of validity generalization and meta-analysis to development communication of knowledge in I/O psychology. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 115–154). Mahwah: Lawrence Erlbaum.

Rothstein, H. R. (2012). Accessing relevant literature. In H. M. Cooper (Ed.), *APA handbook of research methods in psychology*: *Foundations, planning, measures, and psychometrics* (Vol. 1, pp. 133–144). Washington: American Psychological Association.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005a). Publication bias in meta-analyses. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment, and adjustments* (pp. 1–7). West Sussex: Wiley.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005b). *Publication bias in meta analysis: Prevention, assessment, and adjustments*. West Sussex: Wiley.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: a Monte Carlo investigation of statistical power and resistance to type I error. *Journal of Applied Psychology, 71*, 302–310. doi:10.1037/0021-9010.71.2.302.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540. doi:10.1037/0021-9010.62.5.529.

Schmidt, F. L., & Hunter, J. E. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975–2001. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 31–65). Mahwah: Lawrence Erlbaum.

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62*, 97–128. doi:10.1348/000711007x255327.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge: Hogrefe & Huber.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15*, 5–11. doi:10.3102/0013189X015009005.

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to type I and type II errors. *Journal of Applied Psychology, 72*, 3–9. doi:10.1037/0021-9010.72.1.3.

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87*, 96–111. doi:10.1037/0021-9010.87.1.96.

Steel, P. D., & Kammeyer-Mueller, J. D. (2008). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods, 11*, 54–78. doi:10.1177/1094428107300339.

Stone-Romero, E. F., & Anderson, L. E. (1994). Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderating effects. *Journal of Applied Psychology, 79*, 354–359. doi:10.1037/0021-9010.79.3.354.

Sutton, A. J. (2005). Evidence concerning the consequences of publication and related biases. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment, and adjustments* (pp. 175–192). West Sussex: Wiley.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. London: Wiley.

Terrizzi, J. A., Shook, N. J., & McDaniel, M. A. (2013). The behavioral immune system and social conservatism: A meta-analysis. *Evolution and Human Behavior, 34*, 99–108. doi:10.1016/j.evolhumbehav.2012.10.003.

Thomas, H. (1988). What is the interpretation of the validity generalization estimate $S_p^2 = S_r^2 - S_e^2$? *Journal of Applied Psychology, 73*, 679–682. doi:10.1037/0021-9010.73.4.679.

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*, 1559–1573. doi:10.1002/sim.1187.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology, 97*, 499–530. doi:10.1037/a0021196.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26*, 37–52. doi:10.1002/sim.2514.

Visweswaran, C., & Sanchez, J. I. (1998). Moderator search in meta-analysis: A review and cautionary note on existing approaches. *Educational and Psychological Measurement, 58*, 77–87. doi:10.1177/0013164498058001007.

White, H. D. (2009). Scientific communication and literature retrieval. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Research synthesis and meta-analysis* (2nd ed., pp. 51–71). New York: Russell Sage Foundation.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75*, 315–321. doi:10.1037/0021-9010.75.3.315.