META-ANALYSIS: Recent Developments in Quantitative Methods for Literature Reviews

R. Rosenthal and M. R. DiMatteo
Department of Psychology, University of California, Riverside, Riverside, California 92521

Key Words  effect size, graphic displays, contrast analysis, moderator analysis, heterogeneity analysis

Abstract  We describe the history and current status of the meta-analytic enterprise. The advantages and historical criticisms of meta-analysis are described, as are the basic steps in a meta-analysis and the role of effect sizes as chief coins of the meta-analytic realm. Advantages of the meta-analytic procedures include seeing the “landscape” of a research domain, keeping statistical significance in perspective, minimizing wasted data, becoming intimate with the data summarized, asking focused research questions, and finding moderator variables. Much of the criticism of meta-analysis has been based on simple misunderstanding of how meta-analyses are actually carried out. Criticisms of meta-analysis that are applicable are equally applicable to traditional, nonquantitative, narrative reviews of the literature. Much of the remainder of the chapter deals with the processes of effect size estimation, the understanding of the heterogeneity of the obtained effect sizes, and the practical and scientific importance of the effect sizes obtained.

CONTENTS
INTRODUCTION ................................................ 60
ENTER META-ANALYSIS ........................................ 61
THE ADVANTAGES OF META-ANALYSIS ....................... 63
Seeing the Landscape of a Research Enterprise .................. 63
Keeping Statistical Significance In Perspective .................. 63
Toward Wasting No Data ...................................... 64
Intimacy with the Data ........................................ 64
Focused Research Hypotheses ................................... 65
Identifying Moderator Variables .................................. 65
CRITICISMS OF META-ANALYSIS ............................. 66
Bias in Sampling the Findings ................................... 66
Garbage In and Garbage Out ................................... 66
As the twenty-first century unfolds, and scientific research in nearly every field is growing almost explosively, new findings daily “overthrow” old ones, and a “relentless cross fire” (Hunt 1997:1) sometimes occurs as researchers and the public try to make sense of the stories that our scientific data are trying to tell us. Findings are often confusing and conflicting about central issues of theory and practice, not only in psychology more narrowly defined, but in the related domains of education, medicine, and other biopsychological and sociopsychological disciplines. Do vitamins prevent cancer and does exercise extend life span? Does yo-yo dieting pose significant health problems? Does exposure to electromagnetic fields increase the risk of brain cancer and leukemia? Does noncompliance with medical treatment regimens affect outcomes? Does the Head Start program work? What is the relationship between gender and various kinds of social behavior, such as helping, nonverbal communication, and conforming to group values? These, and many other topics, have been addressed by myriad studies that have varying outcomes—some show effects in one direction and some in the opposite, and some show effects that are close to zero.

A resolution of conflicting evidence regarding these outcomes is often necessary for further advance of a field and for any practical application. Further, social and health policy sometimes demands accurate estimation of certain descriptive statistics that, in the available research, show such variability that making intervention decisions based on them is both challenging and precarious. For example, as Hunt (1997) notes, estimates of the occurrence of mental health problems in the homeless range from 1% to 70%, and estimates of drug problems range from 2% to 90%. Patient noncompliance with medical treatment...
regimens is estimated to range between 20% and 70% (DiMatteo & DiNicola 1982).

ENTER META-ANALYSIS

The quantitative procedures of meta-analysis help to address some of the challenges introduced by the existence of multiple answers to a given research question. Meta-analysis allows the combining of numerical results from a few or many studies, the accurate estimate of descriptive statistics (Hedges 1987, Rosenthal 1978) and the explanation of inconsistencies as well as the discovery of moderators and mediators in bodies of research findings.

Meta-analysis allows researchers to arrive at conclusions that are more accurate and more credible than can be presented in any one primary study or in a nonquantitative, narrative review.

Meta-analysis began with a medical problem. This is perhaps appropriate because inquiries in the field of biomedicine often demand immediate answers to complex and multifaceted questions in which existing data may be quite variable and clinical steps depend upon reconciliation of disparate findings. In the case of the first meta-analysis, the year was 1904 and Karl Pearson collected correlation coefficients to determine the extent to which inoculation against smallpox was related to survival (Pearson 1904). The unweighted mean of the correlations between inoculation and survival was 0.63, the weighted mean $r$ was 0.64, and the median was 0.61—a truly huge effect with massive clinical significance. Pearson’s process of combining of research results across many studies was an unusual approach in 1904. Had his technique continued to be regularly employed throughout the first three-quarters of the twentieth century, cumulative research may have been brought to bear on some frustrating dilemmas in clinical treatment (Hunt 1997: Ch. 4, Robin 1984). Instead, it was not until the latter two decades of the twentieth century that meta-analysis became popular in fields such as biomedicine, the behavioral sciences, the interface of the two (the fields of health psychology, medical psychology, and behavioral medicine), and others. This popularity has come about partly because these disciplines generate too much information to manage easily, and methods are needed to synthesize that information. In medicine, for example, over two million medical articles are published every year. When different researchers, or even the same researchers, try to study a phenomenon more than once, they are bound to find different results from study to study, making it difficult to make sense of a body of research using narrative methods of synthesis. As Hall and Rosenthal (1995) have noted, there has been a shift in perspective recently such that a broader and more objective view of research is emerging and the “landscape” or distribution of results has become of greater interest than the results of individual studies.

Reviews of research have been valuable to many fields, but when presented and described only qualitatively, the results of conflicting studies can be confusing.
Qualitative or narrative methods approach controversy by listing and describing conflicting findings, and sometimes by trying to group or otherwise configure those that have various types of results or outcomes. Yet, it may be all too tempting for authors of narrative reviews consciously or unconsciously to select and describe studies to support their own understanding of the literature and/or their own established theoretical positions.

In an effort to avoid such influence—specifically preconceptions about the effectiveness of psychotherapy—Gene Glass and Mary Lee Smith conducted an extensive, systematic review of the entire literature on its outcomes, including every empirical article that provided an effect size (or the data to compute it) estimating the magnitude of the relationship between psychotherapy and an outcome (Smith & Glass 1977, Smith et al 1980). This research marks the beginning of the meta-analytic movement in psychology; the term meta-analysis was first used by Gene Glass in his 1976 presidential address to the American Educational Research Association. In this work, it became clear that meta-analysis is more than a statistical technique; it is a methodology for systematically examining a body of research, carefully formulating hypotheses, conducting an exhaustive search and establishing inclusion/exclusion criteria for articles, recording and statistically synthesizing and combining data and effect sizes from these studies, searching for moderator and mediator variables to explain effects of interest, and reporting results. Writing in the *Annual Review of Psychology*, Green & Hall (1984:52) foretold the future of this versatile and useful approach to scientific research: “...careful quantitative reviews are likely to play a larger role in further advances in psychology.” Their prediction was accurate. That year, PsychLit catalogued 89 English-language entries with the key word “meta-analysis;” in 1999, that number was 262. In the medical literature database, Medline, only 34 English-language citations with the key word: “meta-analysis” were catalogued in 1984; in 1999, there were 823 entries.

By virtue of its ability to extract fairly clear answers from the research literature, meta-analysis has likely made a big difference in the lives of patients of medicine and psychotherapy by providing answers to clinical questions about their care—answers that might not be available from a morass of conflicting research findings. For example, from a meta-analysis of the effects on mortality rates of bypass surgery versus medical therapy for ischemic heart disease, it was found that mortality owing to bypass was 10.2%, whereas for medical management mortality was 15.8% (Yusuf et al 1994). In another meta-analysis, this one on the use of antibiotics prior to colon surgery, effects combined across 26 clinical trials showed that antibiotic therapy reduced infection from 36% to 22%, and reduced death rates from 11.2% to 4.5% (Baum et al 1981). Finally, a meta-analysis of more than 200 studies on the effects of viewing violent TV demonstrated a greater tendency toward aggressive/antisocial acts after viewing TV violence against another person (Paik & Comstock 1994). The above three research domains had been burdened with controversy because the studies yielded such varying results that an overall message from the data seemed impossible—that is, until meta-analysis.
META-ANALYSIS

THE ADVANTAGES OF META-ANALYSIS

Seeing the Landscape of a Research Enterprise

Meta-analysis has come to occupy a major place in contemporary scientific research partly because, as demonstrated in the above examples, it helps overcome much of the equivocation about research findings in the social sciences and medicine by providing a method for combining research results. Meta-analysis is valuable for several other reasons as well. The methodology requires us to be extremely thorough in our search for relevant research reports and requires careful review and analysis of all of the published, and often the unpublished, data available on a specific research question. Thus, meta-analysis keeps us from relying on the results of a single study or a narrative, nonquantitative review in attempting to understand a phenomenon. A cumulative view of psychology and other sciences provides the opportunity to view the whole picture in a research enterprise, and meta-analysis helps us see the similarities and differences among the methodologies and the results of many studies.

Keeping Statistical Significance In Perspective

In the literature of psychology and medicine, among other fields, researchers often refer to the statistical significance of a finding; significance is considered good, and nonsignificance is considered bad. Yet, the significance of any given effect size will be determined by the size of the sample studied. The simple equation, in prose, is:

\[ \text{Significance Test} = \text{Effect Size} \times \text{Study Size}. \]

The implications of this equation for understanding a body of research are considerable. For example, two studies with exactly the same effect sizes can vary greatly in their significance level depending simply upon the number of participants or other sampling units employed. A focus on significance has often misled us in traditional narrative reviews of the literature, whereas in quantitative reviews, we typically focus on effect sizes. Meta-analysis prevents our reliance on the significance test of any one finding as a measure of its value and helps us realize that repeated results in the same direction across several studies, even if not one is significant, are much more powerful evidence than a single significant result. For example, two results at \( p = .06 \) are much stronger evidence \( (p = 0.014) \) against the null hypothesis than is one 0.05 result; and ten results at \( p = 0.10 \) are stronger evidence \( (p = 0.000025) \) against the null than are five at \( p = 0.05 \ (p = 0.00012) \). Meta-analysis thus provides the opportunity for even small and nonsignificant effects to contribute to the overall picture of the results of a research enterprise. As has been demonstrated elsewhere, in biomedical research in particular, the clinical application of what is learned from the cumulation of even very small effects can save many lives (Rosenthal 1995a).
Toward Wasting No Data

Data collection/acquisition in psychology, and many other fields of science, can range from difficult to intensely frustrating, and from modestly to exceedingly expensive in terms of finances, time, trouble, and opportunity costs. It is fair to suggest that no appropriately collected data from a well-designed study should ever be wasted. In practice, however, data are wasted all the time—such as when researchers fail to write-up results that were not significant, and when journals reject articles with nonsignificant findings. Meta-analysis allows the combination of results from studies with samples so small that they never achieve statistical significance. Approaching the analysis of pilot studies with cumulative techniques of meta-analysis allows researchers to further a field in which the nature of the research precludes large studies. In Science, for example, Cohen (1993) reported on two pilot studies in which experimental monkeys were vaccinated with simian immunodeficiency virus (SIV; akin to HIV). Control monkeys were not vaccinated. Because of the complexity of such research and limits on availability of experimental animals, Study One had only three experimental and three control monkeys. The experimental animals had better health outcomes (two of three improved) than did the control animals (none improved). In the second pilot, experimental animals also did better (two of five improved) than controls (zero of six improved). Neither of these pilot studies showed results even close to traditional levels of significance. Meta-analytic combining of these results, however, showed $p$ noticeably smaller and $r$ dramatically large.

Intimacy with the Data

The process of summarizing a research domain in a quantitative fashion forces a meta-analyst to be complete in finding all the research articles in the literature, and to be precise in extracting the necessary data from them. A meta-analyst cannot read just the abstracts and discussion sections of articles, as interesting and even as accurate as they might be, to obtain what is needed to summarize a research realm. Whereas in a narrative review, one might accumulate conclusions, in doing meta-analysis a researcher must accumulate data by gathering research articles, scrutinizing their methods for inclusion/exclusion criteria and comparison to other studies, and attending carefully to measures and operationalizations of the independent and dependent variables. To extract the information needed to calculate effect sizes, a meta-analyst must become quite familiar with precisely what any given study actually found. These findings might be in the form of means and standard deviations, or tables of counts (that sometimes need to be constructed from textual material), or test statistics such as $t$, $F$, $\chi^2$, or $Z$; or their associated $p$ levels. Thus, meta-analysts are forced to develop a certain intimacy with existing published (and often unpublished) data in a research area. Reading a research paper is quite a thorough enterprise when conducting a meta-analysis.
Focused Research Hypotheses

Conducting a meta-analysis is also an exercise in research precision. Diffuse hypotheses tested with more than one degree of freedom in the numerator of an $F$ test or in a $\chi^2$, although common in the literature, are problematic theoretically as well as statistically. Suppose, for example, a study is done in the area of patient adherence to (or compliance with) antibiotic treatment. The researcher wishes to determine the effect on a patient health outcome (say, eradication of the infection) of “perfect” versus “less than perfect” compliance with a doctor’s recommendation to take four pills a day (referred to as QID, or roughly every six hours). Suppose further that the researcher randomly divides patients into four treatment groups, which respectively actually take one, two, three, or four pills a day. The outcome, appropriately, is the extent to which the patients get better (i.e., are cured of their infections). An analysis of variance “omnibus $F$-test” with 3 $df$ in the numerator might show no significant differences among the four groups, and it could be erroneously concluded that the number of pills a patient actually takes of a QID antibiotic regimen does not matter at all in affecting whether or not an infection clears up. Besides making little intuitive sense, this conclusion can be massively incorrect statistically, and could lead to some serious erroneous clinical choices. A focused research question is essential. One or more contrasts from these three possibilities might be chosen: (a) a contrast between the group taking the prescribed number of pills each day (four) and the average of the other three groups (those taking three, two, or one); (b) a linear contrast, looking at the outcome as a function of increasing numbers of pills taken per day; or (c) a contrast of the average of the groups taking three or four pills per day versus the average of the groups taking one or two pills a day. In all of these cases, the test is a one $df$ $F$-test, or a $t$-test, and it focuses on the real research questions: Is there a better result if one complies with the doctors’ recommended treatment than if one does not? The question that is asked needs to be as scientifically precise as possible, especially given the potentially serious implications of failing to follow treatment advice. Because meta-analysis demands focused one-degree-of-freedom contrasts, it trains researchers to be attentive to the precise formulation of the questions they ask and meticulous in the answers they extract (Rosenthal & Rosnow 1985, Rosenthal et al 2000).

Identifying Moderator Variables

In virtually every research area, there is bound to be variation in the effect sizes discovered in a meta-analytic review. Sometimes there is wide variation, and the questions that are asked seem to be inadequate to deal with it.

For example, meta-analysis of the effects of a particular drug on depression might yield varying estimates, some strongly positive, some close to zero, and even some that are strongly negative. The issue of concern might then appropriately shift, such that the pattern of findings can be examined in relationship to
moderating variables of interest. Perhaps, for example, the drug works well for middle adult patients, but negative effect sizes obtain in geriatric populations. Or the drug works well at low doses, but not at higher ones, or in the first few months of the illness but not later, or the effect is moderated by the extent of patient compliance. The search for important patterns in the quantitative reviews is facilitated by an inquisitive approach, emphasizing exploration instead of confirmation—an approach that allows examination and reconciliation of differences among studies. Correlations between moderator variables and effect sizes sometimes point to associations that are very helpful to understand. Meta-analysis allows us to formulate potential causal influences and to try to understand why various results occurred. Examination of moderator variables (e.g. year of publication, race/ethnicity of subject group, and sex of researcher) adds to theory development and increases the richness of empirical work.

CRITICISMS OF META-ANALYSIS

Bias in Sampling the Findings

Every meta-analysis has some inherent bias by virtue of the inclusion/exclusion criteria and the methods chosen to review the literature. Not every computer-assisted search will be complete, and not every journal article identified. Ideally one would obtain every piece of data ever collected on the topic of concern, but some data are not published, particularly if they yield results that do not achieve statistical significance. These limitations apply, of course, to qualitative and narrative, as well as to quantitative reviews, and concerns about publication bias in favor of significant results can be addressed with a statistical procedure addressing the file drawer problem, i.e. the problem that significant results are published while nonsignificant results are relegated to file drawers (Rosenthal 1979). Other biases are not so straightforward. For example, some researchers provide enough information to compute an effect size, whereas others do not. This difference may reflect a more serious bias in research sophistication.

Garbage In and Garbage Out

A meta-analysis usually includes studies that vary considerably in their sampling units, methods of measuring and operationalizing independent and dependent variables, data-analytic approaches, and statistical findings. Such variation can increase the generalizability of results when the findings are clear, but when they are not, varying theoretical and methodological approaches and an unsuccessful search for moderators can be confusing and can obscure a full understanding of the story the data are trying to tell. In the midst of this method variation is variation in quality: Meta-analysis is sometimes criticized for mixing together good and bad studies. This criticism, known as the “garbage in and garbage out” issue (Hunt 1997:42),
can be dealt with using a weighting technique that takes into account and quantifies the methodological strength of each study in the analysis. Rosenthal (1991) has argued for “quality weighting” of studies, suggesting four-point scales as practical and valuable, and up to nine points as useful in some circumstances. Further, studies can also be blocked according to their type of methodology (randomized clinical trials versus observational; studies with and without control groups) and type of operationalization of dependent and independent variables (see DiMatteo et al 1996).

Singularity and Nonindependence of Effects

This criticism is easy to deal with if meta-analysts remember that effect sizes that are not independent of one another may need to be combined differently from effect sizes that are independent of each other—i.e. from different studies with nonoverlapping samples. If a study has more than one effect size, these can be used individually in analyses of subgroups or in examination of moderating variables, or they can be combined either by conservative averaging or by using less conservative techniques recommended by Rosenthal & Rubin (1986). Nonindependence may be a problem if the same research lab contributes a number of studies and this fact is ignored. It is possible and often valuable to block by laboratory or researcher and examine this as a moderator variable.

An Overemphasis on Individual Effects

Meta-analysis systematically assesses only individual effects, e.g. differences between means (effect size d) or zero-order correlations (effect size r) between independent and dependent variables—without necessarily viewing the big picture. It is true that meta-analytic techniques systematically assess only individual relationships between independent and dependent variables. In most research domains, however, this simple, systematic approach is essential. Consider the example of research on adherence to medical regimens. Many hundreds of empirical articles have been published on social-psychological variables that correlate with whether or not patients follow the recommendations that their health professionals have given them. A multifactorial model is indeed necessary, but before building that model, the individual correlates of adherence need to be taken into account (DiMatteo 2000). Before examining the combination and interaction of these components, meta-analysis is essential for achieving a clear picture of the straightforward operation of each individual component. Then informed studies using multifactorial, longitudinal designs can be built based on what meta-analysis has told us is important to examine. Whereas there is admittedly some loss of information when one concentrates on single effects in meta-analysis, a singular focus helps to target specific questions and to distill the essential elements of a phenomenon under study.
Combining Apples and Oranges

Another evocative image in the list of criticisms of meta-analysis is the apples and oranges argument (Hunt 1997:61). Meta-analysis is sometimes criticized because it involves summarizing results from studies that vary notably in their operationalization and measurement of independent and dependent variables and that employ very different types of sampling units to achieve answers to questions that are similar, though often not identical. It is argued, therefore, that meta-analysis is analogous to taking apples and oranges and averaging such measures as their weights, sizes, flavors, and shelf lives (Hunt 1997). The figures arrived at might be meaningless. It is true that in all reviews of the literature, qualitative and quantitative, we encounter replications that are rarely precisely the same. It can be argued, however, that it is a good thing to mix apples and oranges, particularly if one wants to generalize about fruit, and that studies that are exactly the same in all respects are actually limited in generalizability. Further, when studies vary methodologically, well-done meta-analyses take these differences into account by treating them as moderator variables. Hall and others (Hall et al 1994) do note, of course, that synthesis of very disparate studies requires sensitivity to issues of inference in trying to aggregate very diverse approaches to sampling and operationalization, and awareness of the existence of interesting and relevant moderator variables.

DOING META-ANALYSIS

As Hall & Rosenthal (1995) have noted, there is no single correct way to perform a meta-analysis. There are certain goals that should be addressed, however, and some methods better serve these goals than others. They offer three interrelated basic principles to guide meta-analysis: accuracy, simplicity, and clarity. The simpler a meta-analysis, the more likely it is to be accurate; it is not possible to present one that is too simple. The best quality scientific exploration is often one that poses unadorned, straightforward questions and uses simple statistical techniques for analysis. Alternatively, it is possible to do a meta-analysis, or any statistical analysis for that matter, that suffers from “high-tech statistication.” Such analyses lend an impressive air of sophistication but may be massively inappropriate. Staying simple and staying close to the data helps to avoid serious misconceptions about it (Rosenthal 1995b). This point is underscored in the report of the APA Science Directorate’s Task Force on Statistical Inference (Wilkinson et al 1999), which notes that scientific inquiry should remain logical and straightforward, understanding clearly the differences between correlation and causality, adopting an exploratory orientation, posing clear and straightforward scientific questions, analyzing them with straightforward and well-understood statistical tests, and avoiding the temptation to stuff data into the computer and hope for a sophisticated answer. Statistical analyses, complicated or not, must be used to aid thought about scientific research, not obscure it with mechanical approaches.
Despite developments in methodological and statistical techniques, meta-analysis remains at heart as simple now as it was in 1904. The level of quantitative skill and training required to do meta-analysis is very modest, and researchers who can analyze their own data can learn (easily) the few rather simple calculations needed to carry out a high-quality meta-analysis.

At this point in the extensive proliferation of research in psychology, health, and health psychology, among other fields, anyone who is considering a review of the literature has little justification for not doing it quantitatively. All the valuable aspects of narrative reviews can be preserved in meta-analysis, and quantitative features can be added. In approaching meta-analysis, researchers would do well to carefully examine the following basic bookshelf: Chalmers & Altman (1995), Cook et al (1992), Cooper (1989), Cooper & Hedges (1994), Glass et al (1981), Hedges & Olkin (1985), Hunt (1997), Hunter & Schmidt (1990), Light & Pillemer (1984), Rosenthal (1991), Wachter & Straf (1990). Some of the more complex procedures are described by Hedges & Olkin (1985) and by Hunter & Schmidt (1990); those that are quantitatively less demanding are by Cooper (1989), Hunt (1997), Light & Pillemer (1984), and Rosenthal (1991).

In the remaining pages of this chapter, we review some of the recent developments in the use of meta-analysis as a set of quantitative procedures.

**Basic Steps**

The basic steps in a meta-analysis follow.

1. Define the independent and dependent variables of interest, e.g. the effects of patient depression on patient adherence to medical treatment (DiMatteo et al 2000).

2. Collect the studies in a systematic way, attempting to find all the published (and often the unpublished) research available. Read each article’s method and results very carefully, assessing how independent and dependent variables were operationalized and measured. Hope the researchers have reported effect sizes and \( ns \), and if they have not, scour the articles for the information necessary to calculate these.

3. Examine the variability among the obtained effect sizes informally with graphs and charts. Most approaches to meta-analysis operationalize heterogeneity as a chi-square test of significance. It must be kept in mind, however, that the significance of this chi-square test depends upon sample size and can yield highly significant results even when there is little variation in the effect sizes; the standard deviation is a straightforward measure of the variability in effect sizes that is not dependent upon sample sizes. Variability among effect sizes points to the likelihood that a moderator variable might account for the variability in the effect sizes, and possibilities should be explored.
4. Combine the effects using several measures of their central tendency, i.e. medians and both weighted and unweighted means. When several approaches to central tendency yield different results, the reasons for such differences need to be explored.

5. Examine the significance level of the indices of central tendency. It is almost always useful to employ confidence intervals around the unweighted mean effect size based on a random effects model (i.e. using studies as the unit of analysis) and it is sometimes useful to employ confidence intervals around the weighted mean effect size based on a fixed effects model. The latter fixed effects model employs subjects nested within studies as the units of analysis, and yields a more powerful test of an overall null hypothesis, a null that is probably always false in any case (Cohen 1994). The disadvantage of the fixed effect model is that it does not permit generalization to studies other than those already in the sample. The random effects approach, though less powerful, does permit generalization to studies not yet in the sample, and if only one approach were to be used it would be the one we prefer. A new statistical procedure called the counternull value of the effect size is often helpful in meta-analytic work as well as in the analysis of individual studies. The counternull gives that value of the effect size that is greater than the one obtained and has exactly the same probability level as does the null value. For example, if we obtain \( r = 0.10 \), not significant at e.g. \( p = 0.05 \), before we decide \( r \) must, therefore, really be 0.00, the counternull tells us that the true value of \( r \) could as easily be 0.20 as it could be 0.00 (Rosenthal & Rubin 1994).

6. Using an examination of the binomial effect size display (see below), evaluate the importance of the obtained effect size.

EFFECT SIZES: Chief Coins of the Meta-Analytic Realm

There are two main families of effect sizes, the \( r \) family and the \( d \) family.

The \( r \) family of product moment correlations includes Pearson \( r \) when both variables are continuous, \( \phi \) when both variables are dichotomous, point biserial \( r \) when one variable is continuous and one is dichotomous, and \( \rho \) when both variables are in ranked form, as well as \( Z_r \), the Fisher transformation of \( r \).

This family also includes the various squared indices of \( r \) and related quantities, such as \( r^2 \), \( \omega^2 \), \( \epsilon^2 \), and \( \eta^2 \). Squared indices are problematic, however, because they lose their directionality (although this can be retrieved through careful analysis of the findings), and the practical magnitude of these indices is often misinterpreted. In an example regarding the latter problem, it may be concluded that one percent of the variance in a dependent variable owing to the independent variable is too little to matter. However, if the independent variable is a very inexpensive and safe intervention, and the dependent variable...
involves saving lives [as was the case in research on prevention of heart attacks with low-dose aspirin (Rosenthal & Rosnow 1991)], the percentage of variance explained may be very small, but its implications might be quite substantial.

The three main members of the \(d\) family of effect sizes are Cohen’s \(d\), Hedges’ \(g\), and Glass’s delta. All three employ the same numerator (comparing the difference between two means). The square root of the pooled variance (\(\sigma^2\)) of the two groups is used as the denominator in \(d\), the square root of the pooled variance (\(S^2\)) is used in \(g\), and the denominator of delta is the square root of the control group variance (\(S^2\)) only. The equations are:

\[
\text{Cohen’s } d = \frac{M_1 - M_2}{\sigma_{\text{pooled}}}
\]

\[
\text{Hedges’s } g = \frac{M_1 - M_2}{S_{\text{pooled}}}
\]

\[
\text{Glass’s } \Delta = \frac{M_1 - M_2}{S_{\text{control group}}}
\]

The Advantages of \(r\)

Studies in the psychological literature vary considerably in their reporting of \(r\) or \(d\) effect sizes, and any review of a substantial amount of the literature is bound to reveal data presented with both types of effect sizes (not to mention no effect sizes at all). Both \(r\) and \(d\) estimates can be readily converted to one another, and eventually meta-analytic researchers need to decide to which index they should convert all effect size estimates obtained. Two examples of such conversions are:

\[
r = \sqrt{\frac{d^2}{d^2 + 4}}
\]

\[
d = \frac{2r}{\sqrt{1 - r^2}}
\]

The effect size \(r\) has several advantages over \(d\). First, converting \(d\)'s to \(r\)'s makes sense because \(r\) in its point biserial form represents the relationship between two levels of the independent variable and scores on the dependent variable, but converting the continuous Pearson \(r\) to the dichotomous \(d\) loses information. Furthermore, using 1 df contrasts (see below), \(r\) allows for the analysis of trends across more than two groups, whereas \(d\) is limited to two. The \(r\) index requires no computational adjustment in going from cases of \(t\)-tests of two or more samples, to \(t\)-tests of only a single sample. Also, \(r\) is more simply interpreted in terms of practical importance than are \(d\) or \(g\) (see below for more about practical importance.)
Getting to “r”

Obtaining an effect size $r$ from a given study may be easy or it may be a daunting challenge, depending upon the information the author has presented. Ideally, a researcher publishes a measure of the relationship between the independent and dependent variable in terms of a Pearson $r$, a point biserial $r$, a Spearman rho, or a phi coefficient, depending upon the nature of the independent and dependent variables as continuous variables, as ranks, or as dichotomous variables. The effect size $r$ can be easily computed from $t$ statistics, and from $F$ statistics with 1 df in the numerator using the following:

\[
r = \sqrt{\frac{t^2}{t^2 + df}}
\]

\[
r = \sqrt{\frac{F}{F + df_{error}}}
\]

Effect size $r$'s can also be computed from chi square and from the standard normal deviate $Z$.

\[
r = \sqrt{\frac{\chi^2(1)}{N}}
\]

\[
r = \frac{Z}{\sqrt{N}}
\]

If an article contains nothing but $p$ values, we can proceed as follows: Convert $p$ to its associated one-tailed standard normal deviate $Z$ and use the equation above. Often, however, a range is given, and the following can be used: For $p < 0.05$, $Z = 1.645$; for $p < 0.01$, $Z = 2.326$; and for $p < 0.001$, $Z = 3.090$. Sometimes the meta-analyst must search a paper to find out what happened in the data and may turn up nothing more specific than the following: “The independent variable had no significant effect on the dependent variable.” In this case the meta-analyst is forced to assign a $Z$ of zero, with a corresponding $r$ of zero, an approach that usually represents a loss of information and an underestimate of the size of the effect. The standardized beta from a multiple regression, as well as a partial correlation, can be used as effect size estimates, but it must be remembered that these represent the relationship between the independent and the dependent variable controlling for other factors (and the meta-analyist might want separately to combine $r$'s and partial $r$'s/standardized betas). It should be noted that an $r$ effect size cannot be computed from kappa, percent agreement, relative risk, risk difference, or the odds ratio unless all the raw data are available, so the meta-analyst can compute the proper index. Also, when meta-analyzing reliabilities, one should always report whether they have been corrected using the Spearman-Brown equations (Rosenthal 1987). Raters’ reliabilities should be reported Spearman-Brown “upped” (the reliability of the set of $k$ raters) as well as Spearman-Brown “downed” (the reliability of a single rater).
The Four $r$’s

In our discussion of the effect size estimate $r$ we have, up until now, been referring to a specific $r$ called $r_{\text{contrast}}$. It is the most generally useful $r$ in meta-analytic work, but it is not the only $r$. Indeed, there is a set of four effect size correlations, all of which provide quite different information, and all of which should ideally be computed in meta-analytic work. The details are given elsewhere (Rosenthal et al. 2000) and here we give only an overview.

1. $r_{\text{alerting}}$: the correlation between the means and their contrast coefficients or weights\(^1\) (which may be constructed from authors’ written hypotheses in the text or even from the meta-analyst’s own). $r_{\text{alerting}}$ ignores within group noise.

2. $r_{\text{effect size}}$: the correlation between an individual’s score on the dependent variable and the contrast weight assigned to the condition to which the individual belongs.

3. $r_{\text{contrast}}$: a special case of $r_{\text{effect size}}$ in which noncontrast between group variation is partialled out. $r_{\text{contrast}}$ tends to be larger than $r_{\text{effect size}}$ because the variation associated with other between groups effects is removed from the error term.

4. $r_{\text{BESD}}$: a more conservative effect size $r$, but one that permits generalization not only to other subjects in the same conditions but also to other levels of the same independent variable.

The four $r$’s can be computed as follows:

\[
\begin{align*}
 r_{\text{alerting}} &= \sqrt{\frac{F_{\text{contrast}}}{F_{\text{contrast}} + F_{\text{noncontrast}}(df_{\text{noncontrast}})}} \\
 r_{\text{effect size}} &= \sqrt{\frac{F_{\text{contrast}}}{F_{\text{contrast}} + F_{\text{noncontrast}}(df_{\text{noncontrast}}) + df_{\text{within}}}} \\
 r_{\text{contrast}} &= \sqrt{\frac{F_{\text{contrast}}}{F_{\text{contrast}} + df_{\text{within}}}} \\
 r_{\text{BESD}}^2 &= \sqrt{\frac{F_{\text{contrast}}}{F + F_{\text{noncontrast}}(df_{\text{noncontrast}} + df_{\text{within}})}}
\end{align*}
\]

\(^1\) Contrast weights (or $\lambda$s) are the predicted results of a study (using, e.g., predictions ranging from 0 to 100) with the restriction that the sum of the weights $= 0$. Subtracting the mean of the predicted values from each prediction achieves this requirement.

\(^2\) When $F_{\text{noncontrast}}$ is less than 1.00, it is entered here as equal to 1.00. $F_{\text{noncontrast}}$ is computed as $F_{\text{between}}(df_{\text{between}} - 1) - F_{\text{contrast}}$. 


Contrasts with Categorical Data

If categorical data are presented in a simple $2 \times 2$ table, the $\phi$ coefficient provides a useful $r$ effect size index; if there are more than two rows and/or columns, however, the following equation can be used for a contrast:

$$Z = \frac{\sum \lambda P}{\sqrt{\sum S_{P}^{2} \lambda^{2}}}$$

where $P$ is the proportion of each column meeting some criterion (e.g. above average performance), $\lambda$ is the contrast weight, and, $S_{P}^{2} = \frac{P(1-P)}{N}$, where $N$ is the total column count (Rosenthal & Rosnow 1985, 1991).

Combining $r$’s

Combining $r$ effect sizes is a straightforward enterprise. First, each $r$ is transformed into the Fisher $Z$ transformation of $r$ in order to normalize the distribution. The unweighted mean of these Fisher $Z$ transformed $r$’s may then be calculated, as well as the mean of the Fisher $Z$ transformed $r$’s weighted by the $N$-3 of each study. These weighted and unweighted average Fisher $Z$ transformed $r$’s are then converted back to $r$, and the weighted and unweighted mean $r$’s reported. Confidence intervals around these estimates reveal the degree to which they differ significantly from zero (i.e. do not cross zero). For the unweighted mean $r$, the random effects confidence interval is usually preferred, yielding wider confidence intervals but allowing generalization to studies other than those in the collected sample. The equation for the 95% confidence interval around the unweighted mean employs the $Z_{r}$ transformation of the correlations:

$$Z_{r} \pm t_{(0.05)}S/\sqrt{k},$$

where $Z_{r}$ is the unweighted mean of the $Z$-transformed $r$’s, $t_{(0.05)}$ is the value of $t$ required for the two-tailed $p$ value of 0.05 for $k-1$ df, $k$ is the number of studies yielding $Z_{r}$s, and $S$ is the standard deviation of the $k Z_{r}$s.

Dealing with Heterogeneity

As noted earlier, it is important to examine the variability of effect sizes. The least useful and least appropriate way is simply to compute a significance test of heterogeneity and give up on combining the effects if the test is significant. This is inappropriate because the significance level depends so heavily on the size of the studies being examined. It is more valuable to examine the standard deviation of the effect sizes, plot them, look for outliers and naturally occurring groupings, and focus on finding blocking variables or moderators that explain the data. The simplest way to examine moderators is to compare average effect sizes in the different subgroups that form the levels of the moderator. Examining moderators in meta-analyses allows for further testing of details of theory, and a better understanding of the research literature. It is not necessary to show that a sample of
effect sizes is significantly heterogeneous in order to look for moderators (Hall & Rosenthal 1991). Just as in analysis of variance, an overall $F$ with more than 1 df in the numerator can be nonsignificant, whereas a planned contrast can be highly significant; a distribution of effect sizes that is not significantly heterogeneous can contain one or more contrasts that are both substantial in magnitude and statistically significant. Effect sizes and significance levels can be readily computed for moderator variables by means of the following equation:

$$Z = \frac{\sum(Z_r \lambda)}{\sqrt{\sum (\frac{\lambda^2}{N-3})}},$$

where $Z_r$ is the $Z$ transformed effect size $r$, $\lambda$ is the contrast weight associated with each of the $k$ studies, and $N$ is the number of subjects or other sampling units on which each $Z_r$ is based.

Of course, in practice it is sometimes the case that no moderator can be found in a group of highly variable effects. For example, in a meta-analysis by DiMatteo et al (2000), across 13 studies, anxiety had an exceptionally variable relationship to patient adherence to treatment: The effect sizes range from $-0.64$ to $0.39$ with great diversity in between ($S = 0.08$, $\chi^2 (12) = 27.58$, $p = 0.0063$). Although the average of these effects was close to zero, it would not have been useful to state that there was no effect of anxiety on patient adherence; one summary statistic simply does not do justice to the apparent complexity of the literature. The authors were, however, unable to find any moderator to account for this variation. This difficulty is partly conceptual. Anxiety itself can be quite heterogeneous and range from panic, which may have no direct effect on adherence, to obsessive-compulsive disorder and generalized anxiety about health, which may improve compliance activities, to anxiety with a depressive overlay, which may reduce adherence considerably (Mineka et al 1998). There were simply not enough studies available yet that had assessed these moderator variables to conduct a thorough analysis; however, future studies may permit the emergence of trends.

**Where to Go From $r$?: Graphics and the Binomial Effect Size Display**

Light et al (1994) present a number of useful methods for displaying data graphically. In one method, the sample size is plotted on the $X$ axis and the effect size on the $Y$ axis, with the data points being those from the studies included in the meta-analysis. In another, odds ratios are graphed in increasing magnitudes along with their 95% confidence intervals. Stem and leaf displays and schematic plots from Tukey’s (1977) exploratory data analysis are recommended as well, and are particularly useful to display the effect of moderator variables in side-by-side schematic plots. In another graphical display, effect size is plotted on the $Y$ axis and levels of a monotonically increasing independent variable are plotted on the $X$ axis. Side-by-side scatterplots for several different dependent variables can
also be used. Whatever graphical displays are used, meta-analysis should clearly convey the distribution of the effect sizes in some informative way.

The binomial effect size display (BESD) is a useful and informative technique for examining the practical importance of any effect indexed by \( r \). The correlation coefficient is shown to be the simple difference in outcome rates of two groups (e.g. experimental/control, female/male) in a standard table with column and row totals of 100 each (Rosenthal & Rubin 1982). We obtain the BESD from any obtained effect size \( r \) by computing the treatment condition success rate as 0.50 plus \( r/2 \), and the control condition success rate as 0.50 minus \( r/2 \). Thus, an \( r \) of 0.20 yields treatment success rates and a BESD as follows (with cell entries multiplied by 100):

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Live</th>
<th>Die</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Variable</td>
<td>Treatment</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>Σ</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

The correlation \( r \) of 0.20 is simply the difference between the success rates of the experimental versus the control group (0.60 − 0.40). In general, BESD involves a \( 2 \times 2 \) table with the cell counts labeled A, B, C, D as follows, and with column and row totals of 100:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Live</th>
<th>Die</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Variable</td>
<td>Treatment</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Control</td>
<td>C</td>
<td>D</td>
<td>C+D</td>
</tr>
<tr>
<td>Σ</td>
<td>A+C</td>
<td>B+D</td>
<td></td>
</tr>
</tbody>
</table>

There are three other indices of effect size that are often useful (and commonplace in biomedical contexts). These are relative risk, odds ratio, and risk difference. In an effort to better capture the implications of the data for the population as a whole, it has been suggested that before computing relative risks,
odds ratios, and risk differences, we compute \( r \), display it as a BESD, and then compute “standardized” relative risks, odds ratios, and risk differences (Rosenthal et al 2000). Relative risk is defined as the ratio of the proportion of control patients at risk, divided by the proportion of treated patients at risk for the bad outcome. Relative risk \( = [A/(A + B)]/[C/(C + D)] \); in the BESD where \( A = D \) and \( B = C \), standardized relative risk \( = A/C \). Suppose a treatment is so important that a possible outcome of noncompliance is death; in fact, renal dialysis is such an example. Suppose that the correlation between noncompliance and death is 0.60, so that the BESD entries are

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Compliant</th>
<th>Noncompliant</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live</td>
<td>80</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>Die</td>
<td>20</td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>Σ</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

Then, the relative risk of death if one is noncompliant is \( 80/20 = 4 \); there is a 4 times greater likelihood that one will die if noncompliant than if compliant. The odds ratio is an index that is very common in biomedical research. It is defined as the ratio of bad outcomes to good outcomes in the control (noncompliant) group, divided by the ratio of bad outcomes to good outcomes in the treated (compliant) group. In the case above, this would be the ratio of the odds of dying if noncompliant (ratio of dying to living, if noncompliant, \( 80/20 \)) to the odds of dying if compliant (ratio of dying to living, if compliant, \( 20/80 \)). The odds ratio, then, is \( 80/20 \) divided by \( 20/80 \), which is 16. Thus, the odds of dying if noncompliant are 16 times greater than the odds of dying if compliant.

A third index is the risk difference, and that is simply the difference (0.60) between the proportion of control or untreated (noncompliant) patients who have a poor outcome (80/100) and the proportion of treated (compliant) patients who have a poor outcome (20/100). The standardized risk difference is equal to the value of \( r \).

**THE EFFECT SIZE \( r \) IN THE LARGER CONTEXT**

**Interpreting the Size of \( r \)**

Is \( r \) ever too small to matter? Answering this question is an important exercise in scientific inquiry and in developing an understanding of a body of research in
a broad context. The importance of an effect size is determined both statistically and theoretically within a given field, and in comparison across fields. Consider, for example, the aspirin trial in the Physicians’ Health Study (Steering Comm. Physicians’ Health Study Res. Group 1988).

Here, the effect size of taking low-dose aspirin in preventing a heart attack was $r = 0.034$ and $r^2 = 0.0012$, indicating that less than 1/8 of 1% of the variance in heart attack was accounted for by using aspirin. Under some circumstances, an effect size $r$ of 0.034 might be seen as unimportant, but examination of the BESD, and a clear understanding of the issues at hand, presents a different picture. According to the BESD, with an effect size $r$ of 0.034 and among persons similar to those represented in the study with comparable risk factors, 34 out of every 1000 would be saved from a heart attack if they used low dose aspirin on a regular basis. Given the ease, safety, and low cost of low dose aspirin therapy, and the high prevalence as well as high cost and potential threat to life and well-being of heart attacks, this finding is, in fact, very important and translates into substantial reductions in morbidity and mortality.

Elsewhere, Rosenthal has compared known effect sizes (Rosenthal 1995a). For example, Smith et al (1980) report the average effect of psychotherapy on improvement to be equivalent to an $r$ of 0.39. Table 1 presents some additional average $r$ effect sizes from 24 meta-analyses in the fields of medicine, behavioral medicine, organizational psychology, and social psychology based on from 5 to 76 studies each. Very few are larger than the effects of psychotherapy.

CONCLUSION

In this chapter, we have explored some of the issues of concern regarding the advantages and criticisms of meta-analysis, and reviewed some of the more recent developments in the statistical procedures of meta-analysis. In concluding, we want to call attention to the role of meta-analyses in drawing causal inferences. If the meta-analysis is based on randomized experiments, strong causal inferences are often warranted. If the meta-analysis is based on observational studies, causal inferences are as risky as they are in the case of individual observational studies. If the meta-analysis includes some randomized experiments and some observational studies, we can meta-analyze them separately and combine their results if they are quite similar, borrowing strength for the randomized experiments from the similar results of the nonrandomized studies. Finally, results of moderator analyses rarely permit strong causal inferences but often suggest fresh studies permitting such inferences.

EPILOGUE

Good researchers who can analyze their own data can conduct meta-analyses. Meta-analysis is not inherently different from primary data analysis; it requires the same basic tools, thought processes, and cautions. Meta-analytic procedures are straightforward enough to carry out with a statistical calculator, with extensive
### TABLE 1  Average effect size estimates from 24 meta-analyses conducted in three domains of research

<table>
<thead>
<tr>
<th>Reference</th>
<th>Effects of:</th>
<th>Effects on:</th>
<th>N of studies</th>
<th>r effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devine, 1996</td>
<td>Psycho-educational care</td>
<td>Adult asthmatics’ adherence to treatment</td>
<td>7</td>
<td>0.36</td>
</tr>
<tr>
<td>Devine &amp; Reifschneider, 1995</td>
<td>Psycho-educational care</td>
<td>Adult hypertensives’ medication compliance</td>
<td>17</td>
<td>0.34</td>
</tr>
<tr>
<td>Devine &amp; Pearcy, 1996</td>
<td>Psycho-educational care</td>
<td>Functioning of adult patients with chronic obstructive pulmonary disease (COPD)</td>
<td>8</td>
<td>0.30</td>
</tr>
<tr>
<td>Devine &amp; Reifschneider, 1995</td>
<td>Psycho-educational care</td>
<td>Adult hypertensives’ blood pressure</td>
<td>76</td>
<td>0.28</td>
</tr>
<tr>
<td>Brown, 1990</td>
<td>Educational interventions for diabetics</td>
<td>Dietary compliance</td>
<td>15</td>
<td>0.27</td>
</tr>
<tr>
<td>Devine, 1996</td>
<td>Psycho-educational care</td>
<td>Adult asthmatics’ reduction in asthma attacks</td>
<td>11</td>
<td>0.27</td>
</tr>
<tr>
<td>Devine &amp; Pearcy, 1996</td>
<td>Psycho-educational care</td>
<td>VO₂ (volume oxygen) levels in adult COPD patients</td>
<td>5</td>
<td>0.27</td>
</tr>
<tr>
<td>Mullen, et al. 1992</td>
<td>Cardiac patient education</td>
<td>Blood pressure</td>
<td>5</td>
<td>0.25</td>
</tr>
<tr>
<td>DiFabio, 1995</td>
<td>Comprehensive rehabilitation program and back school</td>
<td>Glycosolated hemoglobin</td>
<td>27</td>
<td>0.20</td>
</tr>
<tr>
<td>Mullen, et al. 1992</td>
<td>Cardiac patient education</td>
<td>Efficacy in pain reduction, increased spinal mobility, increased strength</td>
<td>19</td>
<td>0.14</td>
</tr>
<tr>
<td>Mullen, et al. 1992</td>
<td>Cardiac patient education</td>
<td>Diet</td>
<td>9</td>
<td>0.09</td>
</tr>
<tr>
<td>Mullen, et al. 1992</td>
<td>Cardiac patient education</td>
<td>Exercise</td>
<td>12</td>
<td>0.09</td>
</tr>
<tr>
<td>Durlak, et al. 1991</td>
<td>Cognitive-behavior therapy for dysfunctional children</td>
<td>Expert rating</td>
<td>8</td>
<td>0.44</td>
</tr>
<tr>
<td>Shoham-Salomon &amp; Rosenthal, 1987</td>
<td>Paradoxical intervention in psychotherapy</td>
<td>Psychotherapy outcome</td>
<td>12</td>
<td>0.42</td>
</tr>
<tr>
<td>Ambady &amp; Rosenthal, 1992</td>
<td>Ratings of brief segments of nonverbal behavior</td>
<td>External, objective behavioral criterion</td>
<td>38</td>
<td>0.39</td>
</tr>
<tr>
<td>Eagly, et al. 1991</td>
<td>Physical attractiveness</td>
<td>Attributions of social competence</td>
<td>35</td>
<td>0.32</td>
</tr>
<tr>
<td>Durlak, et al. 1991</td>
<td>Cognitive-behavior therapy for dysfunctional children</td>
<td>Behavioral observation</td>
<td>58</td>
<td>0.27</td>
</tr>
<tr>
<td>Eagly, et al. 1991</td>
<td>Physical attractiveness</td>
<td>Attributions of intellectual competence</td>
<td>38</td>
<td>0.22</td>
</tr>
<tr>
<td>Harris &amp; Schaubroeck, 1988</td>
<td>Source of job performance ratings (peer, supervisor, self)</td>
<td>Correlation between sources of assessment: peer-supervisor</td>
<td>23</td>
<td>0.62</td>
</tr>
<tr>
<td>Conway &amp; Huffcutt, 1997</td>
<td>Source of job performance ratings (subordinate, supervisor, peers)</td>
<td>Mean reliability of job performance ratings: supervisors</td>
<td>69</td>
<td>0.50</td>
</tr>
<tr>
<td>Conway &amp; Huffcutt, 1997</td>
<td>Source of job performance ratings (subordinate, supervisor, peers)</td>
<td>Correlation between sources of assessment: self-peer</td>
<td>26</td>
<td>0.37</td>
</tr>
<tr>
<td>Harris &amp; Schaubroeck, 1988</td>
<td>Source of job performance ratings (peer, supervisor, self)</td>
<td>Mean reliability of job performance ratings: peers</td>
<td>11</td>
<td>0.36</td>
</tr>
<tr>
<td>Harris &amp; Schaubroeck, 1988</td>
<td>Source of job performance ratings (peer, supervisor, self)</td>
<td>Correlation between sources of assessment: self-supervisor</td>
<td>36</td>
<td>0.35</td>
</tr>
<tr>
<td>Conway &amp; Huffcutt, 1997</td>
<td>Source of job performance ratings (subordinate, supervisor, peers)</td>
<td>Mean reliability of job performance ratings: subordinates</td>
<td>28</td>
<td>0.30</td>
</tr>
</tbody>
</table>
tables of $Z$ and $t$. A desire to impose order on chaos may be helpful given the challenges of certain research fields, and patience with the limitations of some research write-ups will certainly help. Intimate communing with the data, particularly within the context of the theory guiding the work, is an essential and rewarding requirement.

Visit the Annual Reviews home page at www.AnnualReviews.org

LITERATURE CITED
