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

...
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 Liebeng (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 weightier 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 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.
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 conventional displays may also be useful under different conditions, 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 regression displays).

**Central tendency.** Several indices of central tendency should be reported and, optionally, the median—and optionally, 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 the weighted mean is based should be reported and, optionally, the confidence limits for the weighted mean. The confidence limits may be obtained by weighting by the square root of the number of observations on which the effect size is based, by weighting by the inverse of the standard error of the effect size, or by any other method.

**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 ($Q_1$). If normally distributed effect sizes are found in Barnett and Lewis (1978), Hedges and Olkin (1985), Hunter and Schmidt (1990), and Light and Pillemer (1984). This can be especially valuable in alerting meta-analysts to "nonsampling error" variability that must 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. In this context, moderator variables are defined simply as variables that are hypothesized to affect the magnitude of the effect size under study. The 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 statistics that should often, if not always, be reported.
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 $r$s, $t$ is the mean $Z$, divided by the square root of the quantity $SD^2/k$, where $SD$ is the standard deviation of $Z$s and $k$ is the number of independent $Z$s. 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 $x^2(1)$ 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 $x^2(1)$ (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 on the total number of sampling units (e.g., research participants, patients, or organisms), but the generalization is not, ordinarily, to other studies.

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

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

Confidence intervals. Confidence intervals should be computed around the mean effect size, preferably using a simple ran-

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 $\times$ 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 ran-
Table 3

<table>
<thead>
<tr>
<th>Source</th>
<th>EMS fixed*b</th>
<th>EMS random*b</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>$\sigma^2 + 200K_T$</td>
<td>$\sigma^2 + 20\sigma_T^2 + 200K_T$</td>
</tr>
<tr>
<td>S</td>
<td>$\sigma^2 + 40K_S$</td>
<td>$\sigma^2 + 40\sigma_S^2$</td>
</tr>
<tr>
<td>TS</td>
<td>$\sigma^2 + 20K_{TS}$</td>
<td>$\sigma^2 + 20\sigma_{TS}$</td>
</tr>
<tr>
<td>U</td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

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.

* These 10 studies are regarded as a “random” sample from a larger population of studies to which the meta-analyst would like to generalize.

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 $\sigma^2$ and $\sigma_{TS}^2$. The estimate of $\sigma^2$ is obtained directly from the mean square for units nested in conditions, and the estimate of $\sigma_{TS}^2$ is obtained in two steps:

(a) $\text{MST}_S - \text{MS}_U = (\sigma^2 + 20\sigma_{TS}^2) - (\sigma^2) = 20\sigma_{TS}^2$

and

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

where 20 is the number of units in each cell. The estimate of $\sigma^2$ gives the basic “noise level” of the dependent variable, whereas the estimate of $\sigma_{TS}^2$ 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 cases 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 may 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 procedure that shows the practical importance of the effect size. The Pearson correlation coefficient $r$, when weighted, unweighted, or trimmed (Tukey, 1977), is a specific effect size estimate, the Pearson $r$, of .8, .6, .4, and .2. An illustration, imagine that three meta-analyses of three treatments were quite similar to one another, the analysis showing the .8 mean $d$ would, of course, be declared the most robust. However, suppose $S$s for the three analyses were 1.00, 0.60, and 0.20, respectively. Then the three CRs would be $.8/1.00 = .8$, $.6/0.60 = 1.0$, and $.4/0.20 = 2.0$. 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.

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 coefficient of robustness, 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 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., $d$, $g$, $\Delta$, $Z_e$). 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 $\bar{Z}$) 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 $2\bar{Z}$, would be just as tenable a conclusion as concluding $\bar{d} = 0$ or $\bar{Z} = 0$.

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 ($Y$) of
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 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.


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*¹
- 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)²
- maximum effect size³
- \( Q_1 \) (75th percentile effect size)
- \( Q_2 \) (50th percentile effect size)
- \( Q_3 \) (25th percentile effect size)
- minimum effect size³
- normal-based \( SD = .75 (Q_3 - Q_1) \)

*¹ Weighting is usually by degrees of freedom; means weighted by study quality or by other weightings should also be reported, if computed.
*² It is also often valuable to report separately the variability "corrected" for sampling variation.
*³ 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 (\( Z \))

Heterogeneity Tests
- \( \chi^2(k - 1) \)
- \( p \) of \( \chi^2 \)
- magnitude of heterogeneity or other indices of magnitude

Appendix C

Checklist of Interpretive Data for the Results Section

Binomial Effect Size Display procedure

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Independent variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Coefficient of robustness: \( M/SD \)

Counternull (especially if overall results not significant)

File Drawer analysis (tolerance for future null results)

Power analysis (if overall results not significant)

* 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.