1 Integrating Research Findings Across Studies

Before we delve into a discussion of methods, we would like to consider a concrete example. The next section presents a set of studies to be reviewed, then a sample narrative review, followed by a critique of this review. It has been our experience that personal experience with the problems of such a review greatly enhances the learning process.

General Problem and an Example

A major task in all areas of science is the development of theory. In many cases, the theorists have available the results of a number of previous studies on the subject of interest. Their first task is to find out what empirical relationships have been revealed in these studies so they can take them into account in theory construction. In developing an understanding of these relationships, it is often helpful in reviewing the studies to make up a table summarizing the findings of these studies. Table 1.1 shows such a summary table put together by a psychologist attempting to develop a theory of the relationship between job satisfaction and organizational commitment. In addition to the observed correlations and their sample sizes, the psychologist has recorded data on (1) sex, (2) organization size, (3) job level, (4) race, (5) age, and (6) geographical location. The researcher believes variables 1, 2, 3, and 4 may affect the extent to which job satisfaction gets translated into organizational commitment. The researcher has no hypotheses about variables 5 and 6 but has recorded them because they were often available.

As an exercise in integrating findings across studies and constructing theory, we would like you to spend a few minutes examining and

interpreting the data in Table 1.1. We would like you to jot down the following:

1. The tentative conclusions you reached about the relationship between job satisfaction and organizational commitment and the variables that do and do not moderate that relationship
2. An outline of your resulting theory of this relationship

A TYPICAL INTERPRETATION OF THE EXAMPLE DATA

A typical report on the findings shown in Table 1.1 would run like this: The correlation between occupational commitment and job satisfaction varies from study to study with the correlation varying between −.10 and .56. Although 19 out of 30 studies found a significant correlation, 11 of 30 studies found no relationship between commitment and satisfaction.

For male work populations, commitment and satisfaction were correlated in 8 studies and not correlated in 7 (i.e., correlated in 53% of the studies), while for women there was a correlation in 11 of 15 cases (or in 73% of the studies). Correlation was found in 83% of the large organizations but in only 50% of the small organizations. Correlation was found in 79% of the blue-collar populations but in only 50% of the white-collar populations. Correlation was found in 67% of the populations that were all white or mixed race, while correlation was found in only 50% of those work populations that were all black. Correlation was found in 83% of the cases in which the workforce was all younger than 30 or a mixture of younger and older workers, while not a single study with only older workers found a significant correlation. Finally, 65% of the studies done in the north found a correlation, while only 58% of the southern studies found a correlation. Each of the differences between work populations could be taken as the basis for a hypothesis that there is an interaction between that characteristic and organizational commitment in the determination of job satisfaction.

If the studies done on older workers are removed, then significant correlation is found for 19 of the remaining 23 studies. Within these 23 studies with younger or mixed-age work populations, all 10 correlations for large organizations were significant.

There are 13 studies of younger or mixed-age work populations in small organizations. Within this group of studies, there is a tendency for correlation between organizational commitment and job satisfaction to be more likely found among women, among blue-collar workers, in all-black work populations, and in the north.

Table 1.1 Correlations between organizational commitment and job satisfaction.

Study	N	r	*Sex*	*Size of Organization*	*White vs. Blue Collar*	*Race*	*Under vs. Over 30*	*North Vs. South*
1	20	.46*	F	S	WC	B	U	N
2	72	.32**	M	L	BC	Mixed	Mixed	N
3	29	.10	M	L	WC	W	O	N
4	30	.45**	M	L	WC	W	Mixed	N
5	71	.18	F	L	BC	W	O	N
6	62	.45**	F	S	BC	W	U	N
7	25	.56**	M	S	BC	Mixed	U	S
8	46	.41**	F	L	WC	W	Mixed	S
9	22	.55**	F	S	WC	B	U	N
10	69	.44**	F	S	BC	W	U	N
11	67	.34**	M	L	BC	W	Mixed	N
12	58	.33**	M	S	BC	W	U	N
13	23	.14	M	S	WC	B	O	S
14	20	.36	M	S	WC	W	Mixed	N
15	28	.54**	F	L	WC	W	Mixed	S
16	30	.22	M	S	BC	W	Mixed	S
17	69	.31**	F	L	BC	W	Mixed	N
18	59	.43**	F	L	BC	W	Mixed	N
19	19	.52*	M	S	BC	W	Mixed	S
20	44	–.10	M	S	WC	W	O	N
21	60	.44**	F	L	BC	Mixed	Mixed	N
22	23	.50**	F	S	WC	W	Mixed	S
23	19	–.02	M	S	WC	B	O	S
24	55	.32**	M	L	WC	W	Mixed	Unknown
25	19	.19	F	S	WC	B	O	N
26	26	.53**	F	S	BC	B	U	S
27	58	.30*	M	L	WC	W	Mixed	S
28	25	.26	M	S	WC	W	U	S
29	28	.09	F	S	BC	W	O	N
30	26	.31	F	S	WC	Mixed	U	S

*$p < .05$.

**$p < .01$.

CONCLUSIONS OF THE REVIEW

Organizational commitment and job satisfaction are correlated in some organizational settings but not in others. In work groups in which all workers are older than 30, the correlation between commitment and satisfaction was never significant. For young or mixed-age work populations, commitment and satisfaction are always correlated in large organizations. For young or mixed-age work populations in small organizations, correlation was found in 9 of 13 studies with no organizational feature capable of perfectly accounting for those cases in which correlation was not found.

These findings are consistent with a model that assumes that organizational commitment grows over about a 10-year period to a maximum value at which it asymptotes. Among older workers, organizational commitment may be so uniformly high that there is no variation. Hence, among older workers, there can be no correlation between commitment and job satisfaction. The finding for large organizations suggests that growth of commitment is slower there, thus generating a greater variance among workers of different ages within the younger group.

CRITIQUE OF THE SAMPLE REVIEW

The preceding review was conducted using review practices that characterize many narrative review articles not only in psychology but in sociology, education, and the rest of the social sciences as well. Yet every conclusion in the review is false. The data were constructed by a Monte Carlo run in which the population correlation was always .33. After a sample size was randomly chosen from a distribution centering about 40, an observed correlation was chosen using the standard distribution for r with mean $\rho = .33$ and variance

$$\frac{(1-\rho^2)^2}{N-1}.$$

That is, the variation in results in Table 1.1 is entirely the result of sampling error. Each study is conducted on a small sample and hence generates an observed correlation that departs by some random amount from the population value of .33. The size of the departure depends on the sample size. Note that the largest and smallest values found in Table 1.1 are all from studies with very small samples. The larger sample size studies tend to show less of a random departure from .33.

The moderator effects appear to make sense, yet they are purely the results of chance (i.e., sampling error). The values for the organizational characteristics were assigned to the studies randomly.

The crucial lesson to be learned from this exercise is this: "Conflicting results in the literature" may be entirely artifactual. The data in Table 1.1

were generated by using one artifact for generating false variation across studies, sampling error. There are other artifacts that are found in most sets of studies: Studies vary in terms of the quality of measurement (reliability) of their scales; researchers make computational or computer errors; people make typographical errors in copying numbers from computer output or in copying numbers from handwritten tables onto manuscripts or in setting tables into print; researchers study variables in settings with greater or smaller ranges of individual differences (range variation); and so on. In our experience (described later), many of the interactions hypothesized to account for differences in findings in different studies are nonexistent; that is, they are apparitions composed of the ectoplasm of sampling error and other artifacts.

Problems With Statistical Significance Tests

In the data set given in Table 1.1, all study population correlations are actually equal to .33. Of the 30 correlations, 19 were found to be statistically significant. However, 11 of the 30 correlations were not significant. That is, the significance test gave the wrong answer 11 out of 30 times, an error rate of 37%. Many people express shock that the error rate can be greater than 5%. The significance test was derived in response to the problem of sampling error, and many believe that the use of significance tests guarantees an error rate of 5% or less. This belief is false. Statisticians have pointed this out for many years; the possibility of high error rates is brought out in discussions of the "power" of statistical tests. However, statistics instructors are well aware that this point is not understood by most students. The 5% error rate is guaranteed only if the null hypothesis is true. If the null hypothesis is false, then the error rate can go as high as 95%.

Let us state this in more formal language. If the null hypothesis is true for the population and our sample data lead us to reject it, then we have made a Type I error. If the null hypothesis is false for the population and our sample data lead us to accept it, then we have made a Type II error. The statistical significance test is defined in such a way that the Type I error rate is at most 5%. However, the Type II error rate is typically left free to be as high as 95%. The question is which error rate applies to a given study. The answer is that the relevant error rate can only be known if we know whether the null hypothesis is true or false for that study. If we know that the null hypothesis is true, then we know that the significance test has an error rate of 5%. Of course, if we know that the null hypothesis is true and we still do a significance test, then we should wear a dunce cap, because if we know the null hypothesis to be true, then we can obtain a 0% error rate by ignoring the data. That is, there is a fundamental circularity to the significance test. If you do not know whether the null hypothesis is true or false, then you do not know whether the relevant error rate is

Type I or Type II; that is, you do not know if your error rate is 5% or some value as high as 95%. There is only one way to guarantee a 5% error rate in all cases: Abandon the significance test and use a confidence interval.

Consider our hypothetical example from Table 1.1. However, let us simplify the example still further by assuming that the sample size is the same for all studies, say $N = 40$. The one-tailed significance test for a correlation coefficient is $\sqrt{N-1}\, r \geq 1.64$; in our case, $\sqrt{39}\, r \geq 1.64$ or $r \geq .26$. If the population correlation is .33 and the sample size is 40, the mean of the sample correlations is .33, while the standard deviation is $(1-\rho^2)/\sqrt{N-1} = (1-.33^2)/\sqrt{39} = .14$. Thus, the probability that the observed correlation will be significant is the probability that the sample correlation will be greater than .26 when the population value is .33 and a standard deviation of .14:

$$P\{r \geq .26\} = P\left\{\frac{r-.33}{.14} \geq \frac{.26-.33}{.14}\right\} = P\{z \geq -.50\} = .69.$$

That is, if all studies were done with a sample size of 40, then a population correlation of .33 would mean an error rate of 31% (i.e., 1 − .69 = .31).

Suppose we alter the population correlation in our hypothetical example from .33 to .20. Then the probability that the observed correlation will be significant drops from .69 to

$$P\{r \geq .26\} = P\left\{z \geq \frac{.26-.20}{.15} = .39\right\} = .35.$$

That is, the error rate rises from 31% to 65%. In this realistic example, we see that the error rate can be over 50%. A two-to-one majority of the studies can find the correlation to be not significant despite the fact that the population correlation is always .20.

Error rates of 50% or higher have been shown to be the usual case in many research literatures. Thus, reviewers who count the number of significant findings are prone to incorrectly conclude that a relationship does not exist when it does. Furthermore, as Hedges and Olkin (1980) pointed out, this situation only gets worse as more studies are conducted. The reviewer will become ever more convinced that the majority of studies show no effect and that the effect thus does not exist. Statistical power has been examined in much of the research literature in psychology, starting with Cohen (1962) and extending up to the present. In most literatures, the mean statistical power is in the .40 to .60 range and is as low as .20 in some areas (Hunter, 1997; Schmidt, 1996; Schmidt & Hunter, 2003; Sedlmeier & Gigerenzer, 1989). And, surprisingly, over time there has been little or no increase in statistical power in published studies (Sedlmeier & Gigerenzer, 1989. Maxwell (2004) explores the reasons why this is the case. Other possible reasons are given in Chapter 13.

If the null hypothesis is true in a set of studies, then the base rate for significance is not 50% but 5%. If more than 1 in 20 studies finds significance, then the null hypothesis must be false in some studies. We must then avoid an error made by some reviewers who know the 5% base rate. For example, if 35% of the findings are significant, some have concluded that "Because 5% will be significant by chance, this means that the number of studies in which the null hypothesis is truly false is 35 – 5 = 30%." Our hypothetical example shows this reasoning to be false. If the population correlation is .20 in every study and the sample size is always 40, then there will be significant findings in only 35% of the studies, even though the null hypothesis is false in all cases.

The typical use of significance test results leads to gross errors in traditional review studies. Most such reviews falsely conclude that further research, focused on moderator variables, is needed to resolve the "conflicting results" in the literature. These errors in review studies can only be eliminated if errors in the interpretation of significance tests can be eliminated. Yet those of us who have been teaching power to generation after generation of graduate students have been unable to change the reasoning processes and the false belief in the 5% error rate (Sedlmeier & Gigerenzer, 1989).

This example illustrates a critical point. Traditional reliance on statistical significance tests in interpreting studies leads to false conclusions about what the study results mean; in fact, the traditional approach to data analysis makes it virtually impossible to reach correct conclusions in most research areas (Hunter, 1997; Schmidt, 1996, 2010).

A common reaction to the preceding critique of traditional reliance on significance testing goes something like this: "Your explanation is clear, but I don't understand how so many researchers (and even some methodologists) could have been so wrong so long on a matter as important as the correct way to analyze data? How could psychologists and other researchers have failed to see the pitfalls of significance testing?" Over the years, a number of methodologists have addressed this question (Carver, 1978; Cohen, 1994; Guttman, 1985; Meehl, 1978; Oakes, 1986; Rozeboom, 1960; Schmidt & Hunter, 1997). In their statistics classes, young researchers have typically been taught a lot about Type I error and very little about Type II error and statistical power. Thus, they are unaware that the error rate is very large in the typical study; they tend to believe the error rate is the alpha level used (typically .05 or .01). In addition, empirical research suggests that most researchers believe that the use of significance tests provides them with many nonexistent benefits in understanding their data. For example, most researchers believe that a statistically significant finding is a "reliable" finding in the sense that it will replicate if a new study is conducted (Carver, 1978; Oakes, 1986; Schmidt, 1996; Schmidt & Hunter, 1997). For example, they believe that if a result is significant at the .05 level, then the probability of replication in subsequent studies (if conducted) is 1.00 − .05 = .95. This belief is completely false. The probability of replication is the statistical

power of the study and is almost invariably much lower than .95 (e.g., typically .50 or less). Killeen (2005a, 2005b) proposed a statistic called *P-rep* that he claimed gives the probability that a research finding would be replicated in a new study. For a few years, this statistic was widely used. However, Trafimow, MacDonald, Rice, and Carlson (2010) demonstrated mathematically that the *P-rep* statistic did not in fact provide this probability. Today *P-rep* is rarely used.

Most researchers also believe that if a result is nonsignificant, one can conclude that it is probably just due to chance, another false belief, as illustrated in our example in which all nonsignificant results were Type II errors. There are other widespread but false beliefs about the usefulness of information provided by significance tests (Carver, 1978; Oakes, 1986). Discussion of these beliefs can be found in Schmidt (1996) and Schmidt and Hunter (1997).

Another fact is relevant at this point: The physical sciences, such as physics and chemistry, do not use statistical significance testing in interpreting their data (Cohen, 1990). Instead, they use confidence intervals. It is no accident, then, that these sciences have not experienced the debilitating problems described here that are inevitable when researchers rely on significance tests. Given that the physical scientists regard reliance on significance testing as unscientific, it is ironic that so many psychologists defend the use of significance tests on grounds that such tests are the objective and scientifically correct approach to data analysis and interpretation. In fact, it has been our experience that psychologists and other behavioral scientists who attempt to defend significance testing usually equate null hypothesis statistical significance testing with scientific hypothesis testing in general. They argue that hypothesis testing is central to science and that the abandonment of significance testing would amount to an attempt to have a science without hypothesis testing. They falsely believe that significance testing and hypothesis testing in science are one and the same thing. This belief is tantamount to stating that physics, chemistry, and the other physical sciences are not legitimate sciences because they do not test their hypotheses using statistical significance testing. Another logical implication of this belief is that prior to the introduction of null hypothesis significance testing by R. A. Fisher (1932) in the 1930s, no legitimate scientific research was possible. The fact is, of course, that there are many ways to test scientific hypotheses—and that significance testing is one of the least effective methods of doing this (Schmidt & Hunter, 1997).

Is Statistical Power the Solution?

Some researchers believe that the only problem with significance testing is low power and that if this problem could be solved there would be no problems with reliance on significance testing. These individuals see the solution as larger sample sizes. They believe that the problem would be

solved if every researcher, before conducting each study, would calculate the number of subjects needed for "adequate" power (usually taken as power of .80) and then would use that sample size. What this position overlooks is that this requirement would make it impossible for most studies ever to be conducted. At the start of research in a given area, the questions are often of the form "Does Treatment A have an effect?" (e.g., Does interpersonal skills training have an effect? Does cognitive behavior therapy work?). If Treatment A indeed has a substantial effect, the sample size needed for adequate power may not be prohibitively large. But as research develops, subsequent questions tend to take the form "Is the effect of Treatment A larger than the effect of Treatment B?" (e.g., Is the effect of the new method of training larger than that of the old method? Is Predictor A more valid than Predictor B?). The effect size then becomes the *difference* between the two effects. Such effect sizes will often be small, and the required sample sizes are therefore often quite large—1,000 or 2,000 or more (Schmidt & Hunter, 1978). And this is just to attain power of .80, which still allows a 20% Type II error rate when the null hypothesis is false—an error rate most would consider high. Many researchers cannot obtain that many subjects, no matter how hard they try; either it is beyond their resources or the subjects are just not available at any cost. Thus, the upshot of this position would be that many—perhaps most—studies would not be conducted at all.

People advocating the power position say this would not be a loss. They argue that a study with inadequate power cannot support a research conclusion and therefore should not be conducted. Such studies, however, contain valuable information when combined with others like them in a meta-analysis. In fact, accurate meta-analysis results can be obtained based on studies that *all* have inadequate statistical power individually, because meta-analysis can provide precise estimates of average effect size. The information in these studies is lost if these studies are never conducted.

The belief that such studies are worthless is based on two false assumptions: (1) the assumption that every individual study must be able to justify a conclusion on its own, without reference to other studies, and (2) the assumption that every study should be analyzed using significance tests. One of the contributions of meta-analysis has been to show that no single study is adequate by itself to answer a scientific question. Therefore, each study should be considered as a data point to be contributed to a later meta-analysis. In addition, individual studies should be analyzed using not significance tests but point estimates of effect sizes and confidence intervals.

How, then, *can* we solve the problem of statistical power in individual studies? Actually, this problem is a pseudoproblem. It can be "solved" by discontinuing the significance test. As Oakes (1986, p. 68) noted, statistical power is a legitimate concept only within the context of statistical significance testing. If significance testing is not used, then the concept of

statistical power has no place and is not meaningful. In particular, there need be no concern with statistical power when point estimates and confidence intervals are used to analyze data in studies and meta-analysis is used to integrate findings across studies.

Our critique of the traditional practice of reliance on significance testing in analyzing data in individual studies and in interpreting research literature might suggest a false conclusion, namely, that if significance tests had never been used, the research findings would have been consistent across different studies examining a given relationship. Consider the correlation between job satisfaction and job performance in Table 1.1. Would these studies have all had the same findings if researchers had not relied on significance tests? Absolutely not: The correlations would have varied widely (as indeed they did). The major reason for such variability in correlations is simple sampling error—caused by the fact that the small samples used in individual research studies are randomly unrepresentative of the populations from which they are drawn. Most researchers severely underestimate the amount of variability in findings that is caused by sampling error.

The law of large numbers correctly states that large random samples are representative of their populations and yield parameter estimates that are close to the actual population values. Many researchers seem to believe that the same law applies to small samples. As a result, they erroneously expect statistics computed on small samples (e.g., 50 to 300) to be close approximations to the real (population) values. In one study we conducted (Schmidt, Ocasio, Hillery, & Hunter, 1985), we drew random samples (small studies) of $N = 30$ from a much larger single data set ($N = 1{,}455$; $r = .22$) and computed results on each $N = 30$ sample. The resulting validity estimates varied dramatically from "study" to "study," ranging from –.21 to .61, with all this variability being due solely to sampling error (Schmidt, Ocasio, et al., 1985). Yet when we showed these data to researchers, they found it hard to believe that each "study" was a random draw from the same larger study. They did not believe simple sampling error could produce that much variation. They were shocked because they did not realize how much variation simple sampling error produces in research studies.

There are two alternatives to the significance test. At the level of review studies, there is meta-analysis. At the level of single studies, there is the confidence interval.

Confidence Intervals

Consider Studies 17 and 30 from our hypothetical example in Table 1.1. Study 17, with $r = .31$ and $N = 69$, finds the correlation to be significant at the .01 level. Study 30, with $r = .31$ and $N = 26$, finds the correlation to be not significant. That is, two authors with an identical finding, $r = .31$, come

to opposite conclusions. Author 17 concludes that organizational commitment is highly related to job satisfaction, while Author 30 concludes that they are unrelated. Thus, two studies with identical findings can lead to a review author claiming "conflicting results in the literature."

The conclusions are quite different if the results are interpreted with confidence intervals. Author 17 reports a finding of $r = .31$ with a 95% confidence interval of $.10 \leq \rho \leq .52$. Author 30 reports a finding of $r = .31$ with a 95% confidence interval of $-.04 \leq \rho \leq .66$. There is no conflict between these results; the two confidence intervals overlap substantially.

Consider now Studies 26 and 30 from Table 1.1. Study 26 finds $r = .53$ with $N = 26$, which is significant at the .01 level. Study 30 finds $r = .31$ with $N = 26$, which is not significant. That is, we have two studies with the same sample size but apparently widely divergent results. Using significance tests, one would conclude that there must be some moderator that accounts for the difference. This conclusion is false.

Had the two studies used confidence intervals, the conclusion would have been different. The confidence interval for Study 26 is $.25 \leq \rho \leq .81$, and the confidence interval for Study 30 is $-.04 \leq \rho \leq .66$. It is true that the confidence interval for Study 30 includes $\rho = 0$, while the confidence interval for Study 26 does not; this is the fact registered by the significance test. The crucial thing, however, is that the two confidence intervals show an overlap of $.25 \leq \rho \leq .66$. Thus, consideration of the two studies together leads to the correct conclusion that it is possible that both studies could imply the same value for the population correlation ρ. Indeed, the overlapping intervals include the correct value, $\rho = .33$.

Two studies with the same population value can have non-overlapping confidence intervals, but this is a low-probability event (about 5%). But, then, confidence intervals are not the optimal method for looking at results across studies; that distinction belongs to meta-analysis.

Confidence intervals are more informative than significance tests for two reasons. First, the interval is correctly centered on the observed value rather than on the hypothetical zero value of the null hypothesis. Second, the confidence interval gives the researcher a correct picture of the extent of uncertainty in small sample studies. It may be disconcerting to see a confidence interval as wide as $-.04 \leq \rho \leq .66$, but that is far superior to the frustration produced over the years by the false belief in "conflicting results."

Confidence intervals can be used to generate definitions for the phrase "small sample size." Suppose we want the confidence interval for the correlation coefficient to define the correlation to the first digit, that is, to have a width of $\pm.05$. Then, for small population correlations, the minimum sample size is approximately 1,538. For a sample size of 1,000 to be sufficient, the population correlation must be at least .44. Thus, under this standard of accuracy, for correlational studies, "small sample size" includes all studies with less than a thousand persons and often extends above that.

There is a similar calculation for experimental studies. If the statistic used is the *d* statistic (by far the most frequent choice), then small effect sizes will be specified to their first digit only if the sample size is 3,076. If the effect size is larger, then the sample size must be even greater than 3,076. For example, if the difference between the population means is .30 standard deviations, then the minimum sample size that yields accuracy to within ±.05 of .30 is 6,216. Thus, given this standard of accuracy, for experimental studies, "small sample size" begins with 3,000 and often extends well beyond that. Now think about the fact that many, perhaps most, experimental studies in behavior labs have total *N*s between 20 and 50.

Since the publication of the first edition of this book in 1990, recognition of the superiority of confidence intervals and point estimates of effect sizes over significance tests has grown exponentially. The report of the task force on significance testing of the American Psychological Association (APA) (Wilkinson & The APA Task Force on Statistical Inference, 1999) stated that researchers should report effect size estimates and confidence intervals. The fifth and sixth editions of the APA *Publication Manual* stated that it is almost always necessary for primary studies to report effect size estimates and confidence intervals (American Psychological Association, 2001, 2009). Twenty-one research journals in psychology and education now require that these statistics be reported (B. Thompson, 2002). Some have argued that information on the methods needed to compute confidence intervals is not widely available. However, there are now helpful and informative statistics textbooks designed around point estimates of effect size and confidence intervals instead of significance testing (Cumming, 2012; Kline, 2004; Lockhart, 1998; Smithson, 2000). The Cumming (2012) book includes excellent online computer programs that make calculations easy and that illustrate critical statistical facts and principles. B. Thompson (2002) presents considerable information on computation of confidence intervals and cites many useful references that provide more detail (e.g., Kirk, 2001; Smithson, 2001). The August 2001 issue of *Educational and Psychological Measurement* was devoted entirely to methods of computing and interpreting confidence intervals. There are many other such publications (e.g., Borenstein, 1994).

Despite these developments, most published articles still use significance tests. How this can be is something of a mystery, given the fact that this practice has been completely discredited. Orlitzky (2011) argues that the problem is that the evidence against significance testing has not been *institutionalized.* Articles discrediting significance testing have been aimed at inducing individual researchers to change their statistical practices, not at a broader, more systematic or institutional change. But it very difficult for individual researchers to go against what has become an institutionalized practice in most journals. What is needed, he contends, is top-down disciplinary-wide changes in the research culture. This is a broad recommendation. For example, he says that urging individual journal editors to

require effect sizes and confidence intervals will produce little change. There must be an enforceable agreement at the level of the entire discipline that proper data analysis procedures must be used in primary studies, along with major changes in the way research methods are taught in graduate programs. This is a long-term proposition for culture change. Fortunately, as we will see next, meta-analysis makes it possible to make progress in developing cumulative knowledge in the interim even if significance testing continues to be used in individual primary studies.

Meta-Analysis

Is there a quantitative analysis that would have shown that all the differences in Table 1.1 might stem from sampling error? Suppose we compute the variance of the correlations, weighting each by its sample size. The value we obtain is .02258 (*SD* = .150). We can also compute the variance expected solely from sampling error. The formula for the sampling error variance of each individual correlation r_i is

$$(1 - .331^2)^2/(N_i - 1),$$

where .331 is the sample size–weighted mean of the correlations in Table 1.1. If we weight each of these estimates by its sample size (as we did when we computed the observed variance), the formula for variance expected from sampling error is

$$S_e^2 = \frac{\sum_{i=1}^{i=30} \left[\frac{N_i \left(1-.331^2\right)^2}{N_i - 1} \right]}{\sum N_i}.$$

This value is .02058 (*SD* = .144). The ratio of variance expected from sampling error to actual (observed) variance is .02058/.02258 = .91. Thus, sampling error alone accounts for an estimated 91% of the observed variance in the correlations. The square root of the .91 is the correlation between the sampling errors and the observed correlations. This correlation is .95 and is a more informative index than the percent of variance accounted for (Schmidt, 2010). The best conclusion is that the relationship between job satisfaction and organizational commitment is constant across sexes, races, job levels, ages, geographical locations, and size of organization. (The difference between a correlation of 1.00 and our value of .95 is due to second-order sampling error, which is discussed in Chapter 9.) The best estimate of this constant value is .331, the sample size–weighted mean of the 30 correlations. When in our oral presentations researchers analyzed

the data from these 30 studies qualitatively, different people came to different conclusions. In contrast, all researchers applying the quantitative method used here would (barring computational errors) come to exactly the same conclusion.

For theoretical purposes, the value .331 is not the one we want, because it is biased downward by unreliability in both measures. The effect of measurement error is to reduce all the observed correlations, and hence the mean correlation, below the actual correlation between the two constructs. What we are interested in is the construct-level correlation, because this correlation reflects the underlying science. Suppose from information in the 30 studies we estimate the average reliability of job satisfaction measures at .70 and the average reliability of organizational commitment measures at .60. Then the estimated correlation between true scores on the measures is $.331/\sqrt{.70(.60)} = .51$. This value is the best estimate of the construct-level correlation. Schmidt, Le, and Oh (in press) have shown that true scores and construct scores typically correlate about .98, so true score correlations are good estimates of construct correlations. The necessity of correcting for measurement error is discussed by Hedges (2009b, chap. 3).

Most artifacts other than sampling error that distort study findings are systematic rather than random. They usually create a downward bias in the obtained study *r* or *d* value. For example, all variables in a study must be measured and all measures of variables contain measurement error. There are no exceptions to this rule. The effect of measurement error is to downwardly bias every correlation or *d* value. Measurement error can also cause *differences* between studies: If the measures used in one study have more measurement error than those used in another study, the observed *r*s or *d*s will be smaller in the first study. Thus, meta-analysis must correct both for the downward bias and for the artifactually created differences between different studies. Corrections of this sort are discussed in Chapters 2 to 7.

Traditional review procedures are inadequate to integrate conflicting findings across large numbers of studies. As Glass (1976) pointed out, the results of hundreds of studies "can no more be grasped in our traditional narrative discursive review than one can grasp the sense of 500 test scores without the aid of techniques for organizing, depicting and interpreting data" (p. 4). In such areas as the effects of class size on student learning, the relationship of IQ to creativity, and the effects of psychotherapy on patients, literally hundreds of studies can accumulate over a period of only a few years. Glass (1976) noted that such studies collectively contain much more information than can be extracted from them using narrative review methods. He pointed out that because we have not exploited these gold mines of information, "We know much less than we have proven." What is needed are methods that will integrate results from existing studies to reveal patterns of relatively invariant underlying relationships and causalities, the establishment of which will constitute general principles and cumulative knowledge.

At one time in the history of psychology and the social sciences, the pressing need was for more empirical studies examining the problem in question. In many areas of research, the need today is not additional empirical data but some means of making sense of the vast amounts of data that have been accumulated. Because of the increasing number of areas within psychology and the other social sciences in which the number of available studies is quite large and the importance to theory development and practical problem solving of integrating conflicting findings to establish general knowledge, meta-analysis has come to play an increasingly important role in research. Such methods can be built around statistical and psychometric procedures that are already familiar to us. As Glass (1976) stated,

> Most of us were trained to analyze complex relationships among variables in the primary analysis of research data. But at the higher level, where variance, nonuniformity and uncertainty are no less evident, we too often substitute literary exposition for quantitative rigor. The proper integration of research requires the same statistical methods that are applied in primary data analysis. (p. 6)

Role of Meta-Analysis in the Behavioral and Social Sciences

The small-sample studies typical of psychological research produce seemingly contradictory results, and reliance on statistical significance tests causes study results to appear even more conflicting. Meta-analysis integrates the findings across such studies to reveal the simpler patterns of relationships that underlie the research literature, thus providing a basis for theory development. Meta-analysis can correct for the distorting effects of sampling error, measurement error, and other artifacts that produce the illusion of conflicting findings.

The goal in any science is the production of cumulative knowledge. Ultimately, this means the development of theories that explain the phenomena that are the focus of the scientific area. One example would be theories that explain how personality traits develop in children and adults over time and how these traits affect their lives. Another would be theories of what factors cause job and career satisfaction and what effects job satisfaction in turn has on other aspects of one's life. Before theories can be developed, however, we need to be able to precisely calibrate the relationships between variables. For example, what is the relationship between peer socialization and level of extroversion? What is the relationship between job satisfaction and job performance?

Unless we can precisely calibrate such relationships among variables, we do not have the raw materials out of which to construct theories. There is nothing for a theory to explain. For example, if the relationship between extroversion and popularity of children varies capriciously across different

studies from a strong positive to a strong negative correlation and everything in between, we cannot begin to construct a theory of how extroversion might affect popularity. The same applies to the relationship between job satisfaction and job performance.

The unfortunate fact is that most research literatures do show conflicting findings of this sort. Some studies find statistically significant relationships and some do not. In much of the research literature, this split is approximately 50–50 (Cohen, 1962, 1988; Schmidt, 2010; Schmidt & Hunter, 1997; Schmidt, Hunter, & Urry, 1976; Sedlmeier & Gigerenzer, 1989). This has been the traditional situation in most areas of the behavioral and social sciences. Hence, it has been very difficult to develop understanding, theories, and cumulative knowledge.

Today meta-analysis is being widely applied to solve this problem. The extent of the use of meta-analysis is mirrored in the fact that a Google search using this term produces over 50 million hits.

THE MYTH OF THE PERFECT STUDY

Before meta-analysis, the usual way in which scientists attempted to make sense of the research literature was by use of the narrative subjective review. In much of the research literature, however, there were not only conflicting findings but also large numbers of studies. This combination made the standard narrative subjective review a nearly impossible task—one shown by research on human information processing to be far beyond human capabilities. How does one sit down and make sense of, say, 210 conflicting studies?

The answer as developed in many narrative reviews was what came to be called the myth of the perfect study. Reviewers convinced themselves that most—usually the vast majority—of the available studies were "methodologically deficient" and should not even be considered in the review. These judgments of methodological deficiency were often based on idiosyncratic ideas: One reviewer might regard the Peabody Personality Inventory as "lacking in construct validity" and throw out all studies that used that instrument. Another might regard use of that same inventory as a prerequisite for methodological soundness and eliminate all studies *not* using this inventory. Thus, any given reviewer could eliminate from consideration all but a few studies and perhaps narrow the number of studies from 210 to, say, 7. Conclusions would then be based on these seven studies.

It has long been the case that the most widely read literature reviews are those appearing in textbooks. The function of textbooks, especially advanced-level textbooks, is to summarize what is known in a given field. No textbook, however, can cite and discuss 210 studies on a single relationship. Textbook authors would often pick out what they considered to be the one or two "best" studies and then base textbook conclusions on

just those studies, discarding the vast bulk of the information in the research literature. Hence, the myth of the perfect study.

In fact, there are no perfect studies. All studies contain measurement error in all measures used, as discussed later. Independent of measurement error, no study's measures have perfect construct validity. Furthermore, there are typically other artifacts that distort study findings. Even if a hypothetical (and it would have to be hypothetical) study suffered from none of these distortions, it would still contain sampling error—typically a substantial amount of sampling error—because sample sizes are rarely very large. Hence, no single study or small selected subgroup of studies can provide an optimal basis for scientific conclusions about cumulative knowledge. As a result, reliance on "best studies" did not provide a solution to the problem of conflicting research findings. This procedure did not even successfully deceive researchers into believing it was a solution—because different narrative reviewers arrived at different conclusions because they selected a different subset of "best" studies. Hence, the "conflicts in the literature" became "conflicts between the reviews."

SOME RELEVANT HISTORY

By the mid-1970s, the behavioral and social sciences were in serious trouble. Large numbers of studies had accumulated on many questions that were important to theory development and/or social policy decisions. Results of different studies on the same question typically were conflicting. For example, are workers more productive when they are satisfied with their jobs? The studies did not agree. Do students learn more when class sizes are smaller? Research findings were conflicting. Does participative decision making in management increase productivity? Does job enlargement increase job satisfaction and output? Does psychotherapy really help people? The studies were in conflict. As a consequence, the public and government officials were becoming increasingly disillusioned with the behavioral and social sciences, and it was becoming more and more difficult to obtain funding for research. In an invited address to the American Psychological Association in 1970, Senator Walter Mondale expressed his frustration with this situation:

> What I have *not* learned is what we should do about these problems. I had hoped to find research to support or to conclusively oppose my belief that quality integrated education is the most promising approach. But I have found very little conclusive evidence. For every study, statistical or theoretical, that contains a proposed solution or recommendation, there is always another, equally well documented, challenging the assumptions or conclusions of the first. No one seems to agree with anyone else's approach. But more distressing I must confess, I stand with my colleagues confused and often disheartened.

Then, in 1981, the director of the Federal Office of Management and Budget, David Stockman, proposed an 80% reduction in federal funding for research in the behavioral and social sciences. This proposal was politically motivated in part, but the failure of behavioral and social science research to be cumulative created the vulnerability to political attack. This proposed cut was a trial balloon sent up to see how much political opposition it would arouse. Even when proposed cuts are much smaller than a draconian 80%, constituencies can usually be counted on to come forward and protest the proposed cuts. This usually happens, and many behavioral and social scientists expected it to happen. But it did not. The behavioral and social sciences, it turned out, had no constituency among the public; the public did not care (see "Cuts Raise New Social Science Query," 1981). Finally, out of desperation, the American Psychological Association took the lead in forming the Consortium of Social Science Associations to lobby against the proposed cuts. Although this super association had some success in getting these cuts reduced (and even, in some areas, getting increases in research funding in subsequent years), these developments should make us look carefully at how such a thing could happen.

The sequence of events that led to this state of affairs was much the same in one research area after another. First, there was initial optimism about using social science research to answer socially important questions. Do government-sponsored job-training programs work? We will do studies to find out. Does Head Start really help disadvantaged kids? The studies will tell us. Does integration increase the school achievement of black children? Research will provide the answer. Next, several studies on the question are conducted, but the results are conflicting. There is some disappointment that the question has not been answered, but policy makers—and people in general—are still optimistic. They, along with the researchers, conclude that more research is needed to identify the supposed interactions (moderators) that have caused the conflicting findings. For example, perhaps whether job training works depends on the age and education of the trainees. Maybe smaller classes in the schools are beneficial only for lower IQ children. It is hypothesized that psychotherapy works for middle-class but not working-class patients. That is, the conclusion at this point is that a search for moderator variables is needed.

In the third phase, a large number of research studies are funded and conducted to test these moderator hypotheses. When they are completed, there is now a large body of studies, but instead of being resolved, the number of conflicts increases. The moderator hypotheses from the initial studies are not borne out, and no one can make sense out of the conflicting findings. Researchers conclude that the question that was selected for study in this particular case has turned out to be hopelessly complex. They then turn to the investigation of another question, hoping that this time the question will turn out to be more tractable. Research sponsors, government officials, and the public become disenchanted and cynical.

Research funding agencies cut money for research in this area and in related areas. After this cycle has been repeated enough times, social and behavioral scientists themselves become cynical about the value of their own work, and they publish articles endorsing the belief that behavioral and social science research is incapable, *in principle,* of developing cumulative knowledge and providing general answers to socially important questions. Examples of this include Cronbach (1975), Gergen (1982), and Meehl (1978).

Clearly, at this point, there was a critical need for some means of making sense of the vast number of accumulated study findings. Starting in the late 1970s, new methods of combining findings across studies on the same subject were developed. These methods were referred to collectively as *meta-analysis,* a term coined by Glass (1976). Applications of meta-analysis to the accumulated research literature (e.g., Schmidt & Hunter, 1977) showed that research findings were not nearly as conflicting as had been thought and that useful and sound general conclusions could, in fact, be drawn from existing research. The conclusion was that cumulative theoretical knowledge is possible in the behavioral and social sciences, and socially important questions can be answered in reasonably definitive ways. As a result, the gloom and cynicism that had enveloped many in the behavioral and social sciences lifted.

In fact, meta-analysis has even produced evidence that cumulativeness of research findings in the behavioral sciences is probably as great as that in the physical sciences. We have long assumed that our research studies are less consistent than those in the physical sciences. Hedges (1987) used meta-analysis methods to examine variability of findings across studies in 13 research areas in particle physics and 13 research areas in psychology. Contrary to common belief, his findings showed that there is as much variability across studies in physics as there is in psychology. Furthermore, he found that the physical sciences used methods to combine findings across studies that were "essentially identical" to meta-analysis. The research literature in both areas—psychology and physics—yielded cumulative knowledge when meta-analysis was properly applied. Hedges's major finding is that the frequency of conflicting research findings is no greater in the behavioral and social sciences than in the physical sciences. The fact that this finding has been so surprising to many social scientists points up the fact that we have long overestimated the consistency of research findings in the physical sciences. Furthermore, in the physical sciences, no research question can be answered by a single study, and physical scientists must and do use meta-analysis to make sense of their research literature, just as we do. (And, as noted earlier, in analyzing data in individual studies, the physical sciences do not use significance tests; they use point estimates and confidence intervals.)

Other changes have also been produced by meta-analysis. The relative status of reviews has changed dramatically. Journals that traditionally

published only primary studies and refused to publish reviews have now published meta-analytic reviews in large numbers for some years. In the past, research reviews were based on the narrative subjective method, and they had limited status and gained little credit for one in academic raises or promotions. The rewards went to those who conducted primary research. Not only is this no longer the case, but there also has been a more important development. Today, many discoveries and advances in cumulative knowledge are being made not by those who do primary research studies but by those who use meta-analysis to discover the latent meaning of existing research literatures. Today, behavioral or social scientists with the needed training and skills are making major original discoveries and contributions by mining the untapped veins of information in the accumulated research literatures.

The meta-analytic process of cleaning up and making sense of the research literature not only reveals the cumulative knowledge that is there but also provides clearer directions about what the remaining research needs are. That is, we also learn what kinds of primary research studies are needed next. However, some have raised the concern that meta-analysis may be killing the motivation and incentive to conduct primary research studies. Meta-analysis has clearly shown that no single primary study can ever resolve an issue or answer a question. Research findings are inherently probabilistic (Taveggia, 1974), and, therefore, the results of any single study could have occurred by chance. Only meta-analytic integration of findings across studies can control sampling error and other artifacts and provide a foundation for conclusions. And yet meta-analysis is not possible unless the needed primary studies are conducted. In new research areas, this potential problem is not of much concern. The first study conducted on a question contains 100% of the available research information. The second contains roughly 50%, and so on. Thus, the early studies in any area have a certain status. The 50th study, however, contains only about 2% of the available information, and the 100th, about 1%. Will we have difficulty motivating researchers to conduct the 50th or 100th study? If we do, we do not believe this will be due to meta-analysis. When the narrative review was the dominant method of research integration, reviewers did not base their conclusions on single studies but on multiple studies. So no researcher could reasonably hope then—as now—that his or her single study could decide an issue. In fact, meta-analysis represents an improvement for the primary researcher in one respect: All available relevant studies are included in a meta-analysis, and hence, every study has an effect. As we saw earlier, narrative reviewers often threw out most of the relevant studies and based their conclusions on a handful of their favorite studies.

Also, it should be noted that those who raise this question overlook a beneficial effect that meta-analysis has had: It prevents the diversion of valuable research resources into truly unneeded research studies. Meta-analysis applications have revealed that there are questions on

which additional research would waste scientifically and socially valuable resources. For example, already as of 1980, 882 studies based on a total sample of 70,935 had been conducted relating measures of perceptual speed to the job performance of clerical workers. Based on these studies, our meta-analytic estimate of this mean correlation is .47 and its $SD_\rho = .22$; Pearlman, Schmidt, & Hunter, 1980). For other abilities, there were often 200 to 300 cumulative studies. Clearly, further research on these relationships is not the best use of available resources.

If one or more new studies appear after a meta-analysis has been completed, how should the meta-analysis be updated to include these new studies? Schmidt and Raju (2007) showed that the optimal method is to recompute the meta-analysis including those studies (i.e., use the "medical method" of updating a meta-analysis rather than a Bayesian approach).

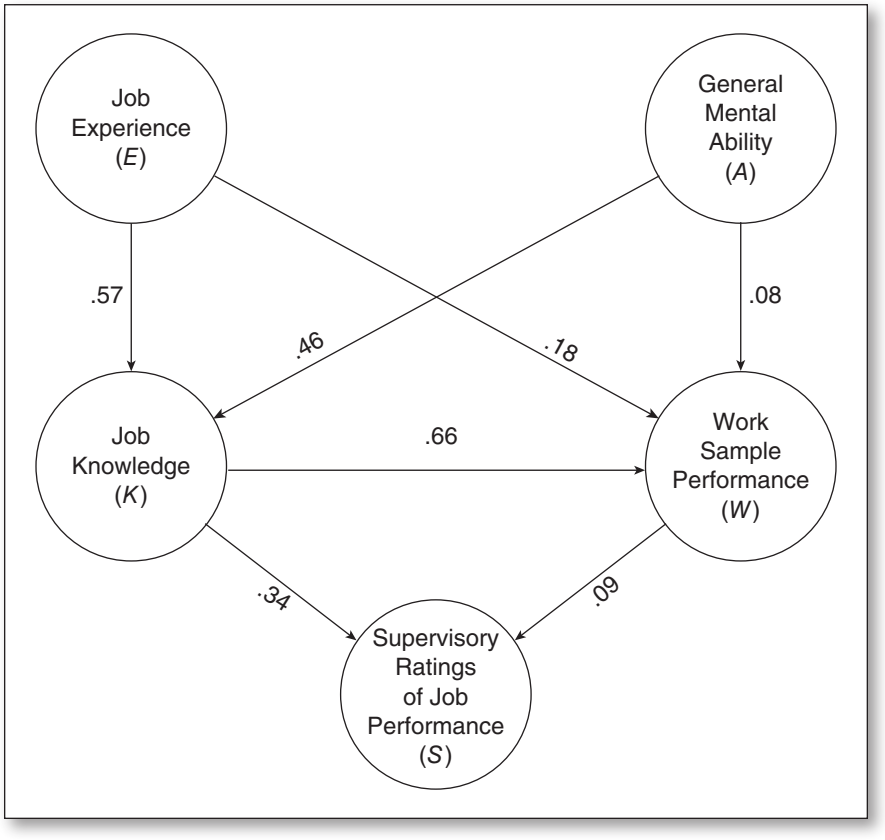
Role of Meta-Analysis in Theory Development

As noted earlier, the major task in the behavioral and social sciences, as in other sciences, is the development of theory. A good theory is simply a good explanation of the processes that actually take place in a phenomenon. For example, what actually happens when employees develop a high level of organizational commitment? Does job satisfaction develop first and then cause the development of commitment? If so, what causes job satisfaction to develop and how does it affect commitment? How do higher levels of mental ability cause higher levels of job performance? Only by increasing job knowledge? Or also by directly improving problem solving on the job? The social scientist is essentially a detective; his or her job is to find out why and how things happen the way they do. To construct theories, however, we must first know some of the basic facts, such as the empirical relationships among variables. These relationships are the building blocks of theory. For example, if we know there is a high and consistent positive population correlation between job satisfaction and organization commitment, this will send us in particular directions in developing our theory. If the correlation between these variables is very low and consistent, theory development will branch in different directions. If the relationship is highly variable across organizations and settings, we will be encouraged to advance interactive or moderator-based theories. Meta-analysis provides these empirical building blocks for theory. Meta-analytic findings tell us what it is that needs to be explained by the theory. Meta-analysis has been criticized because it does not directly generate or develop theory (Guzzo, Jackson, & Katzell, 1986). This is akin to criticizing word processors because they do not generate books on their own. The results of meta-analysis are indispensable for theory construction, but theory construction itself is a creative process distinct from meta-analysis.

As implied in the language used in our discussion, theories are causal explanations. The goal in every science is explanation, and explanation is always causal. In the behavioral and social sciences, the methods of path analysis (see, e.g., Hunter & Gerbing, 1982) and structural equation modeling (SEM) can be used to test causal theories when the data meet the assumptions of the method. The relationships revealed by meta-analysis—the empirical building blocks for theory—can be used in path analysis and SEM to test causal theories. Experimentally determined relationships can also be entered into path analyses along with observationally based relationships. It is only necessary to transform *d* values to correlations (see Chapter 7). Thus, path analyses can be "mixed." Path analysis and SEM cannot demonstrate that a theory is correct but can disconfirm a theory, that is, show that it is not correct. Path analysis can therefore be a powerful tool for reducing the number of theories that could possibly be consistent with the data, sometimes to a very small number, and sometimes to only one theory (Hunter, 1988). For an example, see Hunter (1983a). Every such reduction in the number of possible theories is an advance in understanding.

Application of path analysis or SEM requires either the correlations among the theoretically relevant variables (correlation matrix) or the covariances among the variables (variance-covariance matrix). Meta-analysis can be used to create correlation matrices for the variables of interest. Because each meta-analysis can estimate a different cell in the correlation matrix, it is possible to assemble the complete correlation matrix, even though no single study has included every one of the variables of interest (see, e.g., Viswesvaran & Ones, 1995). If these correlations are appropriately corrected for the biasing effects of artifacts such as measurement error, it is then possible to apply path analysis or SEM to test causal (or explanatory) theories (Cook et al., 1992, pp. 315–316). One example of this is Schmidt, Hunter, and Outerbridge (1986). This study used meta-analysis to assemble the correlations among the variables of general mental ability, job experience, job knowledge, work sample performance, and supervisory evaluations of job performance. (These correlations were homogeneous across studies.) The path analysis results are shown in Figure 1.1. This causal model fit the data quite well. As can be seen in Figure 1.1, both job experience and general mental ability exert strong causal influence on the acquisition of job knowledge, which, in turn, is the major cause of high performance on the job sample measure. The results also indicate that supervisors based their ratings more heavily on employees' job knowledge than on actual performance capabilities. This causal model (or theory) of job performance has since been supported in other studies (Schmidt & Hunter, 1992). Today the research literature contains many studies that use meta-analysis in this manner.

Figure 1.1 Path model and path coefficients.



Note: Adapted from "Impact of Job Experience and Ability on Job Knowledge, Work Sample Performance, and Supervisory Ratings of Job Performance," by F. L. Schmidt, J. E. Hunter, and A. N. Outerbridge, 1986, *Journal of Applied Psychology, 71,* 432–439. Reprinted by permission of the authors.

Becker (1989; 2009, chap. 20) and Becker and Schram (1994) discussed the possibilities of using meta-analysis in this manner. Becker (1992) used this approach in examining a model of the variables affecting the achievement in science of male and female high school and college students. In that case, there were insufficient studies available to obtain meta-analytic estimates of some of the needed correlations; nevertheless, progress was made and the information needed from future research was pinpointed. Becker (1996) and Shadish (1996) provided additional discussion. Although there are technical complexities that must be dealt with in using meta-analysis in this manner (Cook et al., 1992, pp. 328–330), it is a promising approach to accelerating cumulative knowledge in the social sciences. Technical questions related to this use of meta-analysis are discussed in Chapter 5.

Meta-Analysis in Industrial-Organizational Psychology

There have been numerous applications of meta-analysis in industrial-organizational (I/O) psychology. The most extensive and detailed application of meta-analysis in I/O psychology has been the study of the generalizability of the validities of employment selection procedures (Schmidt, 1988; Schmidt & Hunter, 1981, 1998). The findings have resulted in major changes in the field of personnel selection. Validity generalization research is described in more detail in Chapter 4.

Recent meta-analyses in organizational psychology have addressed a broad range of topics across different levels of analysis. At the levels of the organization or business unit, Harter, Schmidt, and Hayes (2002); Harter, Schmidt, Asplund, and Kilham (2010); and Whitman, Van Rooy, and Viswesvaran (2010) demonstrated that unit-level job satisfaction and employee engagement has positive, generalized effects on business unit financial performance and customer satisfaction. Another meta-analysis showed, across 83 different organizations, the generalized positive effects on job performance of the Productivity Measurement and Enhancement System (ProMES; Pritchard, Harrell, DiazGranadaos, & Guzman, 2008), a performance management system designed by organizational psychologists to provide workers and employees with quick and effective feedback on their performance. Meta-analyses of team research continue to be popular, with one meta-analysis summarizing how different teamwork processes affect team effectiveness (LePine, Piccolo, Jackson, Mathieu, & Saul, 2008).

Other meta-analyses focused on individuals as the unit of analysis. One such study examined the relationship between job turnover and the five-factor model (FFM) of personality and found that Emotional Stability is a generalized negative predictor of turnover (Zimmerman, 2008). Another study attempted to untangle the ambiguous causal relationship between job attitudes and job performance (Riketta, 2008). Other studies focused on multicultural and international issues. Dean, Roth, and Bobko (2008) meta-analytically examined ethnic and gender subgroup differences in assessment center ratings to show that gender differences in ratings from these evaluations are smaller than previously thought, but that some ethnic subgroup differences are larger than previously believed. Taras, Kirkman, and Steel (2010) showed that Hofstede's (1980) cultural value dimensions had validity in predicting (in decreasing order) individual emotions, attitudes, behaviors, and job performance. Geyskens, Krishnan, Steenkamp, and Cunha (2009) presented an extensive examination of the use of meta-analysis in management-related research.

Older examples also span a variety of topics and units of analysis. C. D. Fisher and Gitelson's (1983) meta-analysis examined the negative and positive correlates of conflict and ambiguity for members' roles in teams. Meta-analyses of leadership performance were also popular. An

example is a meta-analytic test of Fiedler's contingency theory of leadership, a dominant theory of leadership at the time (L. H. Peters, Harthe, & Pohlman, 1985). Other meta-analyses studied questions about attitudes and beliefs, such as the relatively low accuracy of self-ratings of ability and skill (Mabe & West, 1982) and the negative relationship between job satisfaction and absenteeism (Terborg & Lee, 1982). Other studies focused on more specific interventions and assessments, such as the small but positive effect of realistic job previews in reducing subsequent employee turnover (Premack & Wanous, 1985); the positive, generalizable validity of the LSAT for predicting performance in law school (Linn, Harnisch, & Dunbar, 1981a); and the limited abilities of financial analysts to predict stock growth (Coggin & Hunter, 1983). In short, researchers have pursued and continue to pursue meta-analytic studies across a wide variety of subjects and continue to recognize psychometric meta-analysis as an important research tool.

Additional examples of influential meta-analyses are found in the literature on managerial training. Burke and Day's (1986) meta-analysis on the effectiveness of managerial training prompted a subsequent stream of meta-analytic research on management training, including D. B. Collins and Holton (2004); Taylor, Russ-Eft, and Taylor (2009); and Powell and Yalcin (2010). The conclusions from these studies have repeatedly been that management training programs can be effective in nearly all situations at changing particular behaviors and for the acquisition of knowledge, particularly in the areas of time management and human relations skills. Other meta-analyses have assessed the results of training across a variety of organizational contexts. For example, a recent meta-analysis examined how combinations of training content, trainee attributes, and trainees' affective reactions to training influence the outcomes of a training program (Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008). Other meta-analytic studies investigated the relationships among different training criteria such as behavior, learning, and performance (Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997), or how trainees use their training in applied settings and share knowledge from training with others (Arthur, Bennett, Edens, & Bell, 2003). All three meta-analyses have had powerful impacts on traditional models of learning and training in I/O psychology, resulting in updates to D. L. Kirkpatrick's (2000) widely used model of learning, a demonstration of the effectiveness of lectures as a form of training, and reconsideration of the value of affective reactions to training.

Colquitt, LePine, and Noe (2000) used meta-analysis to further expand on Alliger et al.'s (1997) work and examine how individual differences, situational factors, and job career factors influenced an individual's motivation during training and the subsequent outcomes of this motivation. Colquitt et al.'s (2000) work established the importance of motivation as a critical factor in determining the efficacy of training, how individuals

transfer learning from training to their work performance, and how effective they are in sharing training knowledge with others. In a more recent study of the impact of motivation, Payne, Youngcourt, and Beaubien (2007) have used meta-analytic methods to calibrate the impact of goal orientation (a psychological variable reflecting the motivation of individuals to learn versus the motivation to perform well in front of others) on training outcomes. These two meta-analyses have influenced how researchers conduct subsequent primary studies and created new opportunities for researchers to examine the role of motivation in training. In short, meta-analysis has affected research on management training, motivation in training, and training in general by demonstrating the validity and value of training to organizations across the board.

Meta-analysis has also been used extensively in the leadership literature, a popular topic of study in I/O psychology and organizational behavior (OB). Meta-analysis has provided some clarity on a difficult subject, as well as changing how studies were conducted in light of the findings of the meta-analyses. Judge, Colbert, and Ilies (2004) reported a relatively low relationship between general cognitive ability and leadership, noting that this relationship is weaker than expected based on earlier qualitative reviews. Similar work was done demonstrating that leadership cannot be explained solely as a result of personality traits in the popular FFM framework (Judge, Bono, Ilies, & Gerhardt, 2002). Several years later, this line of research inspired a group of researchers to consider conceptual models of positive ("bright-side") and negative ("dark-side") personality traits in leaders linked to leadership emergence and efficacy (Judge, Piccolo & Kosalka, 2009). Finally, multiple meta-analyses have been used to show the validity, uniqueness, and importance of transformational (charismatic) leadership, a specific set of leadership behaviors highly motivating to employees and a central topic of research in the leadership literature (Bono & Judge, 2004; Eagly, Johannsen-Schmidt, & van Engen, 2003; Eagly, Karau, & Makhijani, 1995). Applications of meta-analysis have produced important changes in research conclusions, fundamental changes in existing theoretical paradigms, and the development of new lines of research.

Wider Impact of Meta-Analysis on Psychology

Some have viewed meta-analysis as merely a set of improved methods for doing literature reviews. Meta-analysis is actually more than that. By quantitatively comparing findings across diverse studies, meta-analysis can discover new knowledge not inferable from any individual study and can sometimes answer questions that were never addressed in any of the individual studies contained in the meta-analysis. For example, no individual study may have compared the effectiveness of a training program for people of higher and lower mental ability; by comparing mean *d* value statistics across different

groups of studies, however, meta-analysis can reveal this difference. That is, moderator variables (interactions) never studied in any individual study can be revealed by meta-analysis, greatly facilitating the development of cumulative knowledge. M. E. Chan and Arvey (2012) provide an analysis of the positive impact of meta-analysis on the development of knowledge in psychology and the social sciences. Richard, Bond, and Stokes-Zoota (2003) review the impact of 322 meta-analyses in social psychology. And Dieckmann, Malle, and Bodner (2009) provide an assessment of the use of meta-analysis in several areas of psychological research. Carlson and Ji (2011) presented an analysis of how meta-analytic studies are used and cited in the wider psychological literature. They found that the frequency of citations to meta-analyses (vs. primary studies) has been rapidly increasing in recent years.

Even though it is much more than that, meta-analysis is indeed an improved method for synthesizing or integrating the research literature. The premier review journal in psychology is *Psychological Bulletin.* Over time since 1980, a steadily increasing percentage of the reviews published in this journal have been meta-analyses and a steadily decreasing percentage have been traditional narrative subjective reviews. It is not uncommon for narrative review manuscripts to be returned by editors to the authors with the request that meta-analysis be applied to the studies reviewed (Cooper, 2003). Most of the remaining narrative reviews published today in *Psychological Bulletin* focus on research literature that is not well enough developed to be amenable to quantitative treatment.

Although a movement toward change began about 2010, most of the meta-analyses that have appeared in *Psychological Bulletin* have employed fixed effects methods, resulting in many cases in overstatement of the precision of the meta-analysis findings (Schmidt, Oh, & Hayes, 2009). (See Chapters 5 and 8 for a discussion of fixed vs. random meta-analysis models; confidence intervals are too narrow when the fixed effect model is used.) Despite this fact, these meta-analyses produced findings and conclusions that are more accurate than those produced by the traditional narrative subjective method. Many other journals have shown the same increase over time in the number of meta-analyses published. Many of these journals had traditionally published only individual empirical studies and had rarely published reviews up until the advent of meta-analysis in the late 1970s. These journals began publishing meta-analyses because meta-analyses came to be viewed not as "mere reviews" but as a form of empirical research. As a result of this change, the quality and accuracy of conclusions from the research literature improved in a wide variety of journals and in a corresponding variety of research areas in psychology. This improvement in the quality of conclusions from the research literature has expedited theory development in many areas in psychology.

The impact of meta-analysis on psychology textbooks has been positive and dramatic. Textbooks are important because their function is to summarize the state of cumulative knowledge in a given field. Most

people—students and others—acquire most of their knowledge about psychological theory and findings from their reading of textbooks. Prior to meta-analysis, textbook authors faced with hundreds of conflicting studies on a single question subjectively and arbitrarily selected a small number of their preferred studies from the literature and based the textbook conclusions on only those few studies. Today, most textbook authors base their conclusions on meta-analysis findings (Myers, 1991), making their conclusions and their textbooks much more accurate. We cannot overemphasize the importance of this development in advancing cumulative knowledge in psychology.

Because multiple studies are needed to solve the problem of sampling error, it is critical to ensure the availability of all studies on each topic. A major problem is that many good replication articles are rejected by the primary research journals. Journals currently put excessive weight on surprising and novel findings in evaluating studies and often fail to consider either sampling error or other important technical problems such as measurement error. Many journals will not even consider "mere replication studies" or "mere measurement studies." Many persistent authors eventually publish such studies in journals with lower prestige, but they must endure many letters of rejection, and publication is often delayed for a long period of time. Problems of this sort, including the general issue of publication bias, are discussed in more detail in Chapter 13.

To us, this clearly indicates that we need a new type of journal—whether hard copy or electronic—that systematically archives all studies that will be needed for later meta-analyses. The American Psychological Association's Experimental Publication System in the early 1970s was an attempt in this direction. However, at that time, the need subsequently created by meta-analysis did not yet exist; the system apparently met no real need at that time and hence was discontinued. Today, the need is so great that failure to have such a journal system in place is retarding our efforts to reach our full potential in creating cumulative knowledge in psychology and the social sciences.

In view of the large number of meta-analyses available in the psychology and social sciences literatures, some readers may wonder why the examples we use in this book to illustrate meta-analysis principles and methods do not employ data from those meta-analyses. The primary reason is that the amount of data (the number of correlations or *d* statistics) is usually so large as to result in cumbersome examples. For pedagogical reasons, we have generally employed examples consisting of small numbers of studies in which the data are hypothetical. As explained in the following chapters, meta-analyses based on such small numbers of studies would not ordinarily yield results that would be optimally stable. (We discuss second-order sampling error in Chapter 9.) However, such examples provide the means to simply and clearly illustrate the principles and methods of meta-analysis, and we believe this is the crucial consideration.

Impact of Meta-Analysis Outside Psychology

IMPACT IN BIOMEDICAL RESEARCH

The impact of meta-analysis may be even greater in biomedical research than in the behavioral and social sciences (Hunt, 1997, chap. 4). Hundreds of meta-analyses have been published in leading medical research journals such as the *New England Journal of Medicine* and the *Journal of the American Medical Association*. Already as of 1995, the medical literature contained between 962 and 1,411 meta-analyses, depending on the method of counting (Moher & Olkin, 1995). This number is much greater today. In medical research, the preferred study is the randomized controlled trial (RCT), in which participants are assigned randomly to receive either the treatment or a placebo, with the researchers being blind as to which treatment the participants are receiving. Despite the strengths of this research design, it is usually the case that different RCTs on the same treatment obtain conflicting results. This is partly because the effect sizes are often small and partly because (contrary perhaps to widespread perceptions) RCTs are often based on small sample sizes. In addition, the problem of information overload is even greater in medicine than in the social sciences; more than a million medical research studies are published every year. No practitioner can possibly keep up with the medical literature in his or her area.

The leader in introducing meta-analysis to medical research was Thomas Chalmers. In addition to being a researcher, Chalmers was also a practicing internal medicine physician who became frustrated with the inability of the vast, scattered, and unfocused medical research literature to provide guidance to practitioners. Starting in the mid-1970s, Chalmers developed his initial meta-analysis methods independently of those developed in the social and behavioral sciences. Despite being well conducted, his initial meta-analyses were not well accepted by medical researchers, who were critical of the concept of meta-analysis. In response, he and his associates developed "sequential meta-analysis"—a technique that reveals the date by which enough information had become available to show conclusively that a treatment was effective. Suppose, for example, that the first RCT for a particular drug had been conducted in 1975 but had a wide confidence interval, one that spans zero effect. Now, suppose three more studies had been conducted in 1976, providing a total of four studies to be meta-analyzed—and the confidence interval for the meta-analytic mean of these studies is still wide and still includes 0. Now, suppose five more RCTs had been conducted in 1977, providing nine studies for a meta-analysis up to this date. Now, if that meta-analysis yields a confidence interval that excludes 0, then we conclude that, given the use of meta-analysis, enough information was already available in 1977 to begin using this drug. On the basis of their meta-analysis findings and statistics on the disease, Chalmers

and his associates then computed how many lives would have been saved to date had use of the drug begun in 1977. It turned out that, considered across different treatments, diseases, and areas of medical practice, a very large number of lives would have been saved had medical research historically relied on meta-analysis. The resulting article (Antman, Lau, Kupelnick, Mosteller, & Chalmers, 1992) is widely considered the most important and influential meta-analysis study ever published in medicine. It was even reported and discussed widely in the popular press (for example, the *New York Times* Science Section). It assured a major role for meta-analysis in medical research from that point on (Hunt, 1997, chap. 4).

Chalmers was also one of the driving forces behind the establishment of the Cochrane Collaboration, an organization that applies sequential meta-analysis in medical research in real time. This group conducts meta-analyses in a wide variety of medical research areas—and then updates each meta-analysis as new RCTs become available. That is, when a new RCT becomes available, the meta-analysis is rerun with the new RCT included. Hence, each meta-analysis is always current. The results of these updated meta-analyses are available on the Internet to researchers and medical practitioners around the world. It is likely that this effort has saved hundreds of thousands of lives by improving medical decision making. The Cochrane Collaboration website is www.cochrane.org. Another major early contributor to the development of meta-analysis methods for medical research is Richard Peto (1987). Later in the book, we discuss in detail methods of correcting for the biasing effects of measurement error, using methods from psychometric theory. Completely independently of these methods from psychometric theory, Peto developed different but equivalent methods of correcting for measurement error in medical research. For an application of these methods, see MacMahon et al. (1990). Examples of the use in medical research of the methods presented in this book include Fountoulakis, Conda, Vieta, and Schmidt (2009) and Gardner, Frantz, and Schmidt (1999). Problems related to publication bias and research fraud in the biomedical sciences are discussed in Chapter 13.

IMPACT IN OTHER DISCIPLINES

Meta-analysis has also become important in research in finance, marketing, sociology, ecology, and even wildlife management. Other areas in which meta-analysis is now being used include (Hafdahl, 2012) higher education, biological psychiatry, physical therapy, nursing practice, neuroscience, cardiology, forestry, and occupational counseling. An example of meta-analysis use in criminal justice is provided in Gendreau and Smith (2007). In fact, today it is difficult to find a research area in which meta-analysis is unknown. In the broad areas of education and social policy, the Campbell Collaboration is attempting to do for the social sciences

what the Cochrane Collaboration (on which it is modeled) has done for medical practice (Rothstein, 2003; Rothstein, McDaniel, & Borenstein, 2001). Among the social sciences, perhaps the last to assign an important role to meta-analysis has been economics. However, meta-analysis has recently become important in economics, too (see, e.g., T. D. Stanley, 1998, 2001; T. D. Stanley & Jarrell, 1989, 1998). Another example in economics is the meta-analysis by Harmon, Oosterbeck, and Walker (2000); their study meta-analyzed a large number of studies on the financial returns to education and found an overall average rate of return of 6.5%. They also found that returns to education have fallen since the 1960s. There now is a doctoral program in meta-analysis in economics (www.feweb.vu.nl/re/Master-Point/). In 2008, an international conference on meta-analysis usage in economics was held in Nancy, France (Nancy-Universite, 2008). Meta-analysis is now also used in political science (see, e.g., Pinello, 1999).

Meta-Analysis and Social Policy

By providing the best available empirically based answers to socially important questions, meta-analysis can influence public policy making (Hoffert, 1997; Hunter & Schmidt, 1996). This can be true for any public policy question for which there is a relevant research literature—which today includes most policy questions. Examples range from the Head Start program to binary chemical weapons (Hunt, 1997, chap. 6). The purpose of the Campbell Collaboration, described previously, is specifically to provide policy-relevant information to policy makers in governments and other organizations by applying meta-analysis to policy-relevant research literatures on social experiments. For more than 20 years, the U.S. General Accounting Office (GAO; now renamed the General Accountability Office), a research and evaluation arm of the U.S. Congress, has used meta-analysis to provide answers to questions posed by senators and representatives. For example, Hunt (1997, chap. 6) described how a GAO meta-analysis of the effects of the Women, Infants, and Children (WIC) program, a federal nutritional program for poor pregnant women, apparently changed the mind of Senator Jesse Helms and made him a supporter of the program. The meta-analysis found evidence that the program reduced the frequency of low-birth-weight babies by about 10%.

This meta-analysis was presented to Senator Helms by Eleanor Chelimsky, for years the director of the GAO's Division of Program Evaluation and Methodology. In that position, she pioneered the use of meta-analysis at GAO. Chelimsky (1994) stated that meta-analysis has proven to be an excellent way to provide Congress with the widest variety of research results that can hold up under close scrutiny under the time pressures imposed by Congress. She stated that the GAO has found that meta-analysis reveals both what is known and what is not known in a

given topic area and distinguishes between fact and opinion "without being confrontational." One application she cited as an example was a meta-analysis of studies on the merits of producing binary chemical weapons (nerve gas in which the two key ingredients are kept separate for safety until the gas is to be used). The meta-analysis did not support the production of such weapons. This was not what the Department of Defense (DOD) wanted to hear, and the DOD disputed the methodology and the results. The methodology held up under close scrutiny, however, and in the end Congress eliminated funds for these binary weapons.

By law, it is the responsibility of the GAO to provide policy-relevant research information to Congress. So the adoption of meta-analysis by the GAO is a clear example of the impact that meta-analysis can have on public policy. Although most policy decisions probably depend as much on political as on scientific considerations, it is possible for scientific considerations to have an impact with the aid of meta-analysis (Hoffert, 1997). Cordray and Morphy (2009) provide an extended discussion of the role of meta-analysis in the formulation of public policy. They emphasize the point that special care must be taken to ensure the objectivity of policy-related meta-analyses. In particular, they examine the meta-analysis conducted by the Environmental Protection Agency on the health effects of secondhand smoke as an example of a biased meta-analysis. This example is important because numerous laws were passed against environmental tobacco smoke based a faulty meta-analytic conclusion that it was harmful to health. Cordray and Morphy (2009) outline the steps necessary to ensure objectivity in meta-analyses.

Meta-Analysis and Theories of Data and Theories of Knowledge

Every method of meta-analysis is of necessity based on a theory of data. It is this theory (or understanding of data) that determines the nature of the resulting meta-analysis methods. A complete theory of data includes an understanding of sampling error, measurement error, biased sampling (range restriction and range enhancement), dichotomization and its effects, data errors, and other causal factors that distort the raw data results we see in research studies. Once a theoretical understanding of how these factors affect data is developed, it becomes possible to develop methods for correcting for their effects. The necessity of doing so is presented in detail in Schmidt, Le, and Oh (2009). In the language of psychometrics, the first process—the process by which these factors (artifacts) influence data—is modeled as the attenuation model. The second process—the process of correcting for these artifact-induced biases—is called the disattenuation model. If the theory of data on which a method of meta-analysis is based is incomplete, that method will fail to correct for some or all of these

artifacts and will thus produce biased results. For example, a theory of data that fails to recognize measurement error will lead to methods of meta-analysis that do not correct for measurement error. Such methods will then perforce produce biased meta-analysis results. Most current methods of meta-analysis do not, in fact, correct for measurement error, as noted in Chapter 11. But in research methodology, the thrust is always in the direction of increased accuracy, so eventually methods for meta-analysis that do not correct for study artifacts that distort empirical findings will have to incorporate these corrections. This has already happened to some extent in that some users of these methods have "appended" these corrections to those methods (e.g., Aguinis, Sturman, & Pierce, 2008; Hall & Brannick, 2002). One's theory of data is also part of one's theory of knowledge or theory of epistemology. Epistemology is concerned with the ways in which we can attain correct knowledge. One requirement of an effective epistemology in empirical research is proper correction for the artifacts that distort the empirical data.

Sampling error and measurement error have a unique status among the statistical and measurement artifacts with which meta-analysis must deal: They are *always* present in all real data. Other artifacts, such as range restriction, artificial dichotomization of continuous variables, or data transcription errors, may be absent in a particular set of studies being subjected to meta-analysis. There is always sampling error, however, because sample sizes are never infinite. Likewise, there is always measurement error, because there are no perfectly reliable measures. In fact, as we will see in subsequent chapters, it is the requirement of dealing simultaneously with both sampling error and measurement error that makes even relatively simple meta-analyses sometimes seem complicated. We are used to dealing with these two types of errors separately. For example, when psychometric texts (e.g., Lord & Novick, 1968; Nunnally & Bernstein, 1994) discuss measurement error, they assume an infinite (or very large) sample size, so that the focus of attention can be on measurement error alone, with no need to deal simultaneously with sampling error. Similarly, when statistics texts discuss sampling error, they implicitly assume perfect reliability (the absence of measurement error), so that they and the reader can focus solely on sampling error. Both assumptions are highly unrealistic because all real data simultaneously contain both types of errors. It is admittedly more complicated to deal with both types of errors simultaneously, yet this is what meta-analysis must do to be successful (see, e.g., Cook et al., 1992, pp. 315–316, 325–328; Matt & Cook, 2009, chap. 28).

The question of what theory of data (and therefore of knowledge) underlies a method of meta-analysis is strongly related to the question of what the general purpose of meta-analysis is. Glass (1976, 1977) stated that the purpose is simply to summarize and describe the studies in the research literature. As we will see in this book, our view (the alternative view) is that the purpose is to estimate as accurately as possible the

construct-level relationships in the population (i.e., to estimate population values or parameters), because these are the relationships of scientific interest. This is an entirely different task; this is the task of estimating what the findings would have been if all studies had been conducted perfectly (i.e., with no methodological limitations). Doing this requires correction for sampling error, measurement error, and other artifacts (if present) that distort study results. Simply describing the contents of studies in the literature requires no such corrections and does not allow estimation of the parameters of scientific interest.

Rubin (1990, 1992) critiqued the common, descriptive concept of the purpose of meta-analysis and proposed the alternative offered in this book and previous editions of this book. He stated that, as scientists, we are not really interested in the population of imperfect studies per se, and hence an accurate description or summary of the results of these studies is not really important. Instead, he argued that the goal of meta-analysis should be to estimate the true effects or relationships—defined as "results that would be obtained in an infinitely large, perfectly designed study or sequence of such studies." According to Rubin (1990),

> Under this view, we really do not care *scientifically* about summarizing this finite population (of observed studies). We really care about the underlying scientific process—the underlying process that is generating these outcomes that we happen to see—that we, as fallible researchers, are trying to glimpse through the opaque window of imperfect empirical studies. (p. 157)

This is an excellent summary of the purpose of meta-analysis as we see it and as embodied in the methods presented in this book.

Conclusion

Until recently, the psychological research literature was conflicting and contradictory. As the number of studies on each particular question became larger and larger, this situation became increasingly frustrating and intolerable. This situation stemmed from reliance on defective procedures for achieving cumulative knowledge: the statistical significance test in individual primary studies in combination with the narrative subjective review of the research literature. Meta-analysis principles have now correctly diagnosed this problem and have provided the solution. In area after area, meta-analytic findings have shown that there is much less conflict between different studies than had been believed; that coherent, useful, and generalizable conclusions can be drawn from the research literature; and that cumulative knowledge is possible in psychology and the social sciences. These methods have also been adopted in other areas such as

medical research. A prominent medical researcher, Thomas Chalmers (as cited in Mann, 1990), has stated, "[Meta-analysis] is going to revolutionize how the sciences, especially medicine, handle data. And it is going to be the way many arguments will be ended" (p. 478). In concluding his oft-cited review of meta-analysis methods, Bangert-Drowns (1986) stated,

> 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)