# What Do Data Really Mean?

## *Research Findings, Meta-Analysis, and Cumulative Knowledge in Psychology*

Frank L. Schmidt *College of Business, University of Iowa*

*How should data be interpreted to optimize the possibilities for cumulative scientific knowledge? Many believe that traditional data interpretation procedures based on statistical significance tests reduce the impact of sampling error on scientific inference. Meta-analysis shows that the significance test actually obscures underlying regularities and processes in individual studies and in research literatures, leading to systematically erroneous conclusions. Meta-analysis methods can solve these problems--and have done so in some areas. However, meta-analysis represents more than merely a change in methods of data analysis. It requires major changes in the way psychologists view the general research process. Views of the scientific value of the individual empirical study, the current reward structure in research, and even the fundamental nature of scientific discovery may change.*

Many today are disappointed in the progress that psychology has made in this century. Numerous explanations have been advanced for why progress has been slow. Faulty philosophy of science assumptions have been cited (e.g., excessive emphasis on logical positivism; Glymour, 1980; Schlagel, 1979; Suppe, 1977; Toulmin, 1979). The negative influence of behaviorism has been discussed and debated (e.g., Koch, 1964; Mackenzie, 1977; McKeachie, 1976; Schmidt, Hunter, & Pearlman, 1981). This article focuses on a reason that has been less frequently discussed: the methods psychologists (and other social scientists) have traditionally used to analyze and interpret their data—both in individual studies and in research literatures. This article advances three arguments: (a) Traditional data analysis and interpretation procedures based on statistical significance tests militate against the discovery of the underlying regularities and relationships that are the foundation for scientific progress; (b) meta-analysis methods can solve this problem—and they have already begun to do so in some areas; and (c) meta-analysis is not merely a new way of doing literature reviews. It is a new way of thinking about the meaning of data, requiring that we change our views of the individual empirical study and perhaps even our views of the basic nature of scientific discovery.

## Traditional Methods Versus Meta-Analysis

Psychology and the social sciences have traditionally relied heavily on the statistical significance test in interpreting the meaning of data, both in individual studies and in research literatures. Following the lead of Fisher (1932), null hypothesis significance testing has been the dominant data analysis procedure. The prevailing decision rule, as Oakes (1986) has demonstrated empirically, has been this: If the statistic (*t*, *F*, etc.) is significant, there is an effect (or a relation); if it is not significant, then there is no effect (or relation). (See also Cohen, 1990, 1992.) These prevailing interpretational procedures have focused heavily on the control of Type I errors, with little attention being paid to the control of Type II errors. A Type I error (alpha error) consists of concluding there is a relation or an effect when there is not. A Type II error (beta error) consists of the opposite—concluding there is no relation or effect when there is. Alpha levels have been controlled at the .05 or .01 levels, but beta levels have by default been allowed to climb to high levels, often in the 50%–80% range (Cohen, 1962, 1988, 1990; Schmidt, Hunter, & Urry, 1976). To illustrate this, let us look at an example from a hypothetical area of experimental psychology.

Suppose the research question is the effect of a certain drug on learning, and suppose the actual effect of a particular dosage is .50 of a standard deviation increase in the amount learned. An effect size of .50, considered medium size by Cohen (1988), corresponds to the difference between the 50th and 69th percentiles in a normal distribution. With an effect size of this magnitude, 69% of the experimental group would exceed the mean of the control group if both were normally distributed. Many reviews of various literatures have found relations of this general magnitude (Hunter & Schmidt, 1990b). Now suppose a large number of studies are conducted on this dosage, each with 15 rats in the experimental group and 15 in the control group.

Figure 1 shows the distribution of effect sizes (*d* values) expected under the null hypothesis. All variability around the mean value of zero is due to sampling error. To be significant at the .05 level (with a one-tailed test), the effect size must be .62 or larger. If the null hypothesis is true, only 5% will be that large or larger. In analyzing

Frederick A. King served as action editor for this article.

An earlier and shorter version was presented as an invited address at the Second Annual Convention of the American Psychological Society, Dallas, TX, June 8, 1990.

I would like to thank John Hunter, Deniz Ones, Michael Judiesch, and C. Viswesvaran for helpful comments on a draft of this article. Any errors remaining are mine.

Correspondence concerning this article should be addressed to Frank L. Schmidt, Department of Management and Organization, College of Business, University of Iowa, Iowa City, IA 52242.



 0003-066X/92/$2.00
Vol. 47, No. 10, 1173-1181

their data, researchers typically focus only on the information in Figure 1. Most believe that their significance test limits the probability of an error to 5%.

Actually, in this example the probability of a Type I error is zero—not 5%. Because the actual effect size is always .50, the null hypothesis is always false, and therefore there is *no possibility* of a Type I error. One cannot *falsely* conclude there is an effect when in fact there is an effect. When the null hypothesis is false, the only kind of error that *can* occur is a Type II error—failure to detect the effect that is present. The only type of error that can occur is the type that is not controlled. A reviewer of this article asked for more elaboration on this point. In any given study, the null hypothesis is either false or not false in the population in question. If the null is false (as in this example), there is a nonzero effect in the population. A Type I error consists of concluding there is an effect when no effect exists. Because an effect does, in fact, exist here, it is not possible to make a Type I error (Hunter & Schmidt, 1990b, pp. 23–31).

Figure 2 shows not only the irrelevant null distribution but also the *actual* distribution of observed effect sizes across these studies. The mean of this distribution is the true value of .50, but because of sampling error there is substantial variation in observed effect sizes. Again, to be significant, the effect size must be .62 or larger. Only 37% of studies conducted will obtain a significant effect size; thus statistical power for these studies is only .37. That is, the true effect size of the drug is *always* .50; it is *never* zero. Yet it is only detected as significant in 37% of the studies. The error rate in this research literature is 63%, not 5% as many would mistakenly believe.

**Figure 1**
*The Null Distribution of d Values in a Series of Experiments*

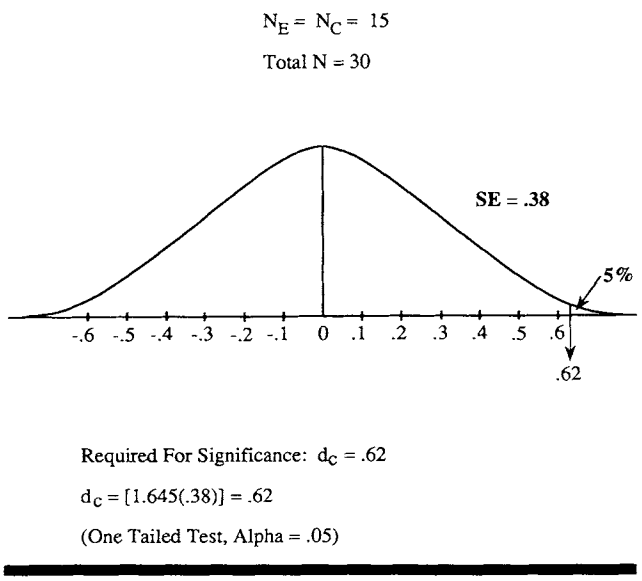
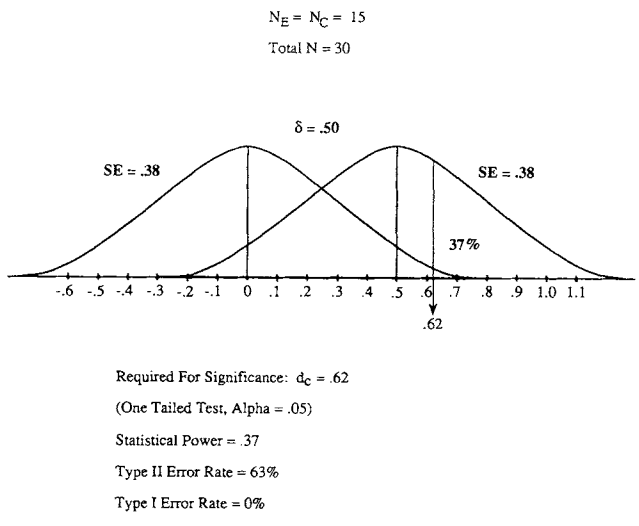



**Figure 2**
*Statistical Power in a Series of Experiments*




Most researchers in experimental psychology would traditionally have used analysis of variance (ANOVA) to analyze these data. This means the significance test would be two-tailed rather than one-tailed, as in our example. With a two-tailed test (i.e., one-way ANOVA), statistical power is even lower—.26 instead of .37. The Type II error rate (and hence the overall error rate) would be 74%. Also, this example assumes use of a *z* test; any researchers not using ANOVA would probably use a *t* test. For a one-tailed *t* test with alpha equal to .05 and degrees of freedom of 28, the effect size (*d* value) must be .65 to be significant. (The *t* value must be at least 1.70, instead of the 1.645 required for the *z* test.) With the *t* test, statistical power would also be lower—.35 instead of .37. Thus both commonly employed alternative significance tests would yield even lower statistical power.

Furthermore, the studies that *are* significant yield distorted estimates of effect sizes. The true effect size is always .50; all departures from .50 are due solely to sampling error. But the minimum value required for significance is .62. The obtained *d* value must be .12 above its true value—24% larger than its real value—to be significant. The *average* of the significant *d* values is .89, which is 78% larger than the true value.

In any study in this example that by chance yields *the correct value* of .50, the conclusion under the prevailing decision rule will be that there is no relationship. That is, it is only the studies that are by chance quite *inaccurate* that lead to the correct conclusion that a relationship exists.

How would this body of studies be interpreted as a research literature? There are two interpretations that would have traditionally been frequently accepted. The first is based on the traditional voting method (Hedges & Olkin, 1980; Light & Smith, 1971). Using this method, one would note that 63% of the studies found "no rela-

**Figure 3**
*Meta-Analysis of the Drug Studies of Figure 1*

**I. Compute Actual Variance of Effect Sizes ($S^2_\delta$)**

1. $S^2_d = .1444$ (Observed Variance of d Values)

2. $S^2_e = .1444$ (Variance Predicted from Sampling Error)

3. $S^2_\delta = S^2_d - S^2_e$

4. $S^2_\delta = .1444 - .1444 = 0$ (True Variance of $\delta$ Values)

**II. Compute Mean Effect Size ($\bar{\delta}$)**

1. $\bar{d} = .50$ (Mean Observed d Value)

2. $\bar{\delta} = .50$

3. $SD_\delta = 0$

**III. Conclusion: There is only one effect size, and its value is .50 standard deviation.**

tionship." Because this is a majority of the studies, the conclusion would be that no relation exists. It is easy to see that this conclusion is false, yet many reviews in the past were conducted in just this manner (Hedges & Olkin, 1980). The second interpretation reads as follows: In 63% of the studies, the drug had no effect. However, in 37% of the studies, the drug did have an effect. Research is needed to identify the moderator variables (interactions) that cause the drug to have an effect in some studies but not in others. For example, perhaps the strain of rat used or the mode of injecting the drug affects study outcomes. This interpretation is also completely erroneous.

How would meta-analysis interpret these studies? Different approaches to meta-analysis use somewhat different quantitative procedures (Bangert-Drowns, 1986; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Hunter & Schmidt, 1990b; Rosenthal, 1984, 1991). I illustrate this example using the methods presented by Hunter et al. (1982) and Hunter and Schmidt (1990b). Figure 3 shows that meta-analysis reaches *the correct conclusion.* Meta-analysis first computes the variance of the observed *d* values. Next, it uses the standard formula for the sampling error variance of *d* values (e.g., see Hunter & Schmidt, 1990b, chap. 7) to determine how much variance would be expected in observed *d* values from sampling error alone. The amount of real variance in population *d* values ($\delta$ values) is estimated as the difference between the two. In our example, this difference is zero, indicating correctly that there is only one population value. This single population value is estimated as the average observed value, which is .50 here, the correct value. If the number of studies is large, the average *d* value will be close to the true (population) value because sampling errors are random and hence average out to zero. Note that these meta-analysis methods do *not* rely on statistical significance tests. Only effect sizes are used, and significance tests are not used in analyzing the effect sizes.

The data in this example are hypothetical. However, if one accepts the validity of basic statistical formulas for sampling error, one will have no reservations about this example. But the same principles do apply to real data, as shown next by an example from personnel selection. Table 1 shows observed validity coefficients from 21 studies of a single clerical test and a single measure of job performance. Each study has $N = 68$ (the median $N$ in the literature in personnel psychology), and every study is a random draw (without replacement) from a single larger validity study with 1,428 subjects. The correlation in the large study (uncorrected for measurement error, range restriction, or other artifacts) is .22 (Schmidt, Ocasio, Hillery, & Hunter, 1985).

The validity is significant in eight (or 38%) of these studies, for an error rate of 62%. The traditional conclusion would be that this test is valid in 38% of the organizations, and invalid in the rest, and that in organizations where it is valid, its mean observed validity is .33 (which is 50% larger than its real value). Meta-analysis of these validities indicates that the mean is .22 and all variance in the coefficients is due solely to sampling error. The meta-analysis conclusions are correct; the traditional conclusions are false.

Reliance on statistical significance testing in psychology and the social sciences has long led to frequent

Table 1
*21 Validity Studies (N = 68 Each)*

| Study | Observed validity | Study | Observed validity |
|---|---|---|---|
| 1 | .04 | 12 | .11 |
| 2 | .14 | 13 | .21 |
| 3 | .31* | 14 | .37* |
| 4 | .12 | 15 | .14 |
| 5 | .38* | 16 | .29* |
| 6 | .27* | 17 | .26* |
| 7 | .15 | 18 | .17 |
| 8 | .36* | 19 | .39* |
| 9 | .20 | 20 | .22 |
| 10 | .02 | 21 | .21 |
| 11 | .23 | | |

* $p < .05$, two-tailed.

serious errors in interpreting the meaning of data (Hunter & Schmidt, 1990b, pp. 29–42, 483–484), errors that have systematically retarded the growth of cumulative knowledge. Yet it has been remarkably difficult to wean social science researchers away from their entrancement with significance testing (Oakes, 1986). One can only hope that lessons from meta-analysis will finally stimulate change. But time after time, even in recent years, I and my research colleagues have seen researchers who have been taught to understand the deceptiveness of significance testing sink back, in weak moments, into the nearly incurable habit of reliance on significance testing. I have occasionally done it myself. The psychology of addiction to significance testing would be a fascinating research area. This addiction must be overcome; but for most researchers, it will not be easy.

In these examples, the only type of error that is controlled—Type I error—is the type that *cannot* occur. It is likely that, in most areas of research, as time goes by and researchers gain a better and better understanding of the processes they are studying, it is less and less frequently the case that the null hypothesis is "true," and more and more likely that the null is false. Thus Type I error decreases in importance, and Type II error increases in importance. This means that researchers should be paying increasing attention to Type II error and to statistical power as time goes by. However, a recent review in *Psychological Bulletin* (Sedlmeier & Gigerenzer, 1989) concluded that the average statistical power of studies in one APA journal had declined from 46% to 37% over a 22-year period, despite the earlier appeal by Cohen (1962) for attention to statistical power. Only two of the 64 experiments reviewed even mentioned statistical power, and none computed estimates of power. The review concluded that the decline in power was due to *increased use of alpha-adjusted procedures* (such as the Newman-Keuls, Duncan, and Scheffé procedures). That is, instead of attempting to reduce the Type II error rate, researchers had been imposing increasingly stringent controls on Type I errors—which probably cannot occur in most studies. The result is a further increase in the Type II error rate—an average increase of 17%.

These examples of meta-analysis have examined only the effects of sampling error. There are other statistical and measurement artifacts that cause artifactual variation in effect sizes and correlations across studies—for example, differences between studies in measurement error, range restriction, and dichotomization of measures. Also, in meta-analysis, *mean d* values and mean correlations must be corrected for attenuation due to such artifacts as measurement error and dichotomization of measures. These artifacts are beyond the scope of this presentation but are covered in detail elsewhere (Hunter & Schmidt, 1990a, 1990b). The purpose here is only to demonstrate that traditional data analysis and interpretation methods logically lead to erroneous conclusions and to demonstrate that meta-analysis can solve these problems.

## Applications of Meta-Analysis in Industrial–Organizational (IO) Psychology

Meta-analysis methods have been applied to a variety of research literatures in I/O psychology. The following are some examples: (a) correlates of role conflict and role ambiguity (Fisher & Gittelson, 1983; Jackson & Schuler, 1985); (b) relation of job satisfaction to absenteeism (Hackett & Guion, 1985; Terborg & Lee, 1982); (c) relation between job performance and turnover (McEvoy & Cascio, 1987); (d) relation between job satisfaction and job performance (Iaffaldono & Muchinsky, 1985; Petty, McGee, & Cavender, 1984); (e) effects of nonselection organizational interventions on employee output and productivity (Guzzo, Jette, & Katzell, 1985); (f) effects of realistic job previews on employee turnover, performance, and satisfaction (McEvoy & Cascio, 1985; Premack & Wanous, 1985); (g) evaluation of Fiedler's theory of leadership (Peters, Harthe, & Pohlman, 1985); and (h) accuracy of self-ratings of ability and skill (Mabe & West, 1982). These applications have been to both correlational and experimental literatures. Sufficient meta-analyses have been published in I/O psychology that a review of meta-analytic studies in this area has now been published. This lengthy review (Hunter & Hirsh, 1987) reflects the fact that this literature is now quite large. It is noteworthy that the review devotes considerable space to the development and presentation of theoretical propositions; this is possible because the clarification of research literatures produced by meta-analysis provides a basis for theory development that previously did not exist.

Although there have now been about 50 such published applications of meta-analysis in I/O psychology, the most frequent application to date has been the examination of the validity of employment tests and other methods used in personnel selection. Meta-analysis has been used to test the hypothesis of *situation-specific validity.* In personnel selection it had long been believed that validity was specific to situations; that is, it was believed that the validity of the same test for what appeared to be the same job varied from employer to employer, region to region, across time periods, and so forth

(Schmidt et al., 1976). In fact, it was believed that the same test could have high validity (i.e., a high correlation with job performance) in one location or organization and be completely invalid (i.e., have zero validity) in another. This belief was based on the observation that obtained validity coefficients and statistical significance levels for similar or identical tests and jobs varied substantially across different studies conducted in different settings; that is, it was based on findings similar to those in Table 1. This variability was explained by postulating that jobs that appeared to be the same actually differed in important ways in the traits and abilities required to perform them. This belief led to a requirement for local (or situational) validity studies. It was held that validity had to be estimated separately for each situation by a study conducted in the setting; that is, validity findings could not be generalized across settings, situations, employers, and the like (Schmidt & Hunter, 1981). In the late 1970s and throughout the 1980s, meta-analyses of validity coefficients, called *validity generalization studies,* were conducted to test whether validity might not in fact be generalizable (Callender and Osburn, 1981; Hirsh, Northrop, & Schmidt, 1986; Pearlman, Schmidt, & Hunter, 1980; Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman, & Shane, 1979). If all or most of the study-to-study variability in observed validities was due to artifacts, then the traditional belief in situational specificity of validity would be seen to be erroneous, and the conclusion would be that validity findings generalized.

To date, meta-analysis has been applied to over 500 research literatures in employment selection, each one representing a predictor–job performance combination. Several slightly different computational procedures for estimating the effects of artifacts have been evaluated and have been found to yield very similar results and conclusions (Callender & Osburn, 1980; Raju & Burke, 1983; Schmidt et al., 1980). In addition to ability and aptitude tests, predictors studied have included nontest procedures, such as evaluations of education and experience (McDaniel, Schmidt, & Hunter, 1988), interviews (McDaniel, Whetzel, Schmidt, & Mauer, 1991), and biodata scales (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990). In many cases, artifacts accounted for all variance in validities across studies; the average amount of variance accounted for by artifacts has been 80%–90% (Schmidt, Hunter, Pearlman, & Hirsh, 1985; Schmidt et al., in press). As an example, consider the relation between quantitative ability and overall job performance in clerical jobs (Pearlman et al., 1980). This substudy was based on 453 correlations computed on a total of 39,584 people. Seventy-seven percent of the variance of the observed validities was traceable to artifacts, leaving a negligible variance of .019. The mean validity was .47. Thus, integration of this large amount of data leads to the general (and generalizable) principle that the correlation between quantitative ability and clerical performance is approximately .47, with very little (if any) true variation around this value. Like other similar findings, this finding shows that the old belief that validities are situationally specific is erroneous (Schmidt & Hunter, 1981).

Today, many organizations—including the federal government, the U.S. Employment Service, and some large corporations—use validity generalization findings as the basis of their selection–testing programs. Validity generalization has been included in standard texts (e.g., Anastasi, 1982, 1988), in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985), and in the *Principles for the Validation and Use of Personnel Selection Procedures* (APA, 1987). Proposals have been made to include validity generalization in the federal government's Uniform Guidelines on Employee Selection Procedures when this document is revised in the near future to reflect changes in the recently enacted Civil Rights Act of 1991. A recent report by the National Academy of Sciences (Hartigan & Wigdor, 1989) devotes a full chapter (chap. 6) to validity generalization and endorses its methods and assumptions.

## Role of Meta-Analysis in Theory Development

The major task in any science 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 have an effect on commitment? As another example, 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. But to construct theories, one must first know some of the basic facts, such as the empirical relations among variables. These relations are the building blocks of theory. For example, if there is a high and consistent population correlation between job satisfaction and organization commitment, this will send theory development in particular directions. If the correlation between these variables is very low and consistent, theory development will branch in different directions. If the relation is known to be highly variable across organizations and settings, the theory developer will be encouraged to advance interactive or moderator-based theories. Meta-analysis provides these empirical building blocks for theory. Meta-analytic findings reveal what it is that needs to be explained by the theory.

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 (e.g., see Hunter & Gerbing, 1982; Kenny, 1979; Loehlin, 1987) 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 to test causal theories even when all the delineated relationships are observational rather than experimental. Experimentally determined relationships can also be entered into path analyses along with observationally based relations. It is necessary only to transform *d* values to correlations (Hunter & Schmidt, 1990b, chap. 7). Thus path analyses can be "mixed." Path analysis can be a useful 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). Every such reduction in the number of possible theories is an advance in understanding.

A recent study (Schmidt, Hunter, & Outerbridge, 1986) is an example of this. We applied meta-analysis to studies reporting correlations among the following variables: General mental ability, job knowledge, job performance capability, supervisory ratings of job performance, and length of experience on the job. General mental ability was measured using standardized group intelligence tests; job knowledge was assessed by content-valid written measures of facts, principles, and methods needed to perform the particular job; job performance capability was measured using work samples that simulated or reproduced important tasks from the job, as revealed by job analysis; and job experience was a coding of months on the job in question. These correlations were corrected for measurement error, and the result was the matrix of meta-analytically estimated correlations shown in Table 2. We used these correlations to test a theory of the joint impact of general mental ability and job experience on job knowledge and job performance capability on the basis of methods described in Hunter and Gerbing (1982). Path analysis results are shown in Figure 4. These results indicate that the major impact of general mental ability on job performance capability (measured by the work sample) is *indirect*: Higher ability leads to increased acquisition of job knowledge, which in turn has a strong effect on job performance capability [(.46)(.66)=.30]. This indirect effect of ability is almost 4 times greater than the direct effect of ability on performance capability. The pattern for job experience is similar: The indirect effect through the acquisition of job knowledge is much greater than the direct effect on performance capability. The analysis also reveals that supervisory ratings of job performance are more heavily determined by employee job knowledge than by employee performance capabilities as measured by the work sample measures. I do not want to dwell at length on the findings of this study. The point is that this study is an example of a two-step process that uses meta-analysis in theory development:

Step 1. Use meta-analysis to obtain precise estimates of the relations among variables. Meta-analysis averages out sampling error deviations from correct values, and it corrects mean values for distortions due to measurement error and other artifacts. Meta-analysis results are not affected by the problems that logically distort the results and interpretations of significance tests.

Step 2. Use the meta-analytically derived estimates of relations in path analysis to test theories.

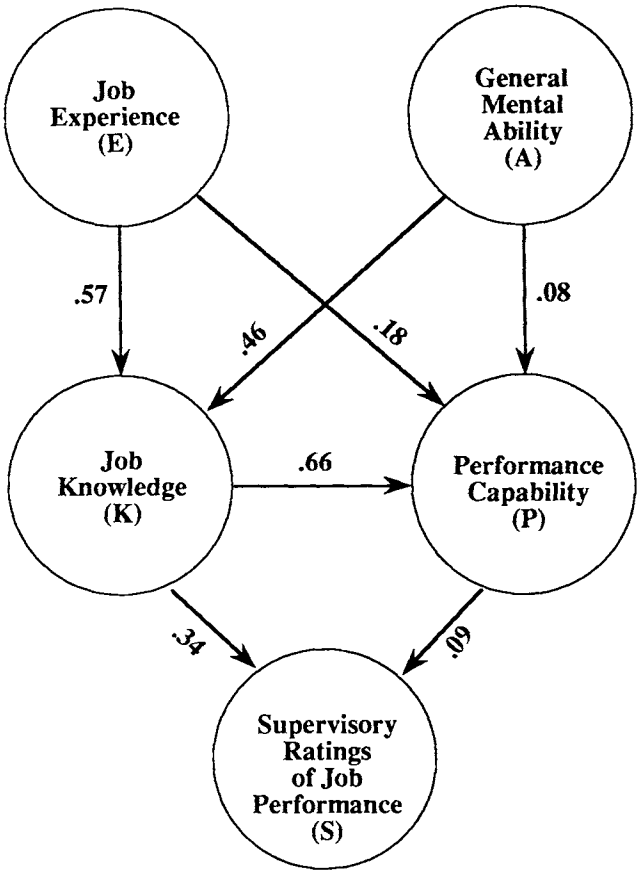
## The Broader Impact of Meta-Analysis

Some have proposed that meta-analysis is nothing more than a new, more quantitative method of conducting literature reviews (Guzzo, Jackson, & Katzell, 1986). If this were true, then its impact could be fully evaluated by

Table 2
*Original Correlation Matrix, Correlation Matrix Reproduced From Path Model, and the Difference Matrix*

| Variable | A | K | P | S | E |
|---|---|---|---|---|---|
| Original matrix | | | | | |
| Ability (A) | — | 0.46 | 0.38 | 0.16 | 0.00 |
| Job knowledge (K) | 0.46 | — | 0.80 | 0.42 | 0.57 |
| Performance capability (P) | 0.38 | 0.80 | — | 0.37 | 0.56 |
| Supervisory ratings (S) | 0.16 | 0.42 | 0.37 | — | 0.24 |
| Job experience (E) | 0.00 | 0.57 | 0.56 | 0.24 | — |
| Reproduced matrix | | | | | |
| Ability (A) | — | 0.46 | 0.38 | 0.19 | 0.00 |
| Job knowledge (K) | 0.46 | — | 0.80 | 0.41 | 0.57 |
| Performance capability (P) | 0.38 | 0.80 | — | 0.36 | 0.56 |
| Supervisory ratings (S) | 0.19 | 0.41 | 0.36 | — | 0.24 |
| Job experience (E) | 0.00 | 0.57 | 0.56 | 0.24 | — |
| Difference matrix (original minus reproduced) | | | | | |
| Ability (A) | — | 0.00 | 0.00 | −0.03 | 0.00 |
| Job knowledge (K) | 0.00 | — | 0.00 | 0.01 | 0.00 |
| Performance capability (P) | 0.00 | 0.00 | — | 0.01 | 0.00 |
| Supervisory ratings (S) | −0.03 | 0.01 | 0.01 | — | 0.00 |
| Job experience (E) | 0.00 | 0.00 | 0.00 | 0.00 | — |

**Figure 4**
*Path Model and Path Coefficients*




merely examining the differences in conclusions of meta-analytic and traditional literature reviews. These differences are major and important, and they show that the conclusions of narrative reviews, based on traditional interpretations of statistical significance tests, are frequently very erroneous. However, meta-analysis is much more than a new method for conducting reviews. The realities revealed about data and research findings by the principles of meta-analysis require major changes in our views of the individual empirical study, the nature of cumulative research knowledge, and the reward structure in the research enterprise.

Meta-analysis has explicated the critical role of sampling error, measurement error, and other artifacts in determining the observed findings and statistical power of individual studies. In doing so, it has revealed how little information there is in any single study. It has shown that, contrary to widespread belief, no single primary study can resolve an issue or answer a question. Consideration of meta-analysis principles suggests that there is a strong cult of overconfident empiricism in the behavioral and social sciences, that is, an excessive faith in data as the direct source of scientific truths and an inadequate appreciation of how misleading most social science data are when accepted at face value and interpreted naively. The commonly held belief that research progress will be made if only we "let the data speak" is sadly erroneous. Because of the effects of artifacts such as sampling error and measurement error, it would be more accurate to say that data come to us encrypted, and to understand their meaning we must first break the code. Doing this requires meta-analysis. Therefore any individual study must be considered only a single data point to be contributed to a future meta-analysis. Thus the scientific status and value of the individual study is necessarily reduced.

The result has been a shift of the focus of scientific discovery from the individual primary study to the meta-analysis, creating a major change in the relative status of reviews. Journals that formerly published only primary studies and refused to publish reviews are now publishing meta-analytic reviews in large numbers. 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 did primary research. Perhaps this was appropriate because it can be seen in retrospect that such reviews often contributed little to cumulative knowledge (Glass et al., 1981; Hedges & Olkin, 1985). Not only is this no longer the case, but there has been a far 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. It is possible for a behavioral or social scientist today with the needed training and skills to make major original discoveries and contributions without conducting primary research studies—simply by mining the information in accumulated research literatures. This process is well under way today. The I/O psychology and organizational behavior research literatures— the ones with which I am most familiar—are rapidly being mined. This is apparent not only in the number of meta-analyses being published but also—and perhaps more importantly—in the shifting pattern of citations in the literature and in textbooks from primary studies to meta-analyses. The same is true in education, social psychology, medicine, finance, marketing, and other areas (Hunter & Schmidt, 1990b, chap. 1).

The meta-analytic process of cleaning up and making sense of research literature not only reveals the cumulative knowledge that is there but also prevents the diversion of valuable research resources into truly unneeded research studies. Meta-analysis applications have revealed that there are questions for which additional research would waste scientifically and socially valuable resources. For example, 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 correlation is .47 ($S_\rho^2$ = .05; Pearlman et al., 1980). For other abilities, there were often 200–300 cumulative studies. Clearly, further research on these relationships is not the best use of available resources.

Only meta-analytic integration of findings across

studies can control chance and other statistical and measurement artifacts and provide a trustworthy foundation for conclusions. And yet meta-analysis is not possible unless the needed primary studies are conducted. Is it possible that meta-analysis will kill the incentive and motivation to conduct primary research studies? 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. But the 50th study contains only about 2% of the available information, the 100th, about 1%. Will we have difficulty motivating researchers to conduct the 50th or 100th study? The answer will depend on the future reward system in the behavioral and social sciences. What will that reward structure be? What should it be?

One possibility—not necessarily desirable—is that we will have a two-tiered research enterprise. One group of researchers will specialize in conducting individual studies. Another group will apply complex and sophisticated meta-analysis methods to those cumulative studies *and will make the scientific discoveries.* Such a structure raises troubling questions. How would these two groups be rewarded? What would be their relative status in the overall research enterprise? Would this be comparable to the division of labor in physics between experimental and theoretical physicists? Experimental physicists conduct the studies, and theoretical physicists interpret their meaning. This analogy may be very appropriate: Hedges (1987) has found that theoretical physicists (and chemists) use methods that are "essentially identical" to meta-analysis. In fact, a structure similar to this already exists in some areas of I/O psychology. A good question is this: Is it the wave of the future?

One might ask, Why can't the primary researchers also conduct the meta-analyses? This can happen, and in some research areas it has happened. But there are worrysome trends that militate against this outcome in the longer term. Mastering a particular area of research often requires all of an individual's time and effort. Research at the primary level is often complex and time consuming, leaving little time or energy to master new quantitative methods such as meta-analysis. A second consideration is that, as they are refined and improved, meta-analysis methods are becoming increasingly elaborated and abstract. In my experience, many primary researchers already felt in the early 1980s that meta-analysis methods were complex and somewhat forbidding. In retrospect, those methods were relatively simple. Our 1982 book on meta-analysis (Hunter et al., 1982) was 172 pages long. Our 1990 book (Hunter & Schmidt, 1990b) is 576 pages long. The meta-analysis book by Hedges and Olkin (1985) is perhaps even more complex statistically and is 325 pages long. Improvements in accuracy of meta-analysis methods are mandated by the scientific need for precision. Yet increases in precision come only at the price of increased complexity and abstractness, and this makes the methods more difficult for generalists to master and apply.

Perhaps the ideal solution would be for the meta-analyst to work with one or two primary researchers in conducting the meta-analysis. Knowledge of the primary research area is desirable, if not indispensable, in applying meta-analysis optimally. But this solution avoids the two-tiered scientific enterprise for only one small group of primary researchers. For the rest, the reality would be the two-tiered structure.

## Conclusion

Traditional procedures for data analysis and interpretation in individual studies and in research literatures have hampered the development of cumulative knowledge in psychology. These procedures, based on the statistical significance test and the null hypothesis, logically lead to erroneous conclusions because they overestimate the amount of information contained in individual studies and ignore Type II errors and statistical power. Meta-analysis can solve these problems. But meta-analysis is more than just a new way of conducting research reviews: It is a new way of viewing the meaning of data. As such, it leads to a different view of individual studies and an altered concept of scientific discovery, and it may lead to changed roles in the research enterprise.

### REFERENCES


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association, Division of Industrial and Organizational Psychology (Division 14). (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

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

Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. *Journal of Applied Psychology, 65,* 543–558.

Callender, J. C. & Osburn, H. G. (1981). Testing the constancy of validity with computer generated sampling distributions of the multiplicative model variance estimates: Results for petroleum industry validation research. *Journal of Applied Psychology, 66,* 274–281.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1,* 98–101.

Fisher, C. D. & Gittelson, R. (1983). A meta-analysis of the correlates of role conflict and ambiguity. *Journal of Applied Psychology, 68,* 320–333.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Edinburgh, Scotland: Oliver & Boyd.

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

Glymour, C. (1980). The good theories do. In E. B. Williams (Ed.), *Construct validity in psychological measurement* (pp. 13–21). Princeton, NJ: Educational Testing Service.

Guzzo, R. A., Jackson, S. E., & Katzell, R. A. (1986). Meta-analysis analysis. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 9). Greenwich, CT: JAI Press.

Guzzo, R. A., Jette, R. D., & Katzell, R. A. (1985). The effects of psychologically based intervention programs on worker productivity: A meta-analysis. *Personnel Psychology, 38,* 275–292.

Hackett, R. D., & Guion, R. M. (1985). A re-evaluation of the absenteeism–job satisfaction relationship. *Organizational Behavior and Human Decision Processes, 35,* 340–381.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist, 42,* 443–455.

Hedges, L. V., & Olkin, I. (1980). Vote counting methods in research synthesis. *Psychological Bulletin, 88,* 359–369.

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

Hirsh, H. R., Northrop, L, & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology, 39,* 399–420.

Hunter, J. E. (1988). *A path analytic approach to analysis of covariance.* Unpublished manuscript, Michigan State University, Department of Psychology.

Hunter, J. E., & Gerbing, D. W. (1982). Unidimensional measurement, second order factor analysis and causal models. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 4). Greenwich, CT: JAI Press.

Hunter, J. E., & Hirsh, H. R. (1987). Applications of meta-analysis. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1987* (pp. 321–357). New York: Wiley.

Hunter, J. E., & Schmidt, F. L. (1990a). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75,* 334–349.

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Iaffaldono, M. T., & Muchinsky, P. M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin, 97,* 251–273.

Jackson, S. E., & Schuler, R. S. (1985). A meta-analysis and conceptual critique of research on role ambiguity and role conflict in work settings. *Organizational Behavioral and Human Decision Processes, 36,* 16–78.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

Koch, S. (1964). Psychology and emerging conceptions of knowledge as unitary. In T. W. Wann (ed.), *Behaviorism and Phenomenology.* Chicago: University of Chicago Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review, 41,* 429–471.

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Erlbaum.

Mabe, P. A., III, & West, S. G. (1982). Validity of self evaluations of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 280–296.

Mackenzie, B. D. (1977). *Behaviorism and the limits of scientific method.* Atlantic Highlands, NJ: Humanities Press.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods of rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–314.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Mauer, S. (1991). *The validity of employment interviews: A review and meta-analysis.* Manuscript submitted for publication.

McEvoy, G. M., & Cascio, W. F. (1985). Strategies for reducing employee turnover: A meta-analysis. *Journal of Applied Psychology, 70,* 342–353.

McEvoy, G. M., & Cascio, W. F. (1987). Do poor performers leave? A meta-analysis of the relation between performance and turnover. *Academy of Management Journal, 30,* 744–762.

McKeachie, W. J. (1976). Psychology in America's bicentennial year. *American Psychologist, 31,* 819–833.

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

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.

Peters, L. H., Harthe, D., & Pohlman, J. (1985). Fiedler's contingency theory of leadership: An application of the meta-analysis procedures of Schmidt and Hunter. *Psychological Bulletin, 97,* 274–285.

Petty, M. M., McGee, G. W., & Cavender, J. W. (1984). A meta-analysis of the relationship between individual job satisfaction and individual performance. *Academy of Management Review, 9,* 712–721.

Premack, S., & Wanous, J. P. (1985). Meta-analysis of realistic job preview experiments. *Journal of Applied Psychology, 70,* 706–719.

Raju, N. S., & Burke, M. J. (1983). Two new procedures for studying validity generalization. *Journal of Applied Psychology, 68,* 382–395.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184.

Schlagel, R. H. (1979, September). *Revaluation in the philosophy of science: Implications for method and theory in psychology.* Invited address at the 87th Annual Convention of the American Psychological Association, New York.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 643–661.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432–439.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology, 66,* 166–185.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38,* 697–798.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology, 32,* 257–381.

Schmidt, F. L., Hunter, J. E., & Urry, V. E. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473–485.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. A. (in press). Refinements in validity generalization procedures: Implications for the situational specificity hypothesis. *Journal of Applied Psychology.*

Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 38,* 509–524.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Suppe, F. (Ed.). (1977). *The structure of scientific theories.* Urbana: University of Illinois Press.

Terborg, J. R., & Lee, T. W. (1982). Extension of the Schmidt–Hunter validity generalization procedure to the prediction of absenteeism behavior from knowledge of job satisfaction and organizational commitment. *Journal of Applied Psychology, 67,* 280–296.

Toulmin, S. E. (1979, September). The cult of empiricism. In P. F. Secord (Chair), *Psychology and Philosophy.* Symposium conducted at the 87th Annual Convention of the American Psychological Association, New York.