Psychological Bulletin
1995, Vol. 118, No. 2, 183–192

0033-2909/95/$3.00

Writing Meta-Analytic Reviews

Robert Rosenthal
Harvard University

This article describes what should typically be included in the introduction, method, results, and discussion sections of a meta-analytic review. Method sections include information on literature searches, criteria for inclusion of studies, and a listing of the characteristics recorded for each study. Results sections include information describing the distribution of obtained effect sizes, central tendencies, variability, tests of significance, confidence intervals, tests for heterogeneity, and contrasts (univariate or multivariate). The interpretation of meta-analytic results is often facilitated by the inclusion of the binomial effect size display procedure, the coefficient of robustness, file drawer analysis, and, where overall results are not significant, the counternull value of the obtained effect size and power analysis.

The purpose of this article is to provide some guidelines for the preparation of meta-analytic reviews of literature. *Meta-analytic reviews* are quantitative summaries of research domains that describe the typical strength of the effect or phenomenon, its variability, its statistical significance, and the nature of the moderator variables from which one can predict the relative strength of the effect or phenomenon (Cooper, 1989; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Light & Pillemer, 1984; R. Rosenthal, 1991).

The goal is not to explain the various quantitative procedures used in meta-analytic practice, for these are described in detail in the textbooks by the authors just cited, in less detail in R. Rosenthal (1993), and in far greater detail in a new handbook edited by Cooper and Hedges (1994). Another goal the writer does not have is to convince readers of the value of meta-analytic research summaries because this too has been addressed in all the previously referenced texts and in many other sources. The heart of this article is a discussion of what should be considered for inclusion in a meta-analytic report. Not all of the suggestions apply equally well to all meta-analytic undertakings, but on average important omissions are likely to be minimized if these suggestions are at least seriously considered.

Who should be thinking of writing meta-analytic reviews? Anyone considering a review of literature, or a specifiable subset of the literature, may as well do it quantitatively as nonquantitatively because all of the virtues of narrative reviews can be preserved in a meta-analysis that merely adds the quantitative features as a bonus. The level of quantitative skill and training required to use basic meta-analytic procedures is so modest that researchers capable of analyzing the results of their own research will be capable of readily learning the small number of calculations required to answer standard meta-analytic questions (e.g., What is the mean and standard deviation of this list of correlation coefficients or other effect size estimates?).

As is the case of data analysis of any study, the analysis of a set of studies can vary greatly in complexity. For example, the texts of the six authors previously listed can be roughly divided into two levels of complexity and completeness. The books by Glass et al. (1981), Hedges and Olkin (1985), and Hunter and Schmidt (1990) are more detailed and more quantitatively demanding than those by Cooper (1989), Light and Pillemer (1984), and R. Rosenthal (1991). There are theoretical differences among these six texts as well, and this article is intended to be useful to meta-analysts working within any of these frameworks. Thus, although some of the more complex procedures described by Hedges and Olkin and by Hunter and Schmidt are not specifically mentioned, researchers working within their frameworks can easily add those analyses to the basics covered in this article. Regardless of how complex the meta-analytic procedures may become in a given review of the literature, reporting the basics makes a meta-analysis easier for the reader to follow and to understand at a deeper level. Reporting the basics also makes it easier for a reader to check the tenability of conclusions drawn by the meta-analyst.

Thus, keeping at least the basic meta-analytic procedures used descriptive, simple, and clear is a positive virtue. In 20 years of reviewing meta-analytic literature syntheses, I have never seen a meta-analysis that was "too simple," but I have often seen meta-analyses that were very fancy and very much in error.

The most important part of a meta-analysis is the descriptive part in which the effect sizes (e.g., correlation coefficients) are displayed and their distribution and central tendency are summarized. Good meta-analytic practice, similar to good data-analytic practice in general, adopts an exploratory orientation toward these displays and summaries (Tukey, 1977); for this valuable enterprise, little "high-tech statistication" is required. Indeed, the computations required for the most basic meta-analytic work are so trivial that in my own meta-analytic work of the last 30 years or so, I have never felt the need to use a software package that "does meta-analysis."

Good software for meta-analytic procedures can, of course, be a great time saver. However, a drawback to the development

Preparation of this article was supported in part by the Spencer Foundation and by a sabbatical award from the James McKeen Cattell Fund and the Faculty of Arts and Sciences of Harvard University.

Correspondence concerning this article should be addressed to Robert Rosenthal, Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, Massachusetts 02138.

of sophisticated software for the computation of meta-analytic (or any other data-analytic) computations is that some researchers who feel less expert than they might like believe the software will "do the analysis." Alas, that is not the case. The software does a variety of computations and it does them fast, but for any given application the computations may be wise or they may be foolish. Staying simple, staying close to the data, and emphasizing description help to avoid most serious errors. It is better to consult with a more experienced colleague who knows exactly what is being computed by the software than to trust the software to do the analysis. That advice applies to all data-analytic undertakings, of course, not merely to meta-analytic procedures.

Without any implication that all good meta-analyses look alike and incorporate all the suggestions to follow, for the remainder of this article I discuss what might be reported in most meta-analyses and what should probably be at least considered for almost all meta-analyses.

Introduction to a Meta-Analytic Review

The introduction to a meta-analysis is not very different strategically from the introduction to any scientific article. It tells readers why they should read the article, what makes it important, and how it achieves what has not been achieved before.

If the literature is made up of several types of study, it is helpful to describe a typical study from each of the types. If the results of the research differ widely—for example, some results strongly favor the treatment condition and some results strongly favor the control condition—it is useful to give examples of studies showing this wide variation in results.

Method Section of a Meta-Analytic Review

Literature Searches

In this section, the meta-analyst should tell readers how the studies summarized were located, what databases were searched, what journals were painstakingly gone through, what research registers were consulted, and what steps were taken to retrieve the "fugitive literature." For those meta-analysts not trained as information scientists, the new *Handbook of Research Synthesis* edited by Harris Cooper and Larry Hedges (1994) may offer considerable help and enlightenment. Most of what any meta-analyst needs to know (and even more) about retrieving the data for a meta-analysis is contained in about 50 pages of the four chapters prepared by White (1994), Reed and Baxter (1994), Dickersin (1994), and M. C. Rosenthal (1994).

The reason for trying to locate all the research on the topic of a meta-analysis is primarily to avoid the biased retrieval of searching only the major journals, which may selectively publish only the results characterized by lower p values and larger effect sizes. If the domain searched has a great many studies, more than the meta-analyst has the resources to analyze, it is better to sample the exhaustive listing of results than to select only the more readily retrievable results.

Criteria for Inclusion

Information available. Not all the reports retrieved are appropriate for inclusion in a meta-analysis. Some turn out to have no data of any kind, some have collected data but report on the data so poorly that they are unusable. Some are borderline cases where the meta-analyst is given enough data that good detective work allows him or her to obtain at least an approximate effect size estimate and significance level. Many studies, for example, simply say "there was no effect of X on Y" or "the effect was not significant." Meta-analysis involves the summarization of data, not of an author's conclusions, so the previous statements are of little help to the meta-analyst. However, if the meta-analyst has the relevant means and standard deviations, he or she can compute the effect sizes. If, in addition, sample sizes are given, the meta-analyst can also compute accurate p values.

For studies claiming "no effects" or "no significant effect," the meta-analyst may want to assign an effect size estimate of 0.00 and a one-tailed p of .50 ($Z = 0.00$). Experience suggests that this procedure is conservative and leads to effect size estimates that are too small. The alternative of not using those studies, however, is likely to lead to effect size estimates that are too large and almost surely to p values that are too small, that is, too significant. Confronted with this choice of procedures, it is usually best to "do it both ways" to learn just how much difference it really makes to the overall view of the data. Considerations of alternative approaches to the data are part of the process of "sensitivity analysis" described by Greenhouse and Iyengar (1994).

Study quality. Of the studies retrieved, some may be methodologically exemplary and others may be stunningly bad. Should the meta-analyst include them all or only the good ones? The question of quality criteria for inclusion is really a question of weighting by quality (R. Rosenthal, 1991). Including good studies and excluding bad ones is simply a 1,0 weighting system which is often suspect on grounds of weighter bias. The meta-analyst is too likely to think of his or her own studies, those of his or her students, those of friends, and those of others who successfully replicate his or her work as good studies. In addition, the meta-analyst is too likely to think of the studies of his or her enemies and of those who fail to replicate his or her work as bad studies. As protection against biases, a meta-analyst would do better to evaluate the retrieved studies for quality by some procedure that allows disinterested coders or raters to make the required judgments. Indeed, some workers feel that coders or raters should be blind to the results of the study.

Coding of studies for their quality usually requires only simple judgments of the presence or absence of desirable design features, such as randomized experiment, experimenter blind to hypothesis, or controlled demand characteristics. Quality points can then be assigned on the basis of the number of desirable features present. Rating of studies usually requires a more global, overall assessment of the methodological quality of a study using, for example, a 7-point rating scale. Reliability of coding or rating should be reported. The quality weightings obtained for each study can then be used as (a) an adjustment mechanism in computing average effect size and (b) as a moderator variable to determine whether quality is, in fact, related to obtained effect size. Further details on quality assessment, weighting, and reliability are available in Hall, Tickle-Degnen, Rosenthal, and Mosteller (1994); Rosenthal (1991); and Wortman (1994).

Independence. For a database of any size, the meta-analyst soon discovers that many studies are not independent of one another; that is, the same participants have been used in two or more studies. Perhaps slightly different dependent variables were reported in the multiple reports on the same participants. For example, if responses had been recorded in video, audio, or transcript form, new ideas for dependent variables can be evaluated years later. Although such multiple usage of data archives can be scientifically valuable, they present a problem for the unwary meta-analyst. Most computational procedures dealing with significance testing require that the studies summarized be independent. Treating nonindependent studies as independent leads to significance test errors. These errors can be avoided by treating the several nonindependent studies as a single study with multiple dependent variables (R. Rosenthal, 1991; R. Rosenthal & Rubin, 1986; for a more technical treatment of problems of nonindependence, see Gleser & Olkin, 1994).

Minimum number of studies. What if meta-analytic efforts result in only a few studies retrieved? How few studies are too few for a meta-analysis? Meta-analytic procedures can be applied to as few as two studies; but when there are very few studies, the meta-analytic results are relatively unstable. When there are very few studies available on a given research question, it would be more economical of journal space and editors' and reviewers' time to incorporate the meta-analysis as an extension of the results section of the last study in the series of a few studies. Thus, if my study finds a correlation r between the two variables of interest, I might end my results section by combining and comparing my correlation and my p values with those few obtained earlier by other investigators.

Recorded Variables

Study characteristics. Describe what information was recorded for each study. For example, the number, age, sex, education, and volunteer status of the participants (R. Rosenthal & Rosnow, 1991) might be recorded for each study regardless of whether participants themselves were the sampling unit or whether classrooms, therapists, groups, wards, clinics, or other organizations served as the unit of analysis (e.g., the basis for computing degrees of freedom for the analysis). Was the study conducted in a laboratory or in the field? Was it an observational study or a randomized experiment? What was the year of publication and the form of publication (book, article, chapter, convention report, bachelor's or master's thesis, doctoral dissertation, technical report, or unpublished manuscript)? The particular study characteristics mentioned are just some of what are often included. However, each meta-analysis should also include all the variables that the meta-analyst's knowledge of the literature and intuition suggest may be important correlates of the magnitudes of the obtained effect sizes. More detailed discussions of the selection, coding, and evaluation of study characteristics have recently become available (Lipsey, 1994; Orwin, 1994; Stock, 1994). All of the foregoing study characteristics are used in two ways: as descriptions of the study set retrieved and as potential moderator variables.

Summarizing the characteristics. An overview of the various study characteristics is often valuable. The range and median of ages used in the assembled studies, of dates of published and unpublished studies, and of the proportions of sample participants who were female or male and the proportions found in various types of publication formats, of laboratory or field studies, and of studies that were randomized experiments rather than observational studies are readily summarized statistics that will be useful to readers.

Other moderator variables. All of the study characteristics recorded for each study and summarized for the set of studies can be used as moderator variables, that is, variables correlated with the magnitude of obtained effect size for the different studies. In addition to these fairly standard potential moderators, however, there are specific moderator variables with particular meaning for the specific area of research summarized.

For example, in a recent meta-analysis of "thin slices" of expressive behavior, short periods (under 5 min) of observation of expressive behavior were surprisingly predictive of various objective outcomes (Ambady & Rosenthal, 1992). One of the moderator variables examined was the presence or absence of verbal content accompanying the nonverbal behavior. It was found that studies including verbal content did not yield a higher average effect size of predictive accuracy. Another example of a moderator variable analysis grew out of a meta-analysis of the effects of teachers' expectations on pupils' IQ gains (Raudenbush, 1994). Using the moderator variable of how long teachers had known their pupils before the teachers were given randomly assigned, favorable expectations for pupils' IQ, Raudenbush (1994) found that the longer teachers had known their pupils before the experiment began, the smaller were the effects of experimentally induced teacher expectations.

Effect size estimates. Effect size estimates are the meta-analytic coin of the realm. Whatever else may also be recorded for each study, the estimated effect size should be recorded for each study in the meta-analysis.

The two main families of effect sizes are the r family and the d family. The two most important members of the former are Pearson's product–moment correlations (r) and Z_r, Fisher's r-to-z transformation. The three most important members of the d family are Cohen's d, Hedges's g, and Glass's Δ, all of which are differences between means divided by a standard deviation. Detailed explanations of these and other effect size estimates are given elsewhere (R. Rosenthal, 1991, 1994; for categorical data, see also Fleiss, 1994).

Significance levels. Though far less important than effect size estimates, significance levels should be recorded for each study unless the meta-analyst is certain that questions of statistical significance for the overall results of the meta-analysis will not arise. All such levels should be computed as accurately as possible and recorded as the one-tailed standard normal deviates associated with the p value. Thus, ps of .10, .01, .001, and .000001 are reported as Zs of 1.28, 2.33, 3.09, and 4.75, respectively. Results that are significant in the unpredicted or uncharacteristic direction are reported as negative Zs (e.g., if $p = .01$ one-tailed, but in the wrong direction, it is recorded as -2.33).

Results Section of a Meta-Analytic Review

Descriptive Data

The heart of a meta-analytic review is a description of the obtained effect sizes. Unless the number of studies is very small,

it is often very valuable to provide a visual display of the obtained effect sizes as well as various indices of central tendency and variability.

Visual display. A great many different visual displays may be useful under different conditions, and many of these are described by Cooper (1989); Glass et al. (1981); Greenhouse and Iyengar (1994); Hedges and Olkin (1985); Light and Pillemer (1984); Light, Singer, and Willett (1994); R. Rosenthal and Rosnow (1991); and Tukey (1977). Sometimes a specially prepared graphic can be most useful, one not found in any of these references. It would be instructive in that case to consult some of the general texts on visual displays, for example, those by Cleveland (1985), Kosslyn (1994), and Tufte (1983). However, there is not space here to illustrate even a few of the visual displays that may be instructive (e.g., box plots, funnel plots, and stem-and-leaf displays). As a single example of an often useful visual display, Tukey's stem-and-leaf display is a versatile picture of data that perfectly describes the distribution of results and retains each of the recorded effect sizes. Table 1 is a stem-and-leaf display from a recent meta-analysis of 38 studies on the predictive value of thin slices of nonverbal and verbal behavior. Each of the 38 effect sizes (r) is recorded with the first digit in the "stem" column and the second digit in the "leaf" column. The top three entries of Table 1, therefore, are read as three rs of .87, .73, and .74, respectively.

Central tendency. Several indices of central tendency should be reported, and differences among these indices should be discussed and reconciled. The unweighted mean effect size, the weighted mean effect size, and the median—and optionally, the proportion of studies showing effect sizes in the predicted direction—should be given. The number of independent effect sizes on which these indices are based should be reported and, optionally, the total number of participants on which the weighted mean is based and the median number per obtained effect size. The weighted mean effect size refers to weighting by size of study (e.g., df), but other weightings can also be used. For example, weighting may also be done by the quality of the study or by any other study characteristic likely to be of substantive or methodological interest. In larger meta-analyses, subsets of studies that can be meaningfully grouped together on the basis of study characteristics can be examined separately, subset by subset, with respect to central tendency or other descriptive features.

Table 1
Stem and Leaf Display of 38 Effect Size rs

Stem	Leaf
.9	
.8	7
.7	3, 4
.6	3, 8
.5	0, 2, 2, 3, 4, 4
.4	0, 0, 0, 1, 7
.3	1, 3, 5
.2	1, 1, 1, 2, 3, 3, 4, 5, 6, 6, 7, 8, 9
.1	0, 0, 4, 5, 6, 6
.0	

Note. Effect size rs are based on Ambady and Rosenthal (1992); rs include relationships between two continuous variables (r), two dichotomous variables (phi), and one dichotomous and one continuous variable (point biserial r).

Variability. The most important index of variability of effect sizes is simply their standard deviation. It is also helpful to give the maximum and minimum effect size and the effect sizes found at the 75th percentile (Q_3) and the 25th percentile (Q_1). For normally distributed effect sizes, the standard deviation is estimated at .75 ($Q_3 - Q_1$). Appendix A provides a checklist of descriptive data that should often, if not always, be reported.

Examining the distance (e.g., in units of S) of the maximum and minimum effect sizes from the mean, median, Q_1, and Q_3 of the full distribution of effect sizes is a useful start in data analysis for outliers. Valuable discussions of the problem of outliers are found in Barnett and Lewis (1978), Hedges and Olkin (1985), Hunter and Schmidt (1990), and Light and Pillemer (1984).

Several meta-analysts discuss the separation of the overall variability among effect sizes into components associated with "ordinary sampling error" and variability associated with other sources (Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Light & Pillemer, 1984). This can be especially valuable in alerting meta-analysts to "nonsampling error" variability that must then be investigated. However it should be noted, a conclusion that all the effect size variability is due to "ordinary sampling error" does not mean that meta-analysts cannot or should not investigate the variability by means of moderator variables. Indeed, scientific progress can be defined in terms of scientists' continually reducing the magnitude of sampling error by increasing their understanding of moderator variables.

Inferential Data

Significance testing. A good many procedures are available for testing the significance of an estimate of the typical effect size found in a particular meta-analysis (e.g., Mosteller & Bush, 1954, described 3; R. Rosenthal, 1991, described 9; and Becker, 1994, listed 18). One of the most generally useful of these methods is the Stouffer method in which one needs only to compute the standard normal deviate (Z) associated with each p value in the meta-analysis. Then, one simply adds all Zs (one per study) and divides the sum by $\sqrt{k}$, where k is the number of independent studies, to find the new Z that tests the overall result of the meta-analysis.

A related procedure for significance testing has been described in detail by Hedges, Cooper, and Bushman (1992). This procedure, the lower confidence limit (LCL) method, also yields a standard normal deviate, Z. The LCL Z and the Stouffer Z agree most of the time (nearly 99%); where they disagree, the LCL method may be more powerful unless the smaller studies summarized in the meta-analysis are associated with the larger effect sizes, a likely state of affairs. The LCL method tends to reject the null hypothesis when it is true (a Type I error) more often than does the Stouffer method; but because it may well be that the null hypothesis is essentially never true, that is not a serious problem (Cohen, 1994).

In both the Stouffer and LCL methods to get its magnitude, Z depends on the obtained effect sizes and the size of the studies, and it is interpreted as a fixed effect. That is, generalization of the results is to other participants of the type found in the spe-

cific k studies of the meta-analysis; generalization is not, ordinarily, to other studies.

Because of this limitation of the generalizability of fixed effect analyses, it is desirable also to use a random effects test of significance that permits generalization to other studies from the same population from which the retrieved studies were sampled. A simple one-sample t test on the mean effect size serves this purpose (Mosteller & Bush, 1954). For example, if one is working with Fisher Z-transformed rs, t is the mean Z_r, divided by the square root of the quantity SD^2/k, where SD is the standard deviation of Z_rs and k is the number of independent Z_rs. This t ($df = k - 1$) tends to be more conservative than Stouffer's Z but should nevertheless also be used because of its greater value in generalizing to other studies.

Another random effects approach to significance testing likely to be even more conservative than the one-sample t test is the one-sample $\chi^2(1)$ test of the null hypothesis in which there is no difference in the proportion of studies showing positive effect sizes rather than negative effect sizes. When there are fewer than 10 effect sizes, the binomial test tends to give more accurate p values than $\chi^2(1)$ (Siegel, 1956).

Note the difference between the fixed effect and the random effect view of the obtained results in the meta-analysis. When a meta-analyst adopts a fixed effect view of the results, the significance testing is based on the total number of sampling units (e.g., research participants, patients, or organisms), but the generalization is restricted to other sampling units that might have been assigned only to the same studies of the meta-analysis. The fixed effect good news, therefore, is greater statistical power; the bad news is more limited generalizability. When a meta-analyst adopts a random effect view of the results, the significance testing is based not on the total number of sampling units but only on the total number of studies included; the generalization is beyond the specific studies retrieved to others that can be seen to belong to the same population from which one obtained the studies. The random effect good news, therefore, is somewhat increased generalizability; the bad news is decreased statistical power. One should try not to be overly precise in an application of "random effects" because there is precious little random sampling of studies in meta-analytic work. Indeed, even in the fixed effect model, when one generalizes to other sampling units within the studies, one assumes that the new sampling units will be randomly sampled within the study from the same population from which one sampled the original sampling units. However, it is very seldom that in behavioral or biomedical research one samples participants or patients randomly. Hence, "random" should be thought of as quasi-random at best.

To give an intuitive feel for the fixed versus random effect issue, Tables 2 and 3 have been prepared. Table 2 shows a simple meta-analytic model in which 10 studies have been retrieved, each with a treatment and a control condition of 20 participants in each of the $2 \times 10 = 20$ cells. Table 3 shows the expected mean squares and F tests when studies are regarded as fixed versus random (Snedecor & Cochran, 1989). With treatment always regarded as a fixed effect, the F tests for studies and for the Treatment × Studies interaction are the same whether studies are regarded as fixed or random. However, the treatment effect is tested against different error terms when studies are fixed versus random, and the degrees of freedom for the F test

Table 2
Meta-Analytic Model Illustrating Fixed Versus Random View of Summarized Studies

Study	Condition	
	Treatment	Control
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Note. Assume $n = 20$ for each of the $2 \times 10 = 20$ cells.

are also different. In Tables 2 and 3, when studies are viewed as fixed, the error term is the one expected to be the smallest (variation within cells), and $df = 380$. When studies are viewed as random, the error term will often be larger than when viewed as fixed, to the extent that there are nonzero Treatment × Study interaction effects, and the df will be smaller (9 instead of 380 in this example). The most recent (and more detailed) discussions of the fixed versus random effect issue can be found in Hedges (1994), Raudenbush (1994), and Shadish and Haddock (1994).

Confidence intervals. Confidence intervals should be computed around the mean effect size, preferably using a simple random effects approach. That is, the standard error of the mean effect size estimate (e.g., Z_r) should be computed as $S/\sqrt{k}$, with k being the number of independent effect sizes. The 95% confidence interval should be recorded at least; sometimes it is also useful to give the 90%, the 99%, and other intervals as well.

For example, suppose $k = 25$ independent studies available with an unweighted mean $d = .50$ and a standard deviation of these 25 ds = 1.00. Then the standard error for the 25 ds is $SD/\sqrt{k} = 1.00/\sqrt{25} = .20$. The 95% confidence interval is then given by the rough and ready mean $d \pm 2\,(SE)$ or $.50 \pm 2(.20) = .10$–$.90$. A more accurate interval is obtained by replacing the 2 by the critical .025 one-tailed value of t for the appropriate df, that is, $k - 1$. That critical value of t for $k = 25$ ($df = 24$) is 2.06. Therefore, in this example, the confidence interval is $.50 \pm (2.06)(.20) = .09$–$.91$ (R. Rosenthal & Rubin, 1978). The interpretation of this confidence interval is that if the claim of the effect size for the population (from which the 25 studies must be viewable as a random sample) falls within the 95% confidence interval, the claim will be correct 95% of the time.

This example is based on the conservative random-effects procedure in which studies, not individuals within studies, are used as the sampling unit. It is often useful also to compute confidence intervals in which individuals rather than studies are used as the sampling unit. However, the confidence intervals obtained by such procedures can appear dramatically more optimistic (i.e., narrower) than those based on the previously illustrated random effects procedures. Computational procedures for confidence intervals on the basis of individuals as sampling

Table 3
EMSs and F Tests When Studies Are Viewed as Fixed Versus Random

		Studies fixed[a]		Studies random[b]	
Source	*df*	EMS	*F*	EMS	*F*
T	1	$\sigma^2 + 200K_T^2$	T/U	$\sigma^2 + 20\sigma_{TS}^2 + 200K_T^2$	T/TS
S	9	$\sigma^2 + 40K_S^2$	S/U	$\sigma^2 + 40\sigma_S^2$	S/U
TS	9	$\sigma^2 + 20K_{TS}^2$	TS/U	$\sigma^2 + 20\sigma_{TS}^2$	TS/U
U	380	σ^2		σ^2	

Note. EMS = expected mean square; T = treatment (fixed effect); S = studies; TS = Treatment × Studies interaction; U = units in cells; K = population variance of the effect in question.
[a] These 10 studies are recognized as the entire population of studies that are of interest. [b] These 10 studies are regarded as a "random" sample from a larger population of studies to which the meta-analyst would like to generalize.

units are described in varying degrees of detail by Hedges (1994), Hedges and Olkin (1985), Hunter and Schmidt (1990), and Shadish and Haddock (1994).

Heterogeneity tests. Statistical tests of the heterogeneity of significance levels (R. Rosenthal & Rubin, 1979) and of effect size estimates (Hedges, 1982; R. Rosenthal & Rubin, 1982b) are readily available. Usually one is more interested in the heterogeneity of effect sizes than of significance levels, and it is often useful to present the results of such an analysis. However, two common problems in the use of these tests must be pointed out.

First, there is a widespread belief that a test of heterogeneity must be found to be significant before contrasts can be computed among the obtained effect sizes; this is not the case. Contrasts, particularly planned contrasts, can and should be computed among the obtained effect sizes whether the overall test of heterogeneity is significant or not. The situation is identical to that in a one-way analysis of variance where many investigators believe it is improper to compute contrasts unless the overall F is significant. Actually, planned contrasts should be computed without reference to the overall F, and even unplanned contrasts can be computed with appropriate adjustments of their levels of significance (R. Rosenthal & Rosnow, 1985, 1991). If overall tests of heterogeneity are not to serve as licenses to pursue contrast analyses, why compute them at all? They do provide some useful information. If very significant, they alert the meta-analyst to the likelihood that all the effect sizes are not cut from the same cloth and that he or she should try to find the moderator variables accounting for the significant heterogeneity of the effect sizes. Thus, a very significant χ^2 for heterogeneity "morally" obligates one to search for moderators, but a nonsignificant χ^2 does not preclude the search.

The second common problem in the use of heterogeneity tests is to treat them as though they were estimates of the magnitude of heterogeneity; they are not. They are tests of significance and as with all tests of significance they are a function of the magnitude of the effect and the sample sizes. Thus, the widely varying (SD = .40) effect sizes (r) .80, .40, and .00 may not differ significantly if they are based on small sample sizes (e.g., n = 10), whereas the homogeneous (SD = .05) rs of .45, .40, and .35 may differ significantly if they are based on large sample sizes (e.g., n = 800). The magnitude of the effect size heterogeneity is given by the indices of variability previously described—in particular by the standard deviation of the effect sizes.

Some meta-analysts like to present separately one or both of the ingredients of the standard deviation of the effect size. These two ingredients can be illustrated by examining in Table 3 the expected mean squares for the Treatment × Studies interaction when studies are viewed as random. The two components of variance are σ^2 and σ_{TS}^2. The estimate of σ^2 is obtained directly from the mean square for units nested in conditions, and the estimate of σ_{TS}^2 is obtained in two steps:

$$\text{(a) } MS_{TS} - MS_U = (\sigma^2 + 20\sigma_{TS}^2) - (\sigma^2) = 20\sigma_{TS}^2)$$

and

$$\text{(b) } \sigma_{TS}^2 = \frac{20\sigma_{TS}^2}{20},$$

where 20 is the number of units in each cell. The estimate of σ^2 gives the basic "noise level" of the dependent variable, whereas the estimate of σ_{TS}^2 gives the interaction variation of the study outcomes above that basic noise level.

Contrasts. The statistical significance of the relationship between a moderator variable and the obtained effect sizes is given by the computation of a contrast test (R. Rosenthal, 1991; R. Rosenthal & Rubin, 1982b; or more complex procedures of fitting models to effect size data in the spirit of multiple regression, Hedges & Olkin, 1985). As with the case for tests of heterogeneity, the tests of significance of contrasts do not give a direct indication of the magnitude of the moderator variable's relationship to the obtained effect sizes. Such an indication is readily available, however, simply by correlating the obtained effect sizes with their corresponding "score" on the moderating variable. Such a correlation, in which the sample size is the number of independent studies, reflects a random effects view of the data with generalizability to other potential results drawn from the same population that yielded the obtained results. When the number of studies retrieved is quite small, such correlations of effect sizes with their moderators are not very stable, and a meta-analyst may be forced to take a less generalizable, fixed effect view of the data (Raudenbush, 1994). In such cases, a meta-analyst can get a serviceable indicator of the moderator effect's magnitude by dividing the obtained test of the significance of the contrast, Z, by the square root of the sum of the sample sizes contributing to the computation of Z. This fixed effect type r tends to be smaller than the random effects r but tends to be associated with a more significant test statistic. Appendix B provides a checklist of inferential data that should often, if not always, be reported.

Interpretive Data

In this section, a number of procedures and statistics are summarized that are often useful in helping to understand and interpret the descriptive and inferential data of the meta-analysis. They are described here more as a reminder of their availability and usefulness than as a standard requirement of all meta-analyses.

Binomial effect size display. The binomial effect size display (BESD) is a procedure that shows the practical importance of an effect size (R. Rosenthal & Rubin, 1982a). The input to the BESD is a specific effect size estimate, the Pearson r; but because any other effect size estimate can be converted to r, the BESD can be used to display the mean or median effect size estimate of any meta-analysis.

In a BESD, the Pearson r is shown to be the simple difference in outcome rates (e.g., proportion successful or proportion performing above the overall median) between the experimental and control groups in a standard table, column, and row, totals of which always add up to 100. The BESD is computed from any obtained effect size r by computing the treatment condition success rate as $.50 + r/2$ and the control condition success rate as $.50 - r/2$. Thus, an r of .20 yields a treatment success rate of $.50 + .20/2 = .60$ and a control success rate of $.50 - .20/2 = .40$, or a BESD of

Condition	Success	Failure	Σ
Treatment	60	40	100
Control	40	60	100
Σ	100	100	200.

Had a meta-analyst been given the BESD to examine before knowing r, he or she could easily have calculated it mentally; r is simply the difference between the success rates of the experimental versus control group (.60 − .40 = .20).

Coefficient of robustness. Although the standard error of the mean effect size along with confidence intervals placed around the mean effect size are of great value (R. Rosenthal & Rubin, 1978), it is sometimes helpful to use a statistic that does not increase simply as a function of the increasing number of replications. Thus, if a meta-analyst wants to compare two research areas for their robustness, adjusting for the difference in number of replications in each research area, he or she may prefer the robustness coefficient, which is simply the mean effect size divided by the S of the effect sizes. This metric is the reciprocal of the coefficient of variation (R. Rosenthal, 1990, 1993). The coefficient of robustness (CR) can also be viewed in terms of the one-sample t test on the mean of the set of k effect sizes. Thus, CR is given by $t/\sqrt{k}$, or t adjusted for the number of studies.

The usefulness of this coefficient is based on two ideas—first, that replication success, clarity, or robustness depends on the homogeneity of the obtained effect sizes, and second, that it also depends on the unambiguity or clarity of the directionality of the result. Thus, a set of replications grows in robustness as the variability (S) of the effect sizes (the denominator of the coefficient) decreases and as the mean effect size (the numerator of the coefficient) increases. Incidentally, the mean may be weighted, unweighted, or trimmed (Tukey, 1977). Indeed, it need not be the mean at all but any measure of location or central tendency (e.g., the median).

The CR can be seen as a kind of second-order effect size. As an illustration, imagine that three meta-analyses of three treatments have been conducted with mean effect size ds of .8, .6, and .4, respectively. If the variability (S) of the three meta-analyses were quite similar to one another, the analysis showing the .8 mean d would, of course, be declared the most robust. However, suppose Ss for the three analyses were 1.00, 0.60, and 0.20, respectively. Then the three CRs would be .8/1.00 = .8, .6/.60 = 1.0, and .4/.20 = 2.0. Assuming reasonable and comparable sample sizes and numbers of studies collected for the three analyses, the treatment with the smallest effect size (i.e., .4) would be declared most robust, with the implication that its effect is the most consistently positive.

Counternull. A new statistic was recently introduced to aid the understanding and presentation of research results: the counternull value of the obtained effect size (R. Rosenthal & Rubin, 1994). The counternull statistic is useful in virtually eliminating two common errors: (a) equating failure to reject the null with the estimation of the effect size as equal to zero and (b) equating rejection of a null hypothesis on the basis of a significance test with having demonstrated a scientifically important effect. In most meta-analytic applications, the value of the counternull is simply twice the magnitude of the obtained effect size (e.g., d, g, Δ, Z_r). Thus, with mean $r = .10$ found to be nonsignificant, the counternull value of $r = .20$ is exactly as likely as the null value of $r = .00$. For any effect size with a symmetric reference distribution such as the normal or any t distribution, the counternull value of an effect size can always be found by doubling the obtained effect size and subtracting the effect size expected under the null hypothesis (usually zero). Thus, if meta-analysts found that the overall test of significance of the mean effect size (e.g., $\bar{d}$ or $\bar{z}_r$) did not reach the chosen level (e.g., .05), the use of the counternull would keep them from concluding that the mean effect size was, therefore, probably zero. The counternull value of $2\,\bar{d}$ or $2\bar{z}_r$ would be just as tenable a conclusion as concluding $\bar{d} = 0$ or $z_r = 0$.

File drawer analysis. The *file drawer problem* refers to the well-supported suspicion that the studies retrievable in a meta-analysis are not likely to be a random sample of all studies actually conducted (R. Rosenthal, 1991). The suspicion has been that studies actually published are more likely to have achieved statistical significance than the studies remaining squirreled away in the file drawers (Sterling, 1959). No definitive solution to this problem is available, but reasonable boundaries can be established on the problem, and the degree of damage to any research conclusion that could be done by the file drawer problem can be estimated. The fundamental idea in coping with the file drawer problem is simply to calculate the number of studies averaging null results that must be in the file drawers before the overall probability of a Type I error can be brought to any desired level of significance, say $p = .05$. This number of filed studies, or the *tolerance for future null results,* is then evaluated for whether such a tolerance level is small enough to threaten the overall conclusion drawn by the meta-analyst. If the overall level of significance of the research review is brought down to the *just significant* level by the addition of just a few more null results, the finding is not resistant to the file drawer threat.

Details of the calculations and rationale are given elsewhere (R. Rosenthal, 1991); briefly, a meta-analyst finds the number (X) of new, filed, or unretrieved studies averaging null results required to bring the new overall p to .05 with the following equation: $X = [(\Sigma Z)^2/2.706] - k$, where ΣZ is the sum of the standard normal

deviates associated with the one-tailed ps of all the k studies retrieved.

Meta-analysts should note that the file drawer analysis addresses only the effects of publication bias on the results of significance testing. Very sophisticated graphic (Light & Pillemer, 1984) and other valuable procedures are available for the estimation and correction of publication bias (e.g., Begg, 1994; Hedges & Olkin, 1985; Hunter & Schmidt, 1990).

Power analysis. In large meta-analyses, it is usually the case that the null hypothesis is found to be unlikely at a very low p value. In smaller meta-analyses, however, it can happen that the overall results are not found to be significant. Before concluding that the population value of the effect size is zero, it is helpful to perform a power analysis along with computing the counternull value of the overall obtained effect size. In this application, meta-analysts should assume a population effect size equivalent to the actually obtained overall effect size and simply use Cohen's (1977, 1988) tables to find the power at which the null hypothesis is tested. If that power level is low, the evidence for the null hypothesis is weak and should be reported as such. Appendix C provides a checklist of interpretive data that should often be considered and reported when appropriate.

Discussion Section of a Meta-Analytic Review

The discussion section could begin with a summary of the meta-analytic results, followed by tentative explanations of these results. These explanations may be in terms of the theories of the area in which the meta-analysis was done, or they may require new theory (Hall et al., 1994). The implications for theory—old or new—for practice, if relevant, and for further primary level research could be discussed.

The overall goal of the discussion may be seen as the answer to the question, "Where are we now that this meta-analysis has been conducted?" The meta-analysis is placed into the context of the field, and the field, very often, is placed into the context of the meta-analysis.

References and Appendix to a Meta-Analytic Review

The reference list should include full references for each of the studies included in the meta-analysis, with the following text directly under the heading: "Studies preceded by an asterisk were included in the meta-analysis"; and an asterisk should be inserted before each reference entry.

An appendix in the form of a table should give for each of the included studies the overall effect size, the sample size, the Z corresponding to an accurate p value, and the coded or rated "score" for each study of the primary study characteristics and moderator variables used in the meta-analysis. The journal editor and reviewers will then have important information to guide them in their evaluation of the meta-analysis. If this appendix table makes the article too long, the author note should include where to get a copy of it.

Conclusion

Most reviews of the literature should be quantitative, just as most primary research studies should be quantitative. The statistical procedures used in meta-analyses range from the basic to the very complex, as do the statistical procedures of primary research studies. There is no one way to do a meta-analysis or to report a meta-analysis, any more than there is just one way to do or to report the data analysis of a primary research study. Therefore, the goal of this article was not prescriptive in the sense that every meta-analysis should include everything suggested in this article. The goal instead was to provide some general guidelines that may be considered by meta-analysts following the standard procedures of the various authors of meta-analytic textbooks. My own bias has been to keep it simple, basic, and intuitive. Even when complex analyses are undertaken, their reporting should be kept simple, basic, and intuitive. When one writes a meta-analytic review, after all, it is intended for a far larger audience than the other authors of texts and articles on meta-analytic methodology.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111,* 256–274.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data.* New York: Wiley.

Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 215–230). New York: Russell Sage Foundation.

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

Cleveland, W. S. (1985). *The elements of graphing data.* Monterey, CA: Wadsworth.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

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

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

Dickersin, K. (1994). Research registers. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 71–83). New York: Russell Sage Foundation.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

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

Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 383–398). New York: Russell Sage Foundation.

Hall, J. A., Tickle-Degnen, L., Rosenthal, R., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 17–28). 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. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.

Hedges, L. V., Cooper, H., & Bushman, B. J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin, 111,* 188–194.

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

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

Kosslyn, S. M. (1994). *Elements of graph design.* New York: Freeman.

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. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 439–453). New York: Russell Sage Foundation.

Lipsey, M. W. (1994). Identifying potentially interesting variables and analysis opportunities. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 111–123). New York: Russell Sage Foundation.

Mosteller, F. M., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Vol. 1. Theory and method* (pp. 289–334). Cambridge, MA: Addison-Wesley.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 139–162). New York: Russell Sage Foundation.

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

Reed, J. G., & Baxter, P. M. (1994). Using reference databases. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 57–70). New York: Russell Sage Foundation.

Rosenthal, M. C. (1994). The fugitive literature. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 85–94). New York: Russell Sage Foundation.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality, 5,* 1–30.

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

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues.* Hillsdale, NJ: Erlbaum.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* Cambridge, England: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3,* 377–386.

Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin, 86,* 1165–1168.

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

Rosenthal, R., & Rubin, D. B. (1982b). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400–406.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5,* 329–334.

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

Siegel, S. (1956). *Nonparametric statistics.* New York: McGraw-Hill.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames: Iowa State University Press.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 125–138). New York: Russell Sage Foundation.

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

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

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

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 97–109). New York: Russell Sage Foundation.

(Appendixes follow on next page)

Appendix A

Checklist of Descriptive Data for the Results Section

Visual Displays of Effect Sizes (Often Useful)

stem-and-leaf plots (as in Table 1)
box plots (if many are to be compared)
funnel plots (e.g., to investigate publication bias)
other plots (as needed)

Central Tendency

unweighted mean
weighted mean[A1]
median (repeated for convenience as Q_2 below)
proportion of positive effects
k (the number of independent studies)
N (the number of independent participants)
n (median number of participants per study)

Variability

SD (the standard deviation)[A2]
maximum effect size[A3]
Q_3 (75th percentile effect size)
Q_2 (50th percentile effect size)
Q_1 (25th percentile effect size)
minimum effect size[A3]
normal-based $SD = .75\ (Q_3 - Q_1)$

[A1] Weighting is usually by degrees of freedom; means weighted by study quality or by other weightings should also be reported, if computed.
[A2] It is also often valuable to report separately the variability "corrected" for sampling variation.
[A3] This is useful in a preliminary check for outliers.

Appendix B

Checklist of Inferential Data for the Results Section

Significance Testing

combined (Stouffer) Z (and other such tests as needed)
t test (one-sample)
test of proportion positive (Z)

Confidence Intervals

	From	To
90% (optional)		
95% (almost always desirable)		
99% (optional)		
99.9% (optional)		
standard error ($S/\sqrt{k}$)		

Heterogeneity Tests

$\chi^2(k-1)$
p of χ^2
S (magnitude of heterogeneity or other indices of magnitude not dependent on sample size)

Contrasts

For each contrast or predictor variable give
test of significance
effect size for contrast.

Appendix C

Checklist of Interpretive Data for the Results Section

Binomial Effect Size Display procedure

	Dependent variable		
Independent variable	High	Low	Total
High			100
Low			100
Total	100	100	200

Coefficient of robustness: M/SD[a]
Counternull (especially if overall results not significant)
File Drawer analysis (tolerance for future null results)
Power analysis (if overall results not significant)

[a] Several coefficients may be reported using weighted or unweighted mean or median effect size for the numerator and weighted or unweighted standard deviation for the denominator.

Received January 14, 1994
Revision received January 9, 1995
Accepted January 10, 1995 ■