Review of General Psychology
2009, Vol. 13, No. 2, 101–115

1089-2680/09/$12.00 DOI: 10.1037/a0015107

An Empirical Assessment of Meta-Analytic Practice

Nathan F. Dieckmann
Decision Research & University of Oregon

Bertram F. Malle
University of Oregon

Todd E. Bodner
Portland State University

In the three decades after the publication of the first meta-analyses in the behavioral sciences, hundreds of articles and a number of technical guides have emerged concerning meta-analytic practice and reporting standards. The purpose of the present study is to review the practice and reporting standards of a random sample of published meta-analyses ($n = 100$) in psychology and related disciplines in the decade from 1994 through 2004. We focus on practice and reporting at each stage of the meta-analytic process and explore differences between psychological subdisciplines. These findings suggest that the practice of meta-analysis in the last decade has not yet converged on a set of common standards, though some expert recommendations are beginning to be heeded. Authors should be attentive to proper procedure and reporting in light of the numerous threats to the validity of a meta-analysis. Ironically, even though meta-analysts often struggle with incomplete or inconsistent reporting in primary research they are themselves not entirely consistent in reporting their methods and results.

Keywords: meta-analysis, methodology, research synthesis, literature review

Traditionally, narrative reviews have served to compile and synthesize research literatures although in recent years more quantitative approaches have gained popularity. Meta-analysis can be defined as a cluster of analytic techniques used to quantitatively synthesize results from multiple research reports. Although the word has been used more broadly in some contexts, we reserve the term meta-analysis for research integrations that employ specific quantitative techniques. Although not without problems, meta-analytic techniques offer a more objective approach than narrative reviewing (Beaman, 1991; Cook & Leviton, 1980). The systematic methodology and explicit reporting of method are often cited as the main advantages of the meta-analytic approach.

Literature reviews, and particularly meta-analyses, are playing an increasingly important role in scientific literatures. Literature reviews are widely read and correspondingly have high citation rates compared to primary research reports (Mazela & Malin, 1977; Smith & Caulley, 1981). As Becker (1991) noted, "publication of a review of research in a prestigious journal provides the researcher with an increased potential to influence the thinking of his or her colleagues" (p. 267). In addition, meta-analytic reviews have an increasing role in shaping public policy (Hunt, 1997; Hunter & Schmidt, 1996).

Considering the crucial scientific function and the substantial influence of meta-analytic reviews, it is important to review the methodological rigor and reporting standards of current published meta-analyses. We were also interested in the characteristics of a "typical" meta-analysis in the behavioral sciences. How many primary study results are typically reviewed in a meta-analysis? Are different subdisciplines more likely to use meta-analysis? How is the "typical" meta-analysis conducted and reported and does meta-analytic practice vary by subdiscipline?

The goal of this review is to answer these questions empirically by examining a random sample of published meta-analyses in psychology and related disciplines in the decade from 1994 through 2004. We focus on the extent to which authors adhere to common practice recommendations, as well as whether there are standard reporting practices among meta-analysts. We then compare our empirical results from actual meta-analytic practice to common recommendations for practice and reporting. Where discrepancies arise, we hope to alert methodologists, journal reviewers and editors, and practicing meta-analysts that more focus should be placed on these issues. In addition, particular subdisciplines may be more likely to use meta-analysis and may have different practice and reporting standards. Previous research has shown that the research designs, participants, and measurement methods used in primary research do differ between subdisciplines (Bodner, 2006). A better understanding of how meta-analysis is used by different subdisciplines could help meta-analysts direct improvements in particular subfields.

A few published empirical reviews of meta-analytic practice exist. For example, Steiner, Lane, Dobbins, Schnur, and McConnell (1991) reviewed 35 meta-analyses within the organizational behavior and human resources management literature; Beaman (1991) compared traditional narrative reviews with meta-analyses in a sample drawn from *Psychological Bulletin*, and Light, Singer, and Willett (1994) reviewed visual presentation methods also in a sample from *Psycho-*

Nathan F. Dieckmann, Decision Research, Eugene, Oregon, and Department of Psychology, University of Oregon; Bertram F. Malle, Department of Psychology, University of Oregon; Todd E. Bodner, Department of Psychology, Portland State University.

Bertram Malle is now at Brown University.

Correspondence concerning this article should be addressed to Nathan F. Dieckmann, Decision Research, 1201 Oak Street, Eugene, OR 97401. E-mail: ndieckmann@decisionresearch.org

logical Bulletin.[1] In the present paper we aim to present a broader assessment of meta-analysis, not limited to a particular aspect of meta-analytic practice or to a particular journal or subdiscipline.

Practice and Reporting in Meta-Analytic Work

Table 1 is a brief taxonomy of selected aspects of meta-analytic practice and reporting. We have identified several common guidelines for practice and reporting and tried to identify the aspects of practice that if insufficiently dealt with can threaten the validity of meta-analytic conclusions. Our goal is not to present an exhaustive list of practice recommendations for meta-analysis, as several methodological guides are available (Cook & Leviton, 1980; Cooper, 1989; Cooper & Hedges, 1994; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Light & Pillemer, 1984; Lipsey & Wilson, 2001; Rosenthal, 1991; Wolf, 1986).

There are multiple factors that can threaten the validity of conclusions drawn from a meta-analysis. Some of these threats are unique to meta-analytic methods, whereas others are familiar to primary researchers. Although we do not present an exhaustive assessment of validity threats, we will focus on some of the major threats that appear at each stage of the meta-analytic process. Validity threats have been discussed by many authors and have been summarized by Matt and Cook (1994).

Explicit reporting is cited as one of the main advantages of meta-analysis over narrative reviews (e.g., see Beaman, 1991). Meta-analysts are expected to be explicit about their operational definitions, literature search procedures, and so forth. The first goal of exhaustive reporting in any research report is ". . .to include enough information about the study that the reader can critically examine the evidence" (Jackson, 1980, p. 456). For example, Jackson (1980) points out the importance of comprehensive reporting of the literature search process, and clearly referencing the sources included in the review. This will allow readers to judge the breadth and appropriateness of the sample, and allow future reviewers to add to the existing set of primary reports. The incomplete reporting of procedures can actually threaten the validity of a meta-analysis by reducing the likelihood that the conclusions of the review will be replicated (Cooper, 1989).

Table 1
Selected Common Guidelines for Practice and Reporting of Meta-Analyses

Practice
Exhaustive search of the literature
Include unpublished work
File drawer analysis
Use multiple coders
Account for nonindependence of effects
Weight effects proportional to their precision
Correct for statistical artifacts
Justify statistical model (e.g., fixed vs. random)
Address issues of statistical power
Reporting
Report literature search methods
Report inclusion criteria
Report a complete list of primary reports (or make otherwise available)
Report a complete table of study results (or make otherwise available)
Report the number of coders
Report a measure of coding reliability
Fully report coding scheme
Report CIs
Report a measure of dispersion and/or homogeneity test statistic
Fully discuss limitations of primary data, method, conclusions, etc.

We are not suggesting, however, that practice and reporting standards for meta-analyses should be rigid and that all meta-analyses should look the same. It is clear that the analytic approach and subsequent reporting of a meta-analysis will depend on the specific phenomenon under review and the data available in the primary literature. However, practicing meta-analysts should heed the practice recommendations offered by methodologists. If a meta-analyst deviates from a common practice, it is important to inform the reader how the issue was handled and why the recommendation was not followed. There may be a variety of valid reasons for using alternative meta-analytic techniques, but very few valid reasons for not telling the reader what techniques were used. Authors should not guard against criticism by withholding information about procedures (White, 1994).

In a review of this kind, it is often difficult to distinguish what was practiced from what was reported. For this reason, we will explore several areas of meta-analytic practice while being careful not to assume that something was not practiced just because it was not reported. Of course, this problem does not exist for elements of practice that are consistently reported. Next, we briefly outline the selected practice and reporting recommendations that guided our evaluation of meta-analytic practice.

Searching the Literature and Inclusion Criteria

The main concerns at this stage of a meta-analysis are conducting an exhaustive search of the literature, specifying appropriate inclusion criteria, and addressing possible publication bias. An exhaustive search of the literature is recommended to reduce the possibility of missing critical primary studies and avoiding the criticism of having produced a biased review (Begg, 1994; Rosenthal & Dimatteo, 2001). Any single source is likely to contain only a small proportion of the available studies, and most sources (such as electronic databases) will not include the most recently published or unpublished research (Cooper & Hedges, 1994a). In addition, it is important for researchers to explicitly report the retrieval methods used. In the best case, this will make the thorough search procedures explicit and guard against attacks of a biased sample, and in the worst case, it will appropriately delimit the generalizability of the results.

The meta-analyst must then construct criteria to identify those studies that will be included in the meta-analytic review. Although there may be disagreements about the actual inclusion criteria that should be used, there is general agreement that the inclusion criteria used be made explicit in the meta-analytic report. Differences between reviewers at this stage of the process can greatly change the eventual conclusions drawn from a research literature (see Wanous, Sullivan, & Malinak, 1989, for a discussion of this and other judgment calls in meta-analytic work). In fact, the use of

[1] A well cited article by Jackson (1980) is not discussed here because he assessed the methods in a sample of integrative reviews, and we limited our sample to those reviews that used quantitative methods (meta-analyses).

explicit criteria for including studies has been put forth as one of the advantages of meta-analysis over traditional reviews (see Cook & Leviton, 1980).

Publication bias, also known as the file-drawer problem, can have drastic effects on the conclusions drawn from a meta-analysis (e.g., Begg, 1994; Rosenthal, 1979). The problem stems from the fact that primary authors who find large significant effects are more likely to publish their results than those researchers who do not find such effects. To the extent that unpublished studies exist, this can create a bias in the estimation of effects if only published studies are included in a meta-analytic review.[2] In many cases it may be impossible to locate unpublished reports, and in these cases meta-analysts should at least estimate the potential effect of publication bias on the results by using any one of a number of techniques (see Begg, 1994). We will review the use of only two procedures—namely, preliminary analysis by constructing a "funnel display" (Light & Pillemer, 1984), and the use of the "file-drawer analysis" (Rosenthal, 1979). In addition, to facilitate replication by other researchers a complete list of the primary reports included in the meta-analysis and a complete list of study results extracted from these reports should be reported or made otherwise available by the authors.

Coding the Primary Reports

Extracting the relevant data from the study sample that has been collected is likely to be the most time-consuming and demanding task of the meta-analytic process (e.g., Lipsey & Wilson, 2001; see Cooper & Hedges, 1994, for a detailed discussion). Unreliability in the coding procedures adds additional random variation to the analysis, weakening the reliability and power of the results. At a very basic level, this can be addressed by employing multiple coders and assessing interrater reliability. Unreliable coding processes introduce measurement error that can degrade statistical power to detect effects of interest. We focus on three fundamental recommendations made by the majority of meta-analytic methodologists regarding coding methods and reporting about such methods: (a) using multiple coders and reporting the number of coders used; (b) reporting a measure of coding reliability; and (c) explicitly describing the moderator coding scheme used.

Analysis and Reporting Results

The analytic techniques used for a given meta-analysis depend on the data available for synthesis and the particular questions posed by the researcher. Consequently, the information that is reported in the results section of a meta-analysis will also vary. We focus on some general aspects of meta-analytic practice, as well as basic descriptive reporting and the use of visual displays.

Statistical Independence of Effect Sizes

The lack of statistical independence of study results can affect combined estimates of effect size and their associated error as studies with more effect sizes included in the meta-analysis are implicitly given greater weight and violate the assumptions of many traditional significance tests. There are many ways of dealing with the problem of statistical dependence between effect sizes (see Gleser & Olkin, 1994). Our interest in this review is not what particular technique was used to deal with any dependencies, but whether the problem was addressed at all.

Score Unreliability and Range Restriction

Every study in a meta-analytic sample suffers from imperfections that influence effect size estimates and their associated standard errors. For example, random error is introduced when the implementation or measurement of primary study variables is unreliable, or when the range of the outcome variable is restricted (see Hunter & Schmidt, 1990, for a thorough list of statistical artifacts). Thus, average effects sizes may be biased and heterogeneity of effect sizes may reflect differences in score reliability and range restriction across studies rather than reflecting inherent variability in the true effect sizes. These problems have prompted the development of methods to estimate and correct for these statistical artifacts (Hunter & Schmidt, 1990). There is not universal agreement on this point, however, and some authors have argued that the goal of meta-analysis is to describe the landscape of actual study results, not to describe what would happen if we could conduct a perfect study.

Weighting Effect Sizes by Precision

The failure to weight results by their precision (e.g., weighting individual effect sizes by sample size or the inverse of the variance) can also affect combined estimates of effect size and their associated error (Matt & Cook, 1994). To our knowledge, most methodologists suggest weighting by precision, a notable exception is Hunter and Schmidt (1990) (see Johnson et al., 1995, for a comparison of major meta-analytic approaches). There are circumstances, however, in which weighting studies with larger samples more could lead to misleading average effects. For example, this could happen if the studies with the larger sample sizes used different methodologies than the remaining studies in the sample. Weighting by sample size would lead to average effect sizes that were biased toward specific methodologies. Thus, meta-analysts should always assess whether sample size is confounded with other characteristics of the study sample before weighting by sample size. If there are good reasons not to weight by precision, this should be explicitly reported in the meta-analysis.

Reporting Confidence Intervals

Several authors recommend the use of confidence intervals (CIs) in reporting mean effect sizes (see Cooper, 1989; Halvorsen, 1994; Lipsey & Wilson, 2001; Rosenthal, 1995; Shadish & Haddock, 1994). This issue is related to the continued call for the increased use of confidence intervals as opposed to the exclusive reporting of significance tests in primary data analysis (America Psychological Association, 2001; Wilkinson, 1999). We were interested in how often meta-analysts report CIs in their reports.

[2] Some authors have argued that unpublished reports are more likely to be underpowered and methodologically weaker, and therefore should be excluded from the review (see Chalmers, Levin, Sacks, Reitman, Berrier, & Nagalingam, 1987). In addition, see Kraemer, Yesavage, Gardner, and Brooks (1998) for a discussion of excluding underpowered studies.

Variability of Effect Distributions

Rosenthal (1995) suggests that at least one standard measure of variability (e.g., *SD* or variance) be reported along with measures of central tendency. Many authors also recommend formally testing the homogeneity of the effect distribution (e.g., Halvorsen, 1994; Hedges, 1982; Hedges & Olkin, 1985; Light & Pillemer, 1984). These tests can be helpful in assessing whether the study effects should be viewed as estimating a common population effect size and can highlight additional variability in the effect distribution that can be investigated further with moderator analysis (although significant heterogeneity is not a prerequisite for focused moderator analysis, see discussion in Rosenthal, 1995).

Statistical Model for Effect Sizes

A full discussion of the rationale and the differences between random and fixed effects models is presented in other sources (see Hedges, 1994b; Hedges & Vevea, 1998; Raudenbush, 1994). The important point for the present work is that a meta-analyst should explicitly justify their choice of statistical model. This decision will be based primarily on the researcher's conceptualization of the processes involved in generating the effects of interest, and second on the results of statistical tests applied to the meta-analytic sample. The choice of statistical model will affect both the results of the review as well as the appropriate generalization from the results. Incorrect statistical models challenge the validity of meta-analytic interpretations when the assumptions of the model are not adequately met.

Statistical Power

Many meta-analyses will have large enough sample sizes to generate sufficient statistical power for an overall aggregate effect size (although Hedges & Pigott, 2001, have shown that this is not necessarily the case). Sufficient power becomes more of an issue with smaller meta-analyses and when dealing with effect size distributions of subgroups. Statistical power will also depend on the choice of statistical model. Power increases as the cumulative sample size increases for a fixed-effects model and increases with the number of studies for a random-effects model. Low statistical power challenges the statistical conclusion validity of meta-analytic results.

Visual Displays

Visual displays are indispensable as tools for exploratory data analysis and can provide an extremely clear and powerful way to present research results (see Tufte, 1983; Tukey, 1977). As in primary research, graphs such as histograms, stem-and-leaf, and box plots are excellent ways to represent study results. Beyond effectively displaying the central tendency, variability, and normality of an effect size distribution, these plots are helpful in diagnosing problems such as extreme data points and non-normal distributions (Lipsey & Wilson, 2001). Other helpful graphs include funnel displays, and error-bar plots, the latter graph displaying confidence intervals around the parameters of interest (Lipsey & Wilson, 2001). Finally, a particularly interesting type of graph for meta-analysis is to display study results (usually effect sizes) blocked by important study characteristics (Light et al., 1994). For example, if study results are shown to be systematically related to the use of random assignment, a box plot could be constructed in which studies that used random assignment and those that did not are displayed side-by-side (Light et al., 1994). We were interested in the extent to which meta-analysts take advantage of the many visual displays available.

Discussion of Limitations

As in primary research, meta-analytic authors should discuss the limitations of the research methodology and conclusions drawn from the review (Cooper, 1989; Halvorsen, 1994; Light & Pillemer, 1984; Wolf, 1986). A thorough discussion of the limitations inherent in the methods or primary data is necessary to present a complete picture of the context in which the results should be interpreted. The most common limitation encountered in meta-analytic research is the availability of primary research studies, but other limitations concerning, for example, the generality or boundary conditions of any observed relations should also be discussed (Cooper, 1989).

Theoretical Basis

In many applications of meta-analysis it is the robustness of the overall main effect under study that is of interest, and indeed, this was the primary goal of many early meta-analyses (Cook et al., 1992). This over reliance on investigating only main effects has been one of the criticisms leveled against meta-analysis (see Cook & Leviton, 1980). Many recent meta-analyses go beyond the omnibus effect and "explore some of the method factors, some of the populations and settings, and some of the treatment variants that influence the size of the effect" (Cook et al., 1992, p.14). Although we view all applications of meta-analysis as equally useful and valid, we were interested in the extent to which meta-analysts are driven by theoretical constructions of the phenomenon under study and how often explicit a priori hypothesis are stated and tested within the meta-analytic context. In recent years, many authors have been advocating the use of meta-analysis for theory testing and have been developing the statistical theory and methodology for theory testing in the meta-analytic context (e.g., see Becker & Schram, 1994; Cook et al., 1992; Knight, Fabes, & Higgins, 1996; Miller, Lee, & Carlson, 1991; Miller & Pollock, 1994;Mullen, Salas, & Miller, 1991; Shadish, 1996; Viswesvaran & Ones, 1995). However, theory testing within the meta-analytic context will likely be challenging. Even with a theoretical model in hand, a sufficient number of primary reports needed to test the hypothesized relations may be difficult to find.

Method

Literature Retrieval

The PsycINFO database was used to search for a sample of published meta-analyses. The relatively broad spectrum of fields referenced in this database (e.g., psychology, economics, psychiatry/medicine, etc.) allows loose generalization of our results to the behavioral sciences as a whole. The following keywords were used for the search: *meta-analysis*, *meta analysis*, *research synthesis*,

research integration, *integrative review*. An initial 4,837 references were obtained, including actual meta-analyses (379 of which were dissertations), comments on previously published meta-analyses, methodological papers, and book chapters. The first author scanned the title and abstract from each report in this initial set and separated meta-analyses published in peer-reviewed journals. A meta-analysis was defined here as any study that used statistical techniques to combine multiple data sets. Using this definition a total of 2,792 meta-analytic studies remained in the study population. Further restricting the population to those studies published between 1994 and 2004 left a final population of 1,785 meta-analyses. A random subset of 100 meta-analyses were selected for coding. Three papers from the original random sample were replaced with a random selection when they could not be acquired through any lending libraries, and one paper was replaced because quantitative techniques were not used in the review.

Coding Procedures

The information extracted from the meta-analyses was categorized as being objectively coded or as requiring inference (high inference). Objective codes are those that rely on information explicitly available in each article—for example, sample size, the number of coders, and whether confidence intervals were reported (see Stock, 1982, for a discussion of the coding reliability of such variables). The first author extracted all objective codes from each study in the sample. The high inference variables are subjective in nature and were coded by at least two members of the research team. These variables include the following: the subdiscipline in which the meta-analysis was produced, whether the coding scheme was described adequately, the theoretical basis of the meta-analysis, and the discussion of limitations. We used modifications of coding schemes used by Steiner (1991).

To identify the extent to which meta-analyses in the sample described the coding instrument, ratings were made on a 3-point scale with "1" corresponding to no mention of the coding scheme, "2" representing a study that described some aspects of the coding scheme well but left others unclear, and "3" representing a study that made a substantial effort to detail the entire coding scheme, or it is stated that the coding sheet is readily available from the authors.

In our examination of theoretical basis, meta-analyses were coded a "1" if relevant theory was mentioned only in passing or not at all. This type of meta-analysis was primarily focused on establishing the existence of an effect (possibly with exploratory moderator analyses) in the absence of any theoretical development. Meta-analyses were coded a "2" if theory was explicitly discussed, but any further moderator analyses were exploratory and not designed to test any particular theoretical prediction or hypothesis beyond the omnibus effect(s). Those meta-analyses coded a "3" discussed relevant theory and used additional analyses to test specific a priori hypotheses derived from theory or previous primary research. Meta-analyses were coded a "4" if theory testing was the main focus of the review. These meta-analyses were defined as testing multiple hypothesized relationships within a theoretical construction, with the main goal of assessing the adequacy of the theory or overall model.

In assessing the discussion of limitations, meta-analyses were coded a "1" if there was no mention of limitations because of methods, primary data, or restricted generality of results. A study was coded a "2" if limitations were mentioned and/or discussed but were not directly related to interpretation of the results. For example, this would include general statements about the limitations inherent in all meta-analytic work with no further discussion of how these limitations related to the present results or restricted generalization. A study was coded a "3" if limitations were acknowledged and the conclusions or results were discussed in light of these limitations. The coding reliability for each high inference variable is reported in Table 2. Discrepancies between coders were first checked for clerical errors, then reliability measures were calculated and discussion ensued to reach consensus on the appropriate code.

The complete coding manual, the list of the meta-analyses included in this review and other supporting material has been archived for public download (Dieckmann, Malle, & Bodner, 2009).

Results and Discussion

Meta-analyses that do not fit conceptually or methodologically into particular categories are dropped from the corresponding analyses within that section. For example, if a meta-analyst did not search the literature for primary reports because only studies within a particular lab were being synthesized, the meta-analysis would not be included in the analysis of literature search procedures.

Sample Demographics

Size of the meta-analyses. The number of primary study results included in each meta-analysis was used as a rough measure of the size of the meta-analyses. Figure 1 shows the distribution of the number of study results included in each meta-analyses (M = 118.56; SD = 148.76; Minimum = 5; 25th = 29; 50th = 64; 75th = 140; Maximum = 781).

Subdiscipline. Table 3 shows the number of studies in each of the seven subdisciplines initially used to code the sample and an alternative classification with five categories based on a recent review of primary research (Bodner, 2006).

Using coding scheme B, 40% of the meta-analyses dealt with some aspect of physical or mental health. This is not surprising given the widespread use of meta-analytic methods in fields that focus on outcome research (e.g., Medicine, Psychiatry, Clinical Psychology; Lipsey & Wilson, 2001). The next largest groups are Social/Personality (23%) and Applied (22%), with Cognitive/Neuroscience (11%) and Developmental (4%) at the bottom end.

Meta-analytic work appears to be more common in some subdisciplines than others, but this could simply be a function of the

Table 2
Coding Reliability for High-Inference Variables

Characteristic	% agreement	Cohen's kappa
Sub-discipline	0.90	0.89
Theoretical basis	0.74	0.63
Explanation of coding scheme	0.86	0.75
Discussion of limitations	0.88	0.81

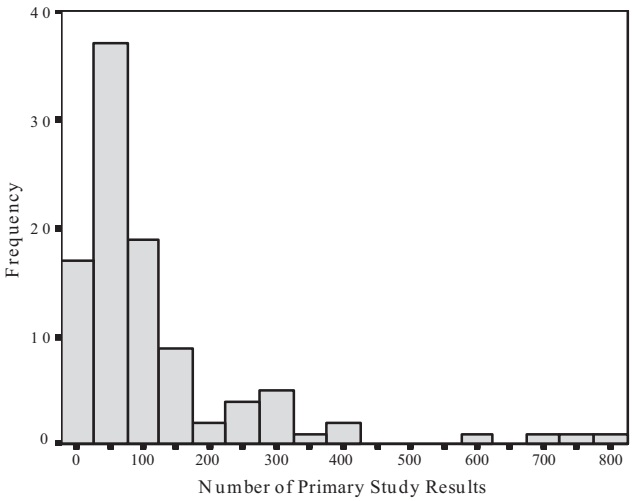

Figure 1. The distribution of primary study results included in the meta-analyses.

size of the research area. The number of meta-analyses produced within a given subdiscipline is likely related to the size of the field as a whole, and the size of a field can be roughly operationalized as the number of publications that are produced. A recent review of the characteristics and methodology of a random sample of primary studies will serve as a convenient baseline. Bodner (2006) extracted a random sample ($N = 200$) of primary research reports from *PsycINFO* for the year of 1999. Coding Scheme B was used to classify these primary reports into subdisciplines. Figure 2 shows the percentage of primary studies and meta-analyses by subdiscipline found in the two reviews. The percentage of meta-analyses per discipline deviated from what would be expected based on the percentage of primary studies, $\chi^2(4) = 14.99$, $p = .005$. The developmental and applied fields are remarkably close to what would be expected, Social/Personality and the Clinical/Health fields produced more meta-analyses, and Cognitive/Neuroscience produced about one third of what would be expected based on the percentage of primary studies.

Searching the Literature and Inclusion Criteria

Exhaustive search. Table 4 shows the percentage of meta-analyses that reported using the different retrieval methods. A large percentage of the studies reported using computer searches (88%), fewer report using the ancestry method (57%), and even fewer report contacting researchers to obtain unpublished reports (40%). Overall, only 27% of the meta-analyses sampled report using all three methods of retrieval. Year of publication was also moderately related to reporting computer searches, $r(92) = .25$, $p = .014$, 95% CI (0.05, 0.43), reporting contacting researchers to obtain unpublished reports, $r(92) = .25$, $p = .013$, 95% CI (0.05, 0.43), and reporting all three search methods, $r(92) = .22$, $p = .035$, 95% CI (0.02, 0.40). Although these effects are relatively small, it appears that more recent meta-analyses are reporting more thorough searches of the literature. Additionally, 26% of authors that reported trying to acquire unpublished studies by contacting researchers explicitly mentioned that they were unsuccessful in their elicitations. Some of this is likely because of the fact that there were no unpublished studies available from these researchers, but it is also questionable at times how much effort is made on the part of the primary researchers that are contacted by a meta-analyst.[3] As the creators of unpublished datasets and the consumers of meta-analytic reviews, we should do our best to help meta-analysts.

Inclusion criteria. The results from this review are encouraging with respect to reporting inclusion/exclusion criteria. We find that 95% of studies at least mention the criteria used for including studies. Meta-analytic authors seem to have made progress in addressing one of the important threats to the validity of meta-analytic conclusions.

Inclusion of unpublished work. Virtually all of the meta-analyses (98%; two studies did not report where their studies came from) in our sample included published studies, while only 53% included both published and some kind of unpublished data. There are several potential explanations for why only about half of the sample included unpublished data. Regardless of the cause, however, authors should use methods for identifying possible publication bias in their sample of primary reports.

Techniques to address publication bias. A funnel display is one of a number of graphical displays designed to facilitate reporting and interpretation of meta-analytic results (Light & Pillemer, 1984). This type of display was exceedingly rare in our sample, with only 5% of meta-analyses reporting them.

Another helpful technique for coping with publication bias is file-drawer analysis (Rosenthal, 1979). The basic procedure is to estimate the number of studies (with average effect sizes at the null value) that would need to be locked away in file drawers before it would challenge the significance of the combined effect size of the meta-analysis. Only 32% of the sample used file-drawer analysis, and only 4% used both funnel plots and file-drawer analysis.

Whether a meta-analysis included only published or both published and unpublished reports was not related to whether they employed either of the interpretive tools. Those studies that are probably most at risk of the detrimental effects of publication bias (those that only include published studies) were not any more likely to employ the tools.

Listing primary reports included in review. Also important is the explicit listing of the primary reports that were included in the review. It is often feasible to include a table or appendix that lists each report included, and in larger meta-analyses the author may note that a complete list of studies is available in an online archive or by request. We find that 92% of meta-analyses explicitly report each primary study in the sample, and this was moderately related to the page length of the journal article, $r(95) = .20$, $p = .048$, 95% CI (0.001, 0.38).

Summary. Overall, these results suggest that very few authors discuss or make efforts to estimate or correct for possible publication bias. This is worrisome given the powerful influence that these effects can have on meta-analytic results. As Begg (1994)

[3] One eminent meta-analytic researcher, Robert Rosenthal, expressed deep concern in a lecture for the inexplicably high rate of misplaced datasets and data-destroying floods, fires, and other natural disasters that plague academic institutions (APA – Advanced Training Institute, June 14-17, 2004).

Table 3
Sample Characteristics

Characteristic	Studies (*n*)
Year of publication	
1994	8
1995	14
1996	11
1997	7
1998	10
1999	5
2000	14
2001	15
2002	5
2003	10
2004 (through June)	1
Total	100
Subdiscipline (Categorization A)	
Clinical/Counseling	15
Social/Personality	23
Cognitive/Neuroscience	11
Developmental	4
IO/Economics	13
Health/Medicine	25
School/Education	9
Subdiscipline (Categorization B)[a]	
Clinical/Counseling/Health/Medicine	40
Social/Personality	23
Cognitive/Neuroscience	11
Developmental	4
Applied (IO/Education/Economics)	22

[a] This is an alternative subdiscipline coding scheme based on Bodner (2006). The first category under this scheme is the combination of Clinical/Counseling and Health/Medicine categories of System A, and the last category is the combination of IO/Economics and School/Education.

stated, "publication bias presents possibly the greatest methodological threat to validity of a meta-analysis" (p. 407). Of course, publication bias is an equally large threat in interpreting primary research findings, although the primary researcher has no methods for estimating the magnitude or effect of this bias. Authors should take advantage of meta-analysis as a systematic method for assessing and controlling publication bias.

Coding the Primary Reports

Multiple coders and reliability. Just as interviewers or judges of behavior in primary research are assessed to make sure that they have clear operational definitions and are measuring the same construct, coders interviewing primary reports in meta-analysis should be assessed as well (Cook et al., 1992). It is generally accepted that multiple coders should be employed and some measure of reliability should be reported, especially as more complex moderator coding schemes are used (e.g., Bullock & Svyantek, 1985; Cooper, 1989; Cooper & Hedges, 1994;Lipsey & Wilson, 2001; Rothstein & McDaniel, 1989; Wolf, 1986). Only 43% of meta-analyses in the sample that coded study characteristics from primary studies report the number of coders that were used. Within this set, the number of reported coders ranged from one ($n = 5$) to eight ($n = 1$) with 88% using more than one coder. Additionally, only 34% of the total sample reported a measure of coding reliability.

We acknowledge that there are some meta-analyses that have very simple moderator coding schemes and only a small sample of studies; in such cases a single coder may not necessitate concern. However, authors in such cases should still make explicit that one coder was used and explain why this was appropriate.

Moderator coding schemes. Along with reporting information about coders and coding reliability, reviews should make explicit the actual procedures used and operational definitions of coded study characteristics. This should be as rigorous as the reporting of study variables in primary research and is critical to the evaluation of any meta-analysis. We find that 62% of studies made a substantial effort to detail the entire coding scheme, 37% described some aspects of the coding scheme, and encouragingly only one study did not mention the coding scheme at all. Because the importance of a complete description of the coding scheme is fundamental to reporting a meta-analysis, our results that only 62% of studies adhere to this recommendation indicates that more careful reporting is needed on this dimension.

Missing codes. Another issue related to the coding of study reports is the generally acknowledged, and often frustrating, problem of insufficient reporting in the primary studies (e.g., see Pigott, 1994, for a detailed discussion of ways to deal with missing data; also see Orwin & Corday, 1985). Over half of the sample (53%) explicitly mentioned insufficient reporting being a problem. The number of reviewers that experienced such problems is probably even higher, but not all authors mentioned it in their reports. Not surprisingly, the consistent call for more complete reporting in primary studies is likely to continue to be heard.

Summary. Overall, there seems to be a relative lack of attention to conducting or at least reporting coding procedures and coding reliability. Meta-analysts should remember that the meta-analytic conclusions are only as good as the information that goes into the review.

Analysis and Reporting the Results

Representing the results of primary reports. In the following, we discuss our findings regarding the attention to statistical inde-

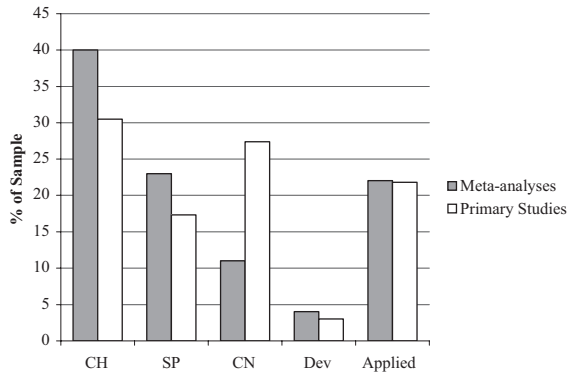

Figure 2. Meta-analyses ($n = 100$; present study), and primary studies ($n = 197$; Bodner, 2006) by subdiscipline.

Table 4
Percentage of Meta-Analyses That Follow Recommendations for Searching the literature, Study Inclusion and Coding

Characteristic	%	95% CI	Total *n*
Searching the literature[a]			
Reports using computer searches	88.3	(80.7, 93.6)	94
Reports using ancestry method	57.4	(47.4, 67.1)	94
Reports using invisible college (IC)	40.4	(30.9, 50.5)	94
If IC used, reports unsuccessful attempts	26.3	(14.4, 41.7)	38
Reports using all three search methods	26.6	(18.5, 36.1)	94
Inclusion criteria			
Described criteria for including studies	94.7	(88.7, 97.9)	94
Included unpublished studies	53.2	(43.1, 63.1)	94
Included a complete list of primary reports[b]	91.8	(85.0, 96.0)	97
Coding the primary reports[c]			
Reported the number of coders used	43.0	(33.3, 53.2)	93
If reported, those that used two or more	87.5	(74.8, 95.1)	40
Reported a measure of coding reliability	34.4	(25.4, 44.4)	93
Description of the moderator coding scheme[d]			
Scheme not mentioned at all	1.1	(0.1, 5.1)	89
Some aspects described	37.1	(27.6, 47.4)	89
Substantial effort was made	61.8	(51.5, 71.4)	89
Insufficient reporting in primary studies[a]	53.2	(43.1, 63.1)	94

Note. Percentages are presented with the total number of meta-analyses within each category. Confidence intervals are equal tailed Jeffrey's prior intervals.
[a] Percentages are calculated out of those meta-analyses that searched the literature for primary reports. [b] Percentages are calculated out of those meta-analyses that included primary reports. [c] Percentages are based on those meta-analyses that extracted information from the primary articles (e.g., as opposed to using raw data). [d] Percentages are based on those meta-analyses that used coding schemes for moderator/mediator variables.

pendence of study results, correction of study artifacts, and presentation of effect size distributions.

About two thirds of the sample (66%) made mention of whether the study results were independent and explained how they treated the dependencies if they existed. We also assessed how many authors corrected effect sizes for study artifacts, following the procedures outlined by Hunter and Schmidt (1990). Only five meta-analyses corrected for systematic study artifacts, and all five came from the applied subdisciplines of Industrial/Organizational ($n = 4$) and School/Education ($n = 1$). It appears that these methods have not been widely used outside of the applied Industrial/Organizational setting.

It is also frequently recommended that meta-analysts display the study results (normally effect sizes) from each primary report used in the analysis (e.g., Bullock & Svyantek, 1985; Halvorsen, 1994; Jackson, 1980). For a small number of primary reports this can be done in a single table, and for larger samples, the author should note that these data are available upon request. The reasons for reporting or making this information available are to fully disclose the meta-analytic methods and facilitate direct replication by other researchers. A little over half of the meta-analyses (54%) directly reported or made this information available.

Combining study results. We find that 71% of the sample weighted individual study results by how well they were expected to estimate the population parameters of interest, typically by weighting each study by the within-study sample size or the inverse of the variance of the effect size estimator. Those analysts that did not weight by precision did not provide an argument for why they did not. Another scheme that has been discussed by some authors (see Shadish & Haddock, 1994) is weighting by the quality of the research methods that generated the result. Two reports weighted by study quality, although 12 reports investigated study quality (assessed by a rating scale) as a moderator variable later in the analysis.

Recent reviews of the use of confidence intervals in primary research find that CIs were generally used more frequently in the health fields than in the psychological disciplines (see Fidler, Thomason, Cumming, Finch, & Leeman, 2004). Similar trends may be present in meta-analytic research. Overall, of those studies that reported a measure of central tendency, 56% also reported confidence intervals. Confidence intervals were more often used in the Health/Medicine subfields (71.4%) compared to Industrial/Organizational (53.8%), School/Education (50%), Clinical/Counseling (46.7%), Social/Personality (40.9%), and Cognitive/Neuroscience (33.3%), although this difference was not statistically reliable, $\chi^2(7, N = 91) = 8.48$, $p = .21$, $\phi_v = .31$. Although a definitive conclusion cannot be made based on these data, we conjecture that psychological subdisciplines are lagging behind the medical sciences in the use of confidence intervals in meta-analytic work, just as they are in primary research. As a hint that things may be improving, more recently published meta-analyses were more likely to report confidence intervals, $r(89) = .26$, $p = .01$, 95% CI (0.06, 0.44).

Assessing and accounting for variability in study results. In addition to deriving an overall omnibus effect, assessing and accounting for variability in the distribution of study results is often a main focus of a research synthesis. Of those meta-analyses that reported a measure of central tendency, only 31% reported a basic measure of variability. Restricting the sample to those that

reported a measure of central tendency, 60% reported either a statistical test or a partition of variance to assess the homogeneity of the effect distributions.

We find that 84% of our sample conducted moderator analysis of some kind, and those that did not either had small restrictive sample sizes or moderator analysis did not fit with the main purpose of the meta-analysis. Overall, moderator analysis was relatively common among meta-analyses in our sample, giving weight to the notion that meta-analytic work is not only concerned with establishing main effects.

While moderator analysis has become common for meta-analyses, some authors have pushed for greater sophistication in using meta-analysis for explanation (see Cook et al., 1992; Shadish, 1996). Among other strategies that are outlined in great detail elsewhere (Cook et al., 1992), many authors have discussed the possibilities of investigating causal mediating processes. This type of analysis was found to be extremely rare in our sample with only one study statistically exploring the effect of a mediating variable.

Statistical model. Regardless of the statistical model that is eventually used for analysis, the author should be explicit about the model specification chosen and discuss the appropriate generalization from that particular model (i.e., fixed, random, or mixed effects model). Only 23% of the sample made the choice of analytic model explicit, and only a minority of these explicitly discussed the appropriate level of generalization (see Table 5). Additionally, in the case that it is difficult to choose between a fixed, random or mixed analytic approach (as it often is), it is recommended that the analyst run multiple models and compare the results. We find that only 9% of the sample explicitly analyzed the meta-analytic dataset with multiple models.

Power analysis. It is often recommended that power analyses be conducted and reported (Hedges & Pigott, 2001; Rosenthal, 1995; see Cohen, 1988). Only one study, from the health and medicine subdiscipline, reported retrospective power analyses on the results of the meta-analysis. This result seems to mirror the relative lack of attention paid to power and Type II errors in primary research.

Visual display. Table 6 shows the number of meta-analyses in our sample by subdiscipline that reported each of these different graphical displays, as well as the number of meta-analyses that presented any graphical displays of results (last column). Overall, relatively few meta-analyses used graphical methods, with only 31% of the sample reporting at least one graph. Not a single review reported a box plot, and error-bar plots were used primarily by studies in the Health/Medical subdisciplines. Consequently, meta-analytic results are much more likely to be reported in tables, averaging almost four tabular displays per meta-analysis ($M = 3.91$, 95% CI (3.38, 4.44); $Mdn = 3$, $SD = 2.69$).

One possible reason for the infrequent use of graphs is that meta-analyses are already relatively long articles, so authors are discouraged to add more space by means of graphical displays. Additionally, some information, such as a list of primary studies and their characteristics, or a list of moderators and their predictive power, is difficult to display in any way but a table. However, some information commonly displayed in a table can be displayed as a graph with better pedagogical effect. For example, instead of (or in addition to) displaying each study and its corresponding effect size in a table, display this information as a histogram to allow the reader to view the shape, central tendency and spread of the distribution. As many others have done, we urge meta-analytic authors to explore the broad range

Table 5
Percentage of Meta-Analyses That Follow Recommendations for Analysis and Reporting the Results

Characteristic	%	95% CI	Total *n*
Representing study results			
Reported whether the study results were independent	66.0	(56.4, 74.7)	100
Corrected for statistical artifacts[a]	5.7	(2.2, 12.0)	88
Included a complete table of effects	54.0	(44.2, 63.5)	100
Combining study results and central tendency			
Weighted by sample size or variance[b]	66.3	(56.6, 75.1)	98
Weighted by study quality	2.0	(0.4, 6.4)	98
Reported confidence intervals[c]	52.7	(42.5, 62.8)	91
Variability of effect distribution[c]			
Reported variance or *SD*	30.8	(22.0, 40.7)	91
Reported homogeneity test statistic	59.3	(49.1, 69.0)	91
Accounting for variability			
Conducted moderator analysis	84.0	(75.9, 90.2)	100
Conducted mediator analysis	1.0	(0.1, 4.6)	100
Fixed, random, and mixed effect models			
Explicitly discussed the type of model used	23.0	(15.6, 31.9)	100
Explicitly mentions comparing multiple analytic models	9.0	(4.6, 15.8)	100

Note. Percentages are presented with the total number of meta-analyses within each category. Confidence intervals are equal tailed Jeffrey's prior intervals.
[a] Because the methods used for adjusting for statistical artifacts were developed for work with effect sizes, percentages are calculated out of all those meta-analyses using effect size based analysis. [b] Weighting schemes based on sample size or variance were not applicable to the two meta-analyses that integrated single-subject research reports. [c] CIs and reporting variance and measures of homogeneity are calculated out of those studies that at least reported a measure of central tendency.

Table 6
Types of Graphical Displays Reported by Subdiscipline

	Type of display						
Subdiscipline	Meta-analyses (*n*)	Histogram or stem-and-leaf	Box plots	Funnel plots	Error-bar plot	ES × Study characteristic (beside funnel)	Any graph
Clinical, counseling	15	1		1	1	2	4
Social personality	23	3		1		2	5
Cognitive neuroscience	11	1				2	3
Developmental	4	1				2	3
IO/Econ	13	2		1		1	3
Health/Medicine	25	1		2	5	4	9
School/Education	9	3				2	4
Total	100	12	0	5	6	15	31

of useful graphical displays available. Simple and powerful graphics can often greatly aid in communicating the important results from a meta-analytic review.

Discussion and Interpretive Tools

As can be seen in Table 7, 37% of the sample did not discuss any limitations, 35% briefly discussed limitations, and only 28% discussed limitations and related them to the conclusions of the meta-analysis. The discussion of limitations has clearly not become standard practice in meta-analytic reviews, as over one third of the sample did not make any mention of them. We do, however, find a moderate positive relationship between year of publication and the extent of the discussion of limitations, $r(98) = .21$, $p = .04$, 95% CI (0.02, 0.39), indicating that this situation may be improving with time.

Interpretive tools. We also investigated the number of meta-analyses that used any of a number of tools that are useful for aiding in the interpretation of meta-analytic results (Rosenthal, 1995)—namely, binomial effect size displays (BESD), coefficient of robustness, and file drawer analysis. Most of these tools, except file drawer analysis (discussed above), appear to be used very infrequently. Rosenthal and Rubin (1982) introduced the binomial effect size display (BESD) as a method of showing the practical importance of an effect size. Of those meta-analyses that used effect sizes to represent study results, only 6% used a BESD as an aid to interpretation. Similarly, only 1% of meta-analyses reported a coefficient of robustness, an alternate effect size metric that can be used to assess the robustness of a domain (or compare domains) after adjusting for the number of studies in a meta-analysis (Rosenthal, 1995).

Theoretical Basis of the Meta-Analyses

Three percent of the sample focused on assessing the adequacy of a model or theory by testing multiple hypothesized relationships (see Table 7). The remainder of the sample was relatively equally dispersed across the remaining three categories, with 27% men-

Table 7
Percentage of Meta-Analyses That Discuss Limitations, Use Interpretive Tools and Are Theoretically Motivated

Characteristic	%	95% CI	Total *n*
Discussion of limitations			
No discussion of limitations	37	(28.0, 46.7)	100
Briefly discussed limitations	35	(26.2, 44.7)	100
Fully discussed limitations	28	(19.9, 37.3)	100
Interpretive tools reported[a]			
Binomial effect size displays (BESD)	5.7	(2.2, 12.0)	88
Coefficient of robustness	1.1	(0.1, 5.2)	88
Retrospective power analysis	1.1	(0.1, 5.2)	88
File drawer analysis	31.8	(22.8, 42.0)	88
Theoretical basis of the meta-analysis			
Theory mentioned in passing or not at all	27	(19.0, 36.3)	100
Theory discussed but analysis exploratory	36	(27.1, 45.7)	100
Theory and specific hypotheses	34	(25.3, 43.6)	100
Theory testing is the main focus	3	(0.9, 7.8)	100

Note. Percentages are presented with the total number of meta-analyses within each category. Confidence intervals are equal tailed Jeffrey's prior intervals.
[a] Because many of the interpretive tools in this section were developed within the context of meta-analytic integration with effect sizes or significance levels, we calculated these percentages out of the total number of meta-analyses with effect size based analyses.

tioning relevant theory in passing or not at all, 36% mentioning relevant theory but any moderator analysis purely was exploratory, and 34% discussing relevant theory and using additional analyses to test specific a priori hypotheses. Even though 84% of the sample conducted some moderator analysis, only 37% tested specific a priori hypotheses derived from theory or previous research.

Additionally, there were some notable differences in theoretical basis between the subdisciplines. There was a relatively high percentage of meta-analyses that mentioned theory in passing or not at all in the Health/Medicine (52%) and School/Education (44.4%) subdisciplines, although the latter had a small sample size ($n = 9$). In Social/Personality this percentage was only 4.3%. Figure 3 shows the mean ratings on theoretical basis with associated 95% CIs by subdiscipline. We find statistically reliable variation across the subdisciplines, $F(6, 93) = 4.28$, $p = .001$, with Social/Personality ($p = .006$) and Industrial/Organizational ($p = .005$) rated as higher than the Health/Medicine subdiscipline.

General Discussion

The typical meta-analysis in our sample integrated 60 to 70 primary study results, although there was a substantial range with the smallest meta-analysis including five study results and the largest including 781. We also found that there were more meta-analyses in the Social/Personality and Health/Clinical and fewer in the Cognitive/Neuroscience subdiscipline. It is unclear why this is the case but suggests that the potential benefits of using meta-analytic techniques are not being realized in some subdisciplines.

The primary goal of most meta-analyses in our sample was to demonstrate the existence of an empirical relation and conduct exploratory moderator analyses. We find that these types of meta-analyses are more common in subdisciplines with a focus on outcome research like medicine, education, and clinical psychology, which have synthesis needs that are particularly well suited to straightforward meta-analytic techniques. This relative lack of a priori hypotheses and theory testing could be related to several factors. One may be the relative poor state of theory in the behavioral sciences as a whole (Meehl, 1978), which carries over to any theories that could be tested in meta-analytic work. Another factor is that meta-analysts are completely dependent on the available data in primary research, and often relations between theoretically interesting variables are simply not available in a sufficient number of primary reports. Several meta-analysts in our sample expressed an interest in testing mediating relationships or multipath models but did not have the primary-level data to test such relationships.

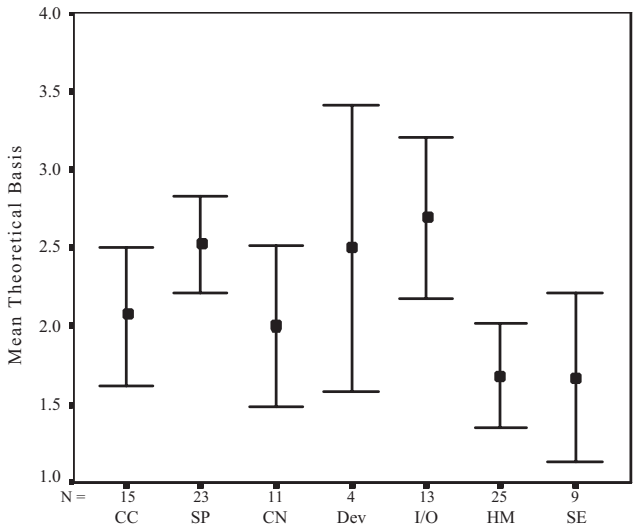

Figure 3. Average theoretical basis of meta-analyses by sub-discipline. Mean values are displayed with corresponding 95% CIs. CC = Clinical/Counseling; SP = Social Personality; CN = Cognitive/Neuroscience; Dev = Developmental; IO = Industrial/Organizational; HM = Health/Medicine; *SE* = School/Education.

We have also described selected aspects of common meta-analytic practice and assessed how consistent and complete authors are in reporting the methods and results of meta-analyses. Unlike previous explorations of meta-analytic practice, we examined a broad sample of meta-analyses with the aim of generalizing to the "typical" meta-analysis conducted in the behavioral sciences. Even though some aspects of practice and reporting were consistently aligned with expert recommendations, these results illustrate deficient reporting and a lack of focus on critical issues at almost every stage of the meta-analytic process.

Figure 4 shows the percentage of meta-analyses that adhered to the common guidelines of practice. This is a snapshot of the implicit weight that has been put on these critical aspects of practice in recent meta-analyses. As the gatekeepers of scientific publication, journal editors and reviewers could make a substantial difference in practice by focusing on these critical aspects, particularly the ones with low adherence. Instructors and meta-analytic methodologists should also be aware that these aspects of practice will likely need further attention.

It is also generally preferable for authors to assess the sensitivity of their meta-analytic conclusions. Sensitivity analysis is a systematic approach to assessing how the meta-analytic conclusions would change if different procedures or analytic techniques were used (see Greenhouse & Iyengar, 1994). It is this general sensitivity approach to analysis that is inherent in some of the recommendations discussed above. For example, conducting exploratory data analysis (e.g., through graphical displays) to identify outliers or distributional features that could affect the results, assessing how additional unpublished studies could affect results and reporting results under alternative analytic assumptions are all examples of attempts to assess the sensitivity of the meta-analytic conclusions.

Figure 5 shows the percentage of meta-analyses that adhered to the common guidelines of reporting. Again, this is a snapshot of the implicit weight that has been put on these aspects of reporting. We do recognize that in some cases fully reporting the procedures and results of a meta-analysis would take up more journal space than is available for the report. In these cases, we suggest putting the additional material in an online database, which can then be assessed by interested readers. Additionally, making all of the materials (e.g., codebooks, coding sheets, etc.) used in the meta-analysis readily available allows other researchers to fully scrutinize the meta-analytic methods.

Our results can be used to focus attention on several aspects of meta-analytic practice and reporting that need improvement. However, it is also important that future meta-analysts have examples

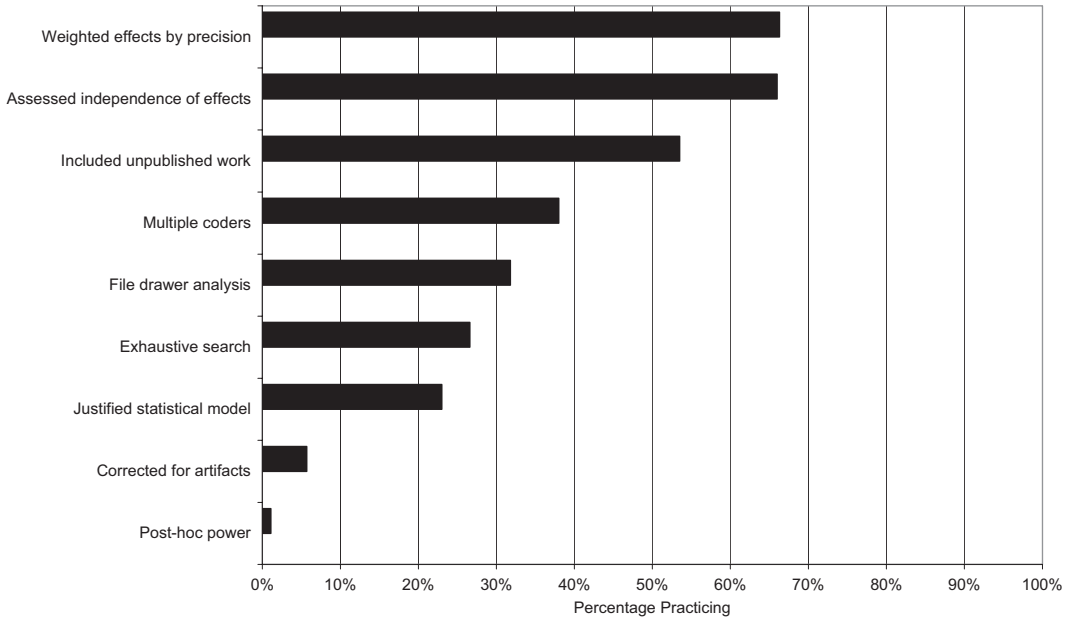

Figure 4. Adherence to common recommendations for meta-analytic practice. Exhaustive Search = those that report using all three methods of literature search (i.e., Computer searchers, ancestry method, and invisible college). Multiple coders = those that reported using two or more coders.

of good practice and reporting to correct these issues. In this paper, we have briefly reviewed a selection of common recommendations and there are several excellent guides to meta-analytic practice and reporting available (Cook & Leviton, 1980; Cooper, 1989; Cooper & Hedges, 1994; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Light & Pillemer, 1984; Lipsey & Wilson, 2001; Rosenthal, 1991, 1995; Wolf, 1986). There were also several excellent examples of meta-analytic practice and reporting in our random sample. Two meta-analyses that could serve as a model for future meta-analysts are by Postmes and Spears (1998) and Tenenbaum and Leaper (2002).

One potential limitation of this review is that the broad selection of meta-analyses sampled may have presented a poorer picture of meta-analytic practice by mixing reviews from journals differing in publication quality. We explored this possibility with journal impact factor (JIF) as a measure of journal quality and found no statistically reliable relationships between JIF and any of the reporting and practice variables reported here.[4] However, when calculating a percentage adherence score for each meta-analysis, we do find a moderate relationship with impact factor, $r(86) = 0.300$, $p = .004$.[5] This suggests that journals with higher impact factors may require stricter adherence to these guidelines for practice and reporting.

Conclusions

It is clear that meta-analytic practice can be improved by more consistent, standardized procedures and reporting practices. Paying particular attention to validity threats and testing the sensitivity of meta-analytic conclusions will substantially strengthen the conclusions of a review. Again, we are not trying to imply that all meta-analyses should use identical methods and should all look the same. There are certainly situations where it is completely justifiable to deviate from common practice recommendations (e.g., not weighting by precision when sample size is confounded with other study characteristics). However, the justification for not heeding common prescriptive advice should be made clear to the reader. This magnifies the importance of the explicit reporting of methods, particularly the subjective judgments that are part of every meta-analysis. Ironically, meta-analysts often struggle with incomplete or inconsistent research reporting and consistently call for standardized reporting in primary research, but as this review shows, meta-analysts are themselves not entirely consistent in reporting their methods and results.

Methods for research synthesis have clearly been helpful for the organization and systematization of ever expanding research literatures. It would be a mistake, however, to think that this has been the biggest contribution of meta-analysis. As Schmidt (1992) notes, meta-analysis is not just another method for reviewing research literatures but has the potential to fundamentally change the way scientists think about individual research results by focusing attention on the "critical role of sampling error, measurement error, and other artifacts in determining the observed findings and statistical power of individual studies" (p. 1179).

[4] This analysis was done on a subset of 88 meta-analyses for which journal impact factor was readily available. Impact factors were taken from Journal Citation reports – Social Sciences & General Science Editions (2004).

[5] The percentage adherence score was calculated as the percentage of reporting and practice recommendations (see Table 1) that each meta-analysis followed.

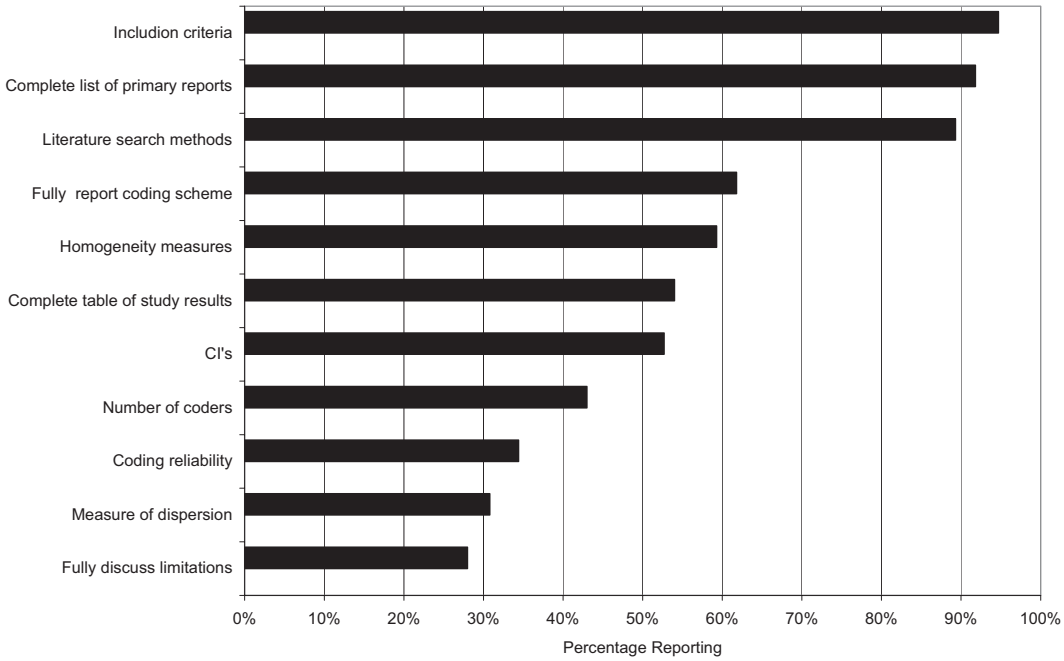

Figure 5. Adherence to common recommendations for meta-analytic reporting. Fully discuss limitations = those that were coded as fully discussing limitations ("3"). Fully report coding scheme = those that were coded as supplying a full description of the coding scheme ("3"). Literature search methods = those that reported any information about search methods.

Meta-analysis also has the potential to contribute to the development of theory and to address questions concerning causation. These potential benefits are not fully realized in the current sample of meta-analyses. All meta-analyses are completely dependent on the primary studies available for synthesis, but asking questions about causation and testing complex theoretical relations require even more from the meta-analytic sample. No amount of procedural excellence or thorough reporting can make up for a dearth of information in the literature.

However, any benefits that are to be had with meta-analysis must begin with a firm methodological foundation and attention to the complete reporting of the procedures and results. In the future, we look forward to advances in meta-analytic methodology and the increased availability of data from primary research reports to further realize the potential of research synthesis. Ideally, the day will come when all contributing authors will be required to upload a complete dataset accompanying a research report, allowing future meta-analysts to directly synthesize research literatures, thereby bypassing many of the methodological difficulties inherent in current meta-analytic practice (Glass, 2000).

References

American Psychological Association. (2001). *Publication manual of the American psychological association* (5th ed). Washington, DC: American Psychological Association.

Beaman, A. L. (1991). An empirical comparison of meta-analytic and traditional reviews. *Personality & Social Psychology Bulletin, 17,* 252–257.

Becker, B. J. (1991). The quality and credibility of research reviews: What Eds. say. *Personality and Social Psychology Bulletin, 17,* 267–272.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357–381). New York: Russell Sage Foundation.

Begg, C. B. (1994). Publication Bias. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399–409). New York: Russell Sage Foundation.

Bodner, T. E. (2006). Designs, participants, and measurement methods in psychological research. *Canadian Psychology, 47,* 263–272.

Bullock, R. J., & Svyantek, D. J. (1985). Analyzing meta-analysis: Potential problems, and unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology, 70,* 108–115.

Chalmers, T. C., Levin, H., Sacks, H. S., Reitman, D., Berrier, J., & Nagalingam, R. (1987). Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large cooperative trials. *Statistics in Medicine, 6,* 315–325.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (Rev. Ed.). New York: Academic Press.*

Cook, T. D., Cooper, H. M., Corday, D. S., Hartmann, H., Hedges, L. V., Light, R. J., et al. (1992). *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality, 48,* 449–472.

Cooper, H. M. (1989). *Integrating research: A guide for literature reviews* (2nd ed.). Newbury Park, CA: Sage.

Cooper, H. M., & Hedges, L. V. (1994a). Research synthesis as a scientific enterprise. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 3–14). New York: Russell Sage Foundation.

Cooper, H. M., & Hedges, L. V. (1994b). Potentials and limitations of research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 512–529). New York: Russell Sage Foundation.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Dieckmann, N. F., Malle, B. F., & Bodner, T. E. (2009). An empirical review of meta-analytic practice [Online supplement]. Available at www.decisionresearch.org

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Eds. can lead researchers to confidence intervals, but can't make them think. *Psychological Science, 15,* 119–126.

Glass, G. V. (2000). *Meta-analysis at 25.* Retrieved November 15th, 2003, from *http://glass.Ed.asu.edu/gene/papers/meta25.html*

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York: Russell Sage Foundation.

Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Halvorsen, K. T. (1994). The reporting format. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 425–437). New York: Russell Sage Foundation.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499.

Hedges, L. V. (1994b). Fixed effects models. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203–217.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law, 2,* 324–347.

Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 438–460.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80,* 94–106.

Knight, G. P., Fabes, R. A., & Higgins, D. A. (1996). Concerns about drawing causal inferences from meta-analyses: An example in the study of gender differences in aggression. *Psychological Bulletin, 119,* 410–421.

Kraemer, H. C., Yesavage, J. A., Gardner, C., & Brooks, J. O. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439–453). New York: Russell Sage Foundation.

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

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Mazela, A., & Malin, M. (1977). A *bibliometric study of review literature.* Philadelphia: Institute for Scientific Information.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Miller, N., Lee, J., & Carlson, M. (1991). The validity of inferential judgments when used in theory-testing meta-analysis. *Personality and Social Psychology Bulletin, 17,* 335–343.

Miller, N., & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 457–483). New York: Russell Sage Foundation.

Mullen, B., Salas, E., & Miller, N. (1991). Using meta-analysis to test theoretical hypotheses in social psychology. *Personality and Social Psychology Bulletin, 17,* 258–264.

Orwin, R. G., & Corday, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin, 97,* 134–147.

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 163–175). New York: Russell Sage Foundation.

Postmes, T., & Spears, R. (1998). Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin, 123,* 238–259.

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

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

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

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Rosenthal, R., & Dimatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

Rothstein, H. R., & McDaniel, M. A. (1989). Guidelines for conducting and reporting meta-analyses. *Psychological Reports, 65,* 759–770.

Schmidt, F. L. (1992). What do data really mean: Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods, 1,* 47–65.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.) *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Smith, N. L., & Caulley, D. N. (1981). The evaluation of educational journals through the study of citations. *Educational Researcher, 10, 11-, 12,* 22–24.

Steiner, D. D., Lane, I. M., Dobbins, G. H., Schnur, A., & McConnell, S. (1991). A review of meta-analyses in organizational behavior and human resource management: And empirical assessment. *Educational and Psychological Measurement, 51,* 609–626.

Stock, W., Okun, M., Haring, M., Miller, W., Kinney, C., & Ceurvorst, R. (1982). Rigor and data synthesis: A case study of reliability in meta-analysis. *Educational Researcher, 11,* 10–14.

Tenenbaum, H. R., & Leaper, C. (2002). Are parents' gender schemas related to their children's gender-related cognitions? A meta-analysis. *Developmental Psychology, 38,* 615–630.

Tufte, E. R. (1983). *The visual display of quantitative information.* Chesire, CT: Graphics Press.

Tukey, J. (1977). *Exploratory data analysis.* Reading, MA: Addison Wesley Publishing Company.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48,* 865–885.

Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74,* 259–264.

White, H. D. (1994). Scientific communication and literature retrieval. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 41–55). New York: Russell Sage Foundation.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage.

Received December 4, 2007
Revision received August 2, 2008
Accepted November 12, 2008 ■