chapter 1

Introduction

1.1 Narrative (Qualitative) and Meta-Analytic (Quantitative) Literature Reviews

> *Science is built up with fact, as a house is with stone. But a collection of fact is no more a science than a heap of stones is a house.*
>
> — Jules Henri Poincare (cited in Olkin, 1990)

> *. . . it is necessary, while formulating the problems of which in our advance we are able to find the solutions, to call into council the views of those of our predecessors who have declared an opinion on the subject, in order that we may profit by whatever is sound in their suggestions and avoid their errors.*
>
> — Aristotle, *De Anima*, Book 1, Chapter 2
> (cited in Cooper & Hedges, 1994)

All scientists acknowledge that their efforts should build upon past work through replication, integration, extension, or reconceptualization. It is, therefore, ironic that the traditional review of scientific data has typically been conducted in an unscientific fashion. In the traditional narrative (qualitative) review, the reviewer uses "mental algebra" to combine the findings from a collection of studies and describes the results verbally. Statisticians were the first scientists to advocate alternative methods for combining research findings. These methods were labeled *meta-analysis* by Gene Glass (1976):

> *Meta-analysis refers to the analysis of analyses . . . the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding literature (p. 3).*

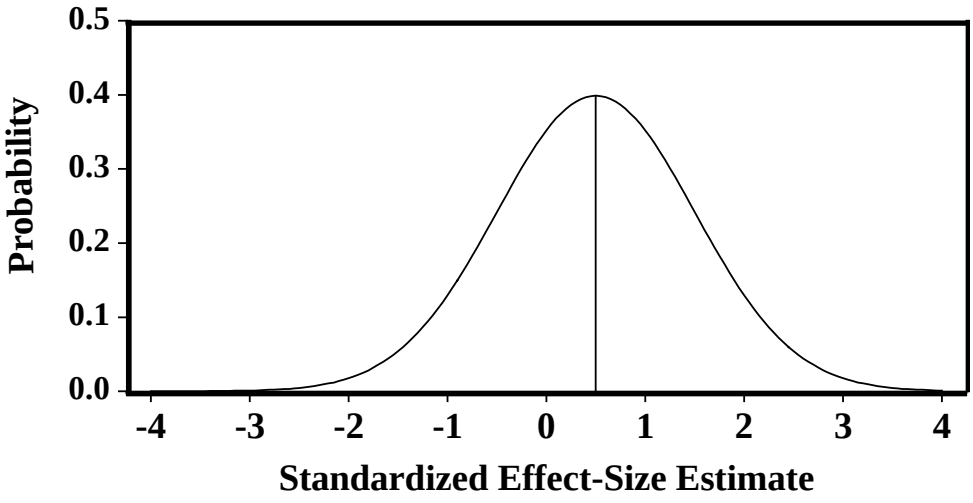
The quantification of research evidence is the key factor that distinguishes a meta-analytic review from a narrative review (Olkin, 1990). In the meta-analytic review, the meta-analyst uses statistical procedures to integrate the findings from a collection of studies and describes the results using numerical effect-size estimates.

One weakness of narrative reviews is that they may be more susceptible to the subjective judgments, preferences, and biases of a particular reviewer's perspective than meta-analytic reviews. As Glass (1976) states:

> *A common method for integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies - those remaining frequently being one's own work or that of one's students or friends - and then advance the one or two "acceptable" studies as the truth of the matter (p. 4)*

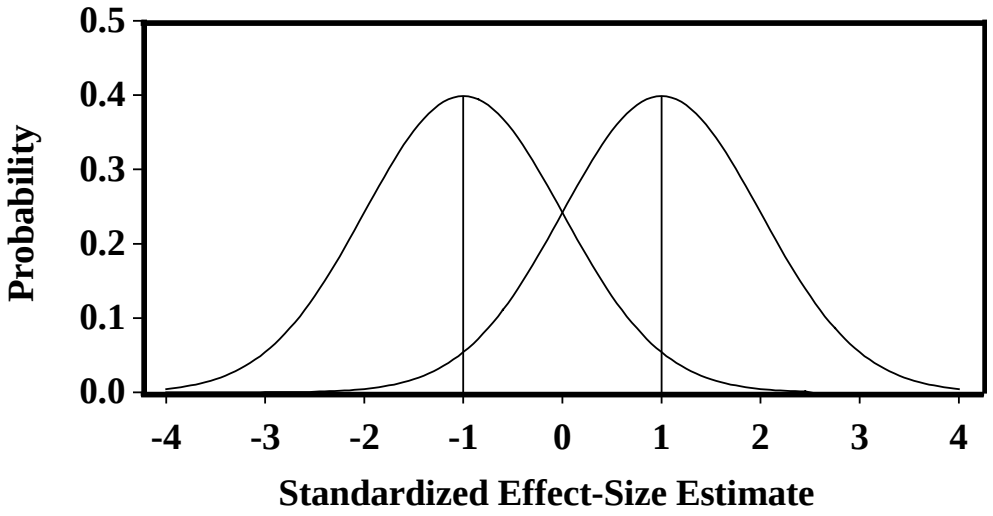
It is worth noting that inconsistent findings in a meta-analytic review are not necessarily problematic. Inconsistent findings may simply reflect opposite tails of the same distribution of effects. Consider, for example, the following distribution of standardized effect-size estimates (that is, effect-size estimate divided by its corresponding estimated standard deviation) that is centered at 0.50 (Cohen's, 1988, conventional value for a medium-sized effect). By random chance some studies (about 31%) should have negative effects even if the true effect-size in the population is 0.50.

Figure 1.1 *Distribution of standardized effect-size estimates centered at 0.50*



Alternatively, inconsistent findings may imply that some variable moderates the treatment effect. A moderator variable influences the strength and/or direction of the relation between the independent variable (that is, the treatment) and the dependent variable (that is, the response; see Baron and Kenny, 1986, for a discussion of moderator variables). In Figure 1.2, the treatment has a negative effect on Group 1 and a positive effect on Group 2. In this example, most negative effects would be found for Group 1, and most positive effects would be found for Group 2. If group is ignored, however, you might conclude that the findings are inconsistent and that the treatment has no effect.

Figure 1.2 *Distribution of standardized effect-size estimates for two different groups that are affected in opposite ways by the treatment*



A second weakness of narrative reviews is that they often ignore the magnitude of the treatment effect. In a narrative review, the reviewer frequently uses *p*-values to draw conclusions by counting the number of studies that found significant treatment effects. But *p*-values cannot be used to determine the magnitude of a treatment effect. Consider the following example in which a treatment group is

compared to a control group. Assume that the experimental and control groups have equal sample sizes. Which treatment effect is largest: (a), (b), or (c)?

(a) $t(256) = 4.0$, $p < .0001$

(b) $t(64) = 2.0$, $p < .05$

(c) $t(4) = 0.5$, $p < .64$

You may be tempted to answer (a) because it has a smaller *p*-value, but this is actually a "trick question." It turns out that the treatment effects are identical for all three tests — the effect-size estimate is 0.50 in each case. A formula for obtaining an effect-size estimate from an independent sample *t*-test is

$$d = \frac{2t}{\sqrt{df}} \tag{1.1}$$

where *d* is the effect-size estimate and *df* are the degrees of freedom (Friedman, 1968). Plugging the values for options (a), (b), and (c) into Equation 1.1, you obtain:

$$d = \frac{2(4.0)}{\sqrt{256}} = \frac{2(2.0)}{\sqrt{64}} = \frac{2(0.5)}{\sqrt{4}} = 0.50.$$

The point is that *p*-values cannot be used as surrogate effect-size estimates.

These weaknesses of narrative reviews can cause their conclusions to be inconsistent with the data. In a study by Cooper and Rosenthal (1980), faculty members and upper-level graduate students in psychology were randomly assigned to use narrative or statistical procedures to review seven studies on sex differences in persistence. None of the reviewers were familiar with meta-analytic techniques. Participants in the statistical group were instructed how to combine the results from the studies. Participants in the narrative group were asked to "employ whatever criteria you would use if this exercise were being undertaken for a class term paper or a manuscript for publication." Participants were asked whether the evidence supported the conclusion that females were more persistent on tasks than males were. Five possible responses were provided (*definitely yes*, *probably yes*,

impossible to say, *probably no*, and *definitely no*). The results showed that 68% of the statistical reviewers were at least considering rejecting the null hypothesis, compared with only 27% of the traditional reviewers. (The null hypothesis should have been rejected at the .05 level because the confidence interval for the effect size excluded the value zero.) Participants also were asked to estimate the magnitude of sex differences in persistence. Six possible responses were provided (*very large*, *large*, *moderate*, *small*, *very small*, and *none at all*). The results showed that 58% of the statistical reviewers estimated at least a small sex difference in persistence, compared with only 27% of the traditional reviewers. (The effect was about equal to Cohen's, 1988, conventional value for a "small" effect.) Thus, participants in the narrative group underestimated the presence and the strength of sex differences in persistence.

In the world outside of the controlled laboratory setting, similar results have been reported. For example, an article in *Science* (Mann, 1994) compares the conclusions drawn from meta-analytic versus traditional literature reviews in five subject areas: (a) psychotherapy, (b) delinquency prevention, (c) school funding, (d) job training, and (e) reducing anxiety in surgical patients. The comparison reveals that narrative reviews underestimate the presence and the strength of treatment effects for each subject area. More recently, Hunt (1997) provided several examples of how narrative reviews underestimate the presence and magnitude of treatment effects. Because of their superiority over narrative reviews, it appears that meta-analytic reviews are here to stay. The next section documents the increasing use of meta-analysis.

1.2 Increasing Use of Meta-Analysis

The use of meta-analysis has increased dramatically in recent years, especially in the social sciences, medicine, and education. For example, we tabulated the number of journal articles in *PsycLit* (a psychological research database) and

Medline (a medical research database) that used the keyword “meta-analysis” from 1976 (the year the term was introduced) to 1995. In both databases, the number of entries has increased rapidly and consistently (see Figures 1.3 and 1.4).

Figure 1.3 *Increase in the use of meta-analysis over time in psychology*

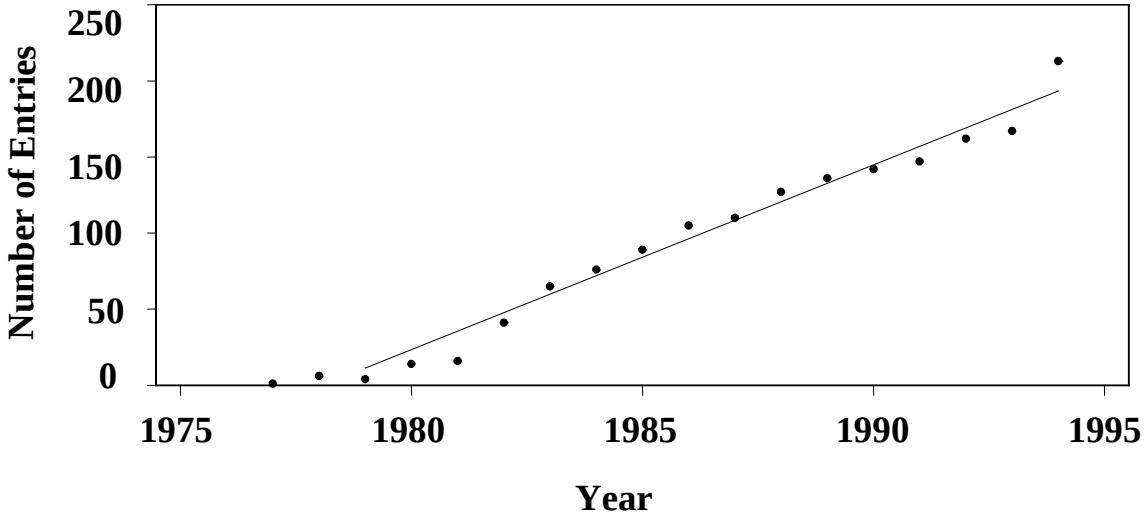
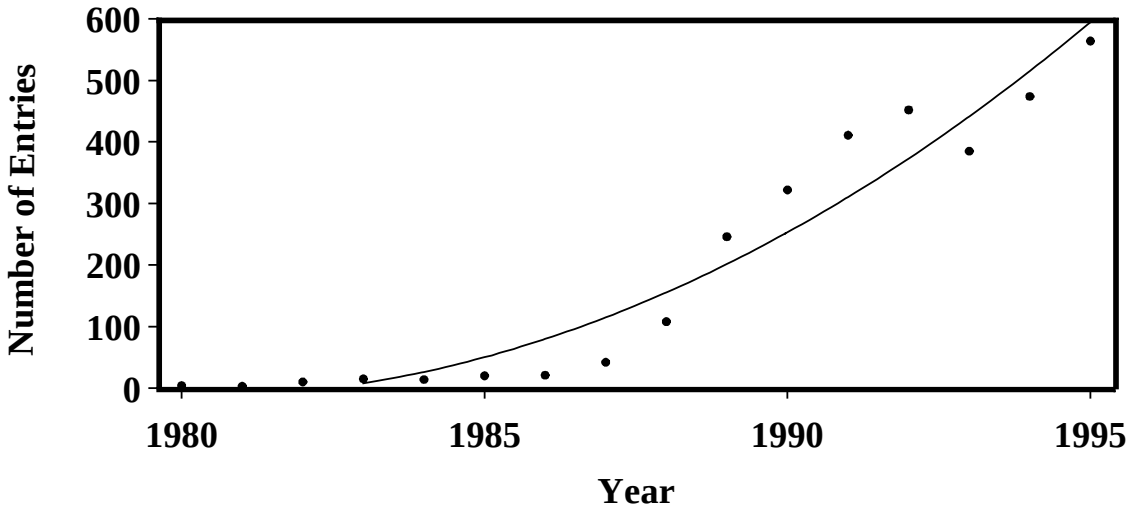


Figure 1.4 *Increase in the use of meta-analysis over time in medicine*



The rapid increase in the use of meta-analysis is likely to continue. In his review of meta-analytic methods, Bangert-Drowns (1986) states the following:

> *Meta-analysis is not a fad. It is rooted in the fundamental values of the scientific enterprise: replicability, quantification, causal and correlational analysis. Valuable information is needlessly scattered in individual studies. The ability of social scientists to deliver generalizable answers to basic questions of policy is too serious a concern to allow us to treat research integration lightly. The potential benefits of meta-analysis method seem enormous. (p. 398)*

1.3 Two Approaches to Conducting a Meta-Analysis

Although the term meta-analysis was coined relatively recently, statisticians have been using these methods for about 100 years. Two different statistical approaches have been used to combine evidence from primary studies. One approach relies on testing the statistical significance of combined results across studies, and the other approach relies on estimating the magnitude of combined results across studies. Fisher (1932), Pearson (1933), and Tippett (1931) were among the first to propose methods for testing the statistical significance of combined results across studies. Consider, for example, the following quotation from the fourth edition of Sir R. A. Fisher's influential text *Statistical Methods for Research Workers*:

> *When a number of quite independent tests of significance have been made, it sometimes happens that although a few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance. It is sometimes desired, taking account only of these probabilities, and not of the detailed composition of the data from which they were derived, which may be of very different kinds, to obtain a single test of the significance of the aggregate, based on the product of the probabilities individually observed (p. 99).*

An early application of this approach was described by Stouffer and his colleagues (Stouffer, Suchman, DeVinney, Star, & Williams, 1949). In three studies, male soldiers rated how much they wanted their sisters to join the United States Army. The ratings were used to determine male soldiers' attitudes toward female soldiers. Some of the male soldiers had female soldiers in their own camp, and some did not. In all three studies, male soldiers were less likely to want their sisters to join the Army when there were female soldiers at their own camp. Stouffer and his colleagues combined the *p*-values from the three studies to obtain an overall significance test.

Significance tests of combined results are sometimes called omnibus or nonparametric tests because they do not depend on the distribution of data. Omnibus tests depend only on the fact that the *p*-values are uniformly distributed between the values 0 and 1.00 when the null hypothesis is true and the treatment has no effect (see Hedges & Olkin, 1985, p. 2). The primary disadvantage of omnibus tests is that they cannot provide estimates of the magnitude of treatment effects across studies.

Birge (1932), Cochran and Yates (Cochran, 1937, 1943; Yates & Cochran, 1938), and Pearson (1904) were among the first to propose methods for estimating the magnitude of treatment effects across studies. For example, Karl Pearson (1904), the famous biometrician, conducted an empirical review of 11 studies that had tested the effectiveness of a typhoid vaccine. Five studies tested whether the vaccine reduced the incidence of typhoid, and the other six studies tested whether the vaccine reduced mortality among those who had contracted typhoid. Pearson computed average correlations of .23 and .19 for typhoid incidence and mortality, respectively. Pearson concluded that these average correlations were too low to warrant adopting the vaccine for British soldiers: "I think the right conclusion to draw would be not that it was desirable to inoculate the whole army, but that improvement in the serum and method of dosing, with a view to a far higher correlation, should be attempted" (p. 1245).

For at least 50 years, social and statistical scientists have questioned the utility of significance testing in research (for example, Bakan, 1966; Berkson, 1938; Carver, 1978; Cohen, 1994; Falk, 1986; Harris, 1991; Hogben, 1957; Kirk, 1996;

Kupfersmid, 1988; Meehl, 1978; Morrison & Henkel, 1970; Nunnally, 1960; Schmidt, 1996; Shaver, 1993). Even Frank Yates (1951), a colleague and friend of R. A. Fisher, said that Fisher's text *Statistical Methods for Research Workers* "has caused scientific research workers to pay undue attention to the results of significance tests . . . and too little (attention) to the estimates of the magnitude of the effects they are estimating" (p. 32). A common theme emerges from these writings: People often use *p*-values as surrogate effect-size estimates (for example, they incorrectly assume that small *p*-values denote large treatment effects). People often misinterpret a *p*-value as the probability that the null hypothesis is false.

Notwithstanding the attacks social and statistical scientists have waged on significance testing, many people continue to "worship" *p*-values (Schulman, Kupst, & Suran, 1976). In a humorous article, Salsburg (1985) concluded that far too many physicians are adherents of a religion called Statistics. According to Salsburg, adherents of this religion engage in the ritual known as "hunting for *p*-values." If the *p*-value is larger than .05, the practitioner must be prepared to suffer the wrath of the angry gods of Statistics. The deep mysterious symbols of this religion are *ns*, * ($p < .05$), ** ($p < .01$) and (mirabile dictu) *** ($p < .001$). The more *'s, the happier are the gods of Statistics. We think that it is a bad idea to worship *p*-values because any treatment effect, no matter how trivial, can achieve statistical significance at any level if the sample size is large enough.

If you accept the need to formally test the null hypothesis (that is, the hypothesis that the treatment has no effect), there is a preferred alternative to significance testing. It involves estimating the magnitude of the treatment effect, called an effect-size estimate, and placing a confidence interval around this estimate (Hedges, Cooper, & Bushman, 1992; Oakes, 1986). This alternative approach can tell not only whether the null hypothesis should be rejected at a given significance level, but also whether the observed treatment effect is large enough to be considered practically important. This book adopts the approach of estimating effect-size estimates and corresponding confidence intervals.

1.4 Operationally Defining Abstract Concepts in Research

Scientific theories are composed of abstract concepts that are linked together in some logical fashion. To test hypotheses derived from theories, researchers must tie abstract concepts to concrete representations of those concepts by means of operational definitions. An operational definition specifies the operations or techniques used to measure the concept. For example, the concept "hunger" might be defined operationally as "depriving an organism of food for 24 hours." Operational definitions are the translation of an abstract concept into a concrete reality.

In a meta-analysis that investigates the same conceptual variables, researchers often use different operational definitions. For example, consider a meta-analysis on the relation between "alcohol" and "aggression" in humans (Bushman & Cooper, 1990). Even though only experimental studies of male social drinkers were included in this meta-analysis, researchers used widely different operational definitions of the concepts "alcohol" and "aggression." Although the concept "alcohol" seems simple enough to define, it was defined in a number of ways. Researchers used different types of alcohol (for example, absolute alcohol; distilled spirits such as vodka, whiskey, rum, and bourbon; beer; wine), different doses of alcohol, and different concentrations of alcohol. The concept "aggression" also was defined in a number of ways. Some researchers used physical measures of aggression (for example, giving electric shocks or noise blasts to another person, taking money away from another person), whereas other researchers used verbal measures of aggression (for example, directing verbally abusive comments to another person, evaluating another person in a negative manner).

Any single operational definition will not fully reflect the more abstract concept that it represents (Gold, 1984). In a meta-analysis, if you find the same relation between concepts, regardless of the operational definitions used in the individual studies, then your confidence in the relation increases. In fact, you might have more confidence in the findings from a meta-analysis of five studies that used different operational definitions than in the findings from a meta-analysis of 50 studies that used the same operational definitions.

1.5 Categorical (Qualitative) and Continuous (Quantitative) Variables

A variable is a qualitative or quantitative entity that can vary or take on different (at least two) values. The value of the variable is the number or label that describes the person or object of interest. In research, variables are used to represent the abstract concepts being studied. One useful distinction is between categorical and continuous variables (for example, Agresti, 1990). A categorical variable simply records which of several distinct categories or groups a person or object falls into. Some examples of categorical variables include political party affiliation, religious denomination, sex, and psychiatric diagnostic groups (for example, schizophrenia, major depression, generalized anxiety disorder). The numbers that are assigned to categorical variables are used only as labels or names; words or letters would work as well as numbers. For the variable SEX, for instance, you could assign the value 1 to males and the value 2 to females. These values do not imply that females are twice as good as males or that you could calculate the "average sex." With categorical variables, you generally calculate the number or the percent of people in each category. The values of a categorical variable are qualitatively different, whereas the values of a continuous variable are quantitatively different. Some examples of continuous variables include temperature, weight, income, and blood alcohol concentration. Mathematical operations (for example, differences, averages) make sense with continuous variables but not with categorical variables. In the SAS language, categorical variables are called classification (CLASS) variables. Variables not specified in a CLASS statement are assumed to be continuous.

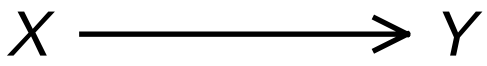
1.5.1 Types of Variables in Research

1.5.1.1 Independent and Dependent Variables

Researchers generally are interested in studying the relations among two or more variables. Suppose that two variables are being studied, a stimulus (X) and a response (Y), and the researcher wants to know whether the stimulus affects the

response. For example, a medical researcher might want to know whether taking aspirin (*X*) reduces the likelihood of a heart attack (*Y*), and a psychological researcher might want to know whether viewing television violence (*X*) increases aggression (*Y*). This relation between variables *X* and *Y* is depicted in Figure 1.5.

Figure 1.5 *Effects of a stimulus (X) on a response (Y)*



If the stimulus (*X*) can be controlled or manipulated by the researcher, it is called the independent variable (treatment or intervention). It is "independent" in the sense that its values are created by the researcher and are not affected by anything else that happens in the study. The corresponding response variable (*Y*) is called the dependent variable (dependent measure or outcome). It is "dependent" in the sense that its values are assumed to depend upon the values of the independent variable.

If the stimulus (*X*) cannot be manipulated by the researcher, it is called a predictor variable. In human participants, individual differences such as sex, age, race, religion, political affiliation, intelligence, ability, personality, risk status (for example, smoker or nonsmoker), and disease status (for example, HIV positive or negative) can be measured but cannot be (ethically) manipulated. The corresponding response variable (*Y*) is called the criterion variable.

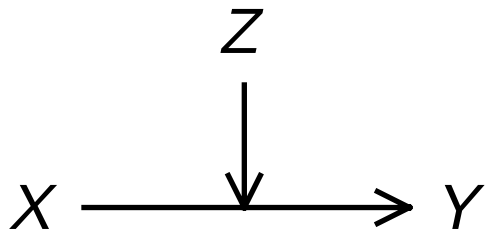
In this book, *X* is called the independent variable or treatment, and *Y* is called the dependent variable or outcome, regardless of whether the researcher manipulated *X*. Although this usage is not technically accurate, it makes for smoother prose and it simplifies discussion considerably.

The relation between variables *X* and *Y* may be influenced by third variables. Two types of third variables, moderator variables and mediator variables, are described respectively in the next sections.

1.5.1.2 Moderator Variables

A moderator variable influences the strength and/or direction of the relation between the independent and dependent variables (Baron & Kenny, 1986). In a study by Stern, McCants, and Pettine (1982), for example, individuals were more likely to become seriously ill if they experienced uncontrollable life events (for example, death of a spouse) than if they experienced controllable life events (for example, being fired from a job). In this example, the type of life event (that is, controllable versus uncontrollable) is the moderator variable. Moderators are typically introduced when there is a weak or inconsistent relation between the independent and dependent variables (Baron & Kenny, 1986). The moderating effects of variable *Z* on the relation between variables *X* and *Y* is depicted in Figure 1.6. In meta-analysis, moderators are any known study characteristics that are associated with differences in effect-size estimates between studies.

Figure 1.6 *Moderating effects of the third variable (Z) on the relation between the stimulus (X) and the response (Y)*



1.5.1.3 Mediator Variables

A mediator variable is the generative mechanism through which the independent variable influences the dependent variable (Baron & Kenny, 1986). Mediator variables are sometimes called intervening variables because they come between the stimulus and the response. Independent variables produce changes in mediator variables that, in turn, produce changes in dependent variables. Berkowitz (1990), for example, proposes that aversive events (for example, provocation, frustration,

hot temperature) increase impulsive aggression because they produce negative affect – an unpleasant emotional response. Berkowitz views negative affect as a possible mediator between aversive events and impulsive aggression. Mediators are typically introduced when there is a strong relation between the independent and dependent variables (Baron & Kenny, 1986). The mediating effect of variable *Z* on the relation between variables *X* and *Y* is depicted in Figure 1.7.

Figure 1.7 *Mediating effects of the third variable (*Z*) on the relation between the stimulus (*X*) and the response (*Y*)*

1.5.2 Effect-Size Measures for Categorical Variables

Suppose that the independent and dependent variables in a study are both dichotomous (that is, both are categorical variables with two levels). For such studies, which are very common in the field of medicine, the odds ratio is the most frequently used effect-size metric. For example, Table 1.1 depicts the results from a large randomized, double-blind, placebo-controlled trial testing whether aspirin reduces mortality from cardiovascular disease (Steering Committee of the Physicians Health Study Group, 1988). The study participants, 22,071 male physicians, took either an aspirin or a placebo every other day. The data from the study at the five-year follow-up are reported here as percentages.

Table 1.1 *Results from a large randomized, double-blind, placebo-controlled trial testing whether aspirin reduces mortality from cardiovascular disease*

	Heart attack	**No heart attack**
Aspirin	0.94%	99.06%
Placebo	1.71%	98.29%

The odds of not having a heart attack in the aspirin group are 99.06 to 0.94 or 105.38 to 1. The odds of not having a heart attack in the placebo group are 98.29 to 1.71 or 57.48 to 1. To compare the aspirin and placebo groups, simply create a ratio of these two odds: 105.38 ÷ 57.48 = 1.83. Thus, physicians in the placebo group are almost twice as likely to have a heart attack as physicians in the aspirin group. An odds ratio of 1.0 means that the aspirin doesn't differ from the placebo in reducing heart attacks. Chapter 4 discusses how to combine odds ratios.

1.5.3 Effect-Size Measures for Continuous Variables

Two measures of effect dominate the meta-analytic literature when the dependent variable is continuous: the standardized mean difference and the Pearson product-moment correlation coefficient. When the primary studies in question compare two groups, either through experimental (treatment) versus control group comparisons or through orthogonal contrasts, the effect-size estimate often is expressed as some form of standardized difference between the group means. For example, suppose that 100 participants in a study are randomly assigned to experimental or control groups. Suppose also that the mean score for the experimental group is higher ($\bar{Y}_E$ = 10) than the mean score for the control group ($\bar{Y}_C$ = 8), but that the variation in scores is about the same for the two groups (pooled standard deviation, S_{POOLED} = 4). To calculate a standardized mean difference, the control group mean is subtracted from the experimental group mean and this difference is divided

by the pooled standard deviation – that is, $(10-8)/4 = 0.5$. According to Cohen (1988), a "small" standardized mean difference is 0.2, a "medium" standardized mean difference is 0.5, and a "large" standardized mean difference is 0.8. Thus, the treatment effect in our hypothetical example is medium sized.

When two continuous variables are related, the Pearson product-moment correlation coefficient (r) is most often used. Values of r can range from +1.0 (a perfect positive correlation) to −1.0 (a perfect negative correlation). A correlation coefficient of 0 indicates that the two variables are not (linearly) related. The sign on the correlation gives the direction of the relation between the two variables – a positive sign indicates that the relation is positive, whereas a negative sign indicates the relation is negative. The value of the correlation indicates the strength of the relation. Most correlations are not perfect. According to Cohen (1988), a "small" correlation is ± .1, a "medium" correlation is ± .3, and a "large" correlation is ± .5. Chapter 5 discusses how to combine standardized mean differences and correlation coefficients.

1.6 Some Issues to Consider When You Conduct a Meta-Analysis

1.6.1 Publication Bias and Study Quality

It is well documented that studies that report statistically significant results are more likely to be published than are studies reporting nonsignificant results (for example, Greenwald, 1975). In meta-analysis, the conditional publication of studies with significant results has been called the "file drawer problem" (Rosenthal, 1979). The most extreme version of this problem would result if only 1 out of 20 studies conducted was published and the remaining 19 studies were located in researchers' file drawers (or garbage cans), assuming that the .05 significance level is used. If publication bias is a problem, then the studies included in a meta-analysis may represent a biased subset of the total number of studies that are conducted on the topic. Chapter 3 describes some graphing procedures that can be used to detect publication bias.

One way to reduce publication bias is to include unpublished studies (for example, theses, dissertations) in the meta-analysis. Including unpublished studies in a meta-analysis, however, raises questions about the qualitative differences between published and unpublished studies. Because most refereed journals have reasonably strict standards for publication, published studies may be more methodologically sound than unpublished studies. Eysenck (1978) argued that when researchers fail to exclude studies of poor design, a meta-analysis becomes an exercise in "mega-silliness" that only demonstrates the axiom "garbage in — garbage out." Our personal belief is that unpublished studies should be included in a meta-analysis, but that studies should be coded on variables related to methodological quality (for example, random assignment, double blind procedures, publication status). You can then test whether the coded variables moderate the treatment effects (see Chapters 8 and 9).

1.6.2 Missing Effect-Size Estimates

Missing data is perhaps the largest problem facing the practicing meta-analyst. Missing effect-size estimates pose a particularly difficult problem because meta-analytic procedures cannot be used at all without a statistical measure for the results of a study (Pigott, 1994). Sometimes research reports do not include enough information (for example, means, standard deviations, statistical tests) to permit the calculation of an effect-size estimate. Unfortunately, the proportion of studies with missing effect-size estimates in a meta-analysis is often quite large, about 25% in psychological studies (Bushman & Wang, 1995, 1996). Vote-counting procedures can be used on studies that don't report enough information to calculate effect-size estimates but do report information about the direction and/or statistical significance of results (Bushman, 1994). Vote-counting procedures are described in Chapter 6.

Currently, the most common "solutions" to the problem of missing effect-size estimates are (a) to omit from the review those studies with missing effect-size estimates and analyze only complete cases, (b) to set the missing effect-size estimates equal to zero, (c) to set the missing effect-size estimates equal to the

mean that is obtained from studies with effect-size estimates, (d) to set studies equal to the conditional mean that is obtained from studies with effect-size estimates (that is, Buck's, 1960, method), and (e) to use the available information in a research report to get a lower limit for the effect-size estimate (Rosenthal, 1994). Unfortunately, all of these procedures have serious problems that limit their usefulness (Bushman & Wang, 1996).

We proposed an alternative procedure for handling missing effect-size estimates (Bushman & Wang, 1996). Our procedure, called the combined procedure, combines sample effect-sizes and vote counts to estimate the population effect size. We believe that the combined procedure, described in Chapter 7, is the method of choice for handling missing effect-size estimates if some studies do not provide enough information to calculate effect-size estimates but do provide information about the direction and/or statistical significance of results.

1.6.3 Fixed- and Random-Effects Models

Effect-size estimates should not be combined unless they are homogeneous or similar in magnitude. You can formally test whether effect-size estimates are too heterogeneous to combine. A statistically significant heterogeneity test implies that variation in effects between-studies is significantly larger than you would expect by random chance. Between-studies variation in effects can be treated as fixed or random (Hedges & Olkin, 1985). The fixed-effects model assumes that the population effect size is a single fixed value, whereas the random-effects model assumes that the population effect size is a randomly distributed variable with its own mean and variance. When between-studies effect-size variation is treated as fixed, the only source of variation treated as random is the within-studies sampling variation. By entering known study characteristics in an analysis of variance (ANOVA) or regression model, the meta-analyst might be able to explain the "extra" variation between-studies (see Hedges, 1994). If the "extra" variation can be explained by a few simple study characteristics, then a fixed-effects model should be used. When a fixed-effects model is used, generalizations can be made to a universe of studies with similar study characteristics. The reviewers should use

random-effects models if the differences between studies are too complicated to be captured by a few study characteristics. When a random-effects model is used, generalizations can be made to a universe of such diverse studies. Although generalizability is higher for random-effects models than for fixed-effects models, statistical power is higher for fixed-effects models than for random-effects models, (Rosenthal, 1995). Consequently, effect-size confidence intervals are narrower for fixed-effects models than for random-effects models. Fixed- and random-effects models are discussed in Chapters 8 and 9, respectively.

1.6.4 Correlated Effect-Size Estimates

Most meta-analytic procedures are based on the assumption that the effect-size estimates that are to be combined are independent. This independence assumption, however, is often violated. Some studies may compare multiple variants of a treatment with a common control. These studies, called multiple-treatment studies (Gleser & Olkin, 1994), will contribute more than one treatment versus control effect-size estimate. Because of the common control group, the effect-size estimates will be correlated. Other studies, called multiple-endpoint studies (Gleser & Olkin, 1994), may include only one treatment and one control but may use multiple dependent variables as endpoints for each participant. A treatment versus control effect-size estimate may be calculated for each endpoint measure. Because measures on each participant are correlated, the effect-size estimates for the measures will be correlated within studies. The best way to combine correlated effect-size estimates is to use multivariate procedures (Gleser & Olkin, 1994; Kalaian & Raudenbush, 1996). We discuss multivariate procedures in meta-analysis in Chapter 10.

1.7 Using the SAS System to Conduct a Meta-Analysis

Although meta-analytic procedures have been around for about 100 years, only since the advent of the digital computer have meta-analytic methods become accessible to practicing meta-analysts. A good meta-analytic software package should have the capability to (a) manage meta-analytical databases, (b) perform numerical calculations based on meta-analytical procedures, (c) use graphical displays to illustrate assumptions about meta-analytic procedures and to present the findings from a meta-analytical review, (d) produce the summary report of a meta-analytical review. None of the existing meta-analytic packages, however, have all of these capabilities (see Normand, 1995, for a review). Although SAS software is not specifically designed to conduct meta-analytic reviews, it has the procedures that are needed to manage databases, analyze data, and graph results. Thus, we believe that SAS is the software of choice for conducting a meta-analytic review. We hope that the SAS code in this book will make meta-analytic methods even more accessible to individuals who want to conduct a meta-analysis.

1.8 References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 1–29.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99*, 388–399.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Berkowitz, L. (1990). On the formation and regulation of anger and aggression: A cognitive-neoassociation analysis. American Psychologist, *45*, 494–503.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526–542.

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review, 40*, 207–227.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B, 22*, 302–303.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 193–213). New York: Russell Sage Foundation.

Bushman, B. J., & Cooper, H. M. (1990). Effects of alcohol on human aggression: An integrative research review. *Psychological Bulletin, 107*, 341–354.

Bushman, B. J. & Wang, M. C. (1995). A procedure for combining sample correlations and vote counts to obtain an estimate and a confidence interval for the population correlation coefficient. *Psychological Bulletin, 117*, 530–546.

Bushman, B. J., & Wang, M. C. (1996). A procedure for combining sample standardized mean differences and vote counts to estimate the population standardized mean difference in fixed effects models. *Psychological Methods, 1*, 66–80.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review; 48*, 378–399.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society, Supplement 4*(1), 102–118.

Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. *Annals of Mathematical Statistics, 14*, 205–216.

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

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

Cooper, H. M., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.) *Handbook of research synthesis* (p. 4). New York: Russell Sage Foundation.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87*, 442–449.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33,* 517.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83–96.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver & Boyd. (Original work published 1925. Final edition published 1951).

Friedman, H. (1968). Magnitude of an experimental effect and a table for its rapid estimation. *Psychological Bulletin, 70,* 245–251.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8.

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

Gold, J. A. (1984). *Principles of psychological research*. Homewood, IL: Dorsey Press.

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

Harris, M. J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory and Psychology, 1*, 375–382.

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

Hedges, L. V., Cooper, H. M., & 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.

Hogben, L. (1957). *Statistical theory*. London: Allen & Unwin.

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

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1*, 227–235.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*, 635–642.

Mann, C. (1994). Can meta-analysis make policy. *Science, 266,* 960–962.

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

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

Normand, J (1995). Meta-analytical software review. *American Statistician, 49,* 352–365.

Nunnally, J. (1960). The place of statistics in psychology. Educational and *Psychological Measurement, 20*, 641–650.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. New York: Wiley.

Olkin, I. (1990). History and goals. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 3–26). New York: Russell Sage Foundation.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal, 2,* 1243–1246.

Pearson, K. (1933). On a method of determining whether a sample of size *n* supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika, 25,* 379–410.

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

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

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. American Statistician, 39, 220–223.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schulman, J. L., Kupst, M. J., & Suran, B. G. (1976). The worship of "p": Significant yet meaningless research results. *Bulletin of the Menninger Clinic, 40,* 134–143.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61,* 293–316.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 318,* 262–264.

Stern, G. S.; McCants, T. R.; Pettine, P. W. (1982). The relative contribution of controllable and uncontrollable life events to stress and illness. *Personality and Social Psychology Bulletin, 8,* 140–145.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). The American soldier: Adjustments during army life, Vol. 1. Princeton, NJ: Princeton University Press.

Tippett, L. H. C. (1931). *The methods of statistics*. London: Williams and Norgate.

Yates, F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *Journal of the American Statistical Association, 46,* 19–34.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science, 28,* 556–580.