I

META-ANALYSIS ANALYSIS

Richard A. Guzzo, Susan E. Jackson and
Raymond A. Katzell

ABSTRACT

This chapter examines meta-analysis, a quantitative form of literature review enjoying a surge of popularity. The chapter focuses on the decisions a meta-analyst makes in conducting such a review and how these decisions affect the conclusions reached through meta-analysis. The goal of the chapter is to create intelligent users and consumers of meta-analysis by making clear the sources of bias and subjectivity in meta-analysis, its problems, ambiguities and ways of dealing with them, and the usefulness of the technique vis-a-vis less-quantitative forms of literature review.

To begin, we consider the purposes of literature reviews, which include assessing the current state of knowledge, identifying directions for future research, advancing theory, and guiding policy decisions. This chapter evaluates meta-analysis as a technique for achieving these purposes, and we find that meta-analysis is less likely to be useful in advancing theory than in fulfilling other purposes. The value of meta-analysis for these other purposes, it is argued, depends in part on the state of the literature to be reviewed.

Further, this chapter evaluates meta-analysis's objectivity. In doing so, we describe numerous judgment calls that must be made by the meta-analyst. These judgment calls introduce considerable subjectivity into the meta-analytic review process. We conclude that, on balance, meta-analysis is not necessarily more objective than traditional narrative literature reviews and that the merit of each form of review depends on the circumstances in which it is used.

Research in Organizational Behavior, volume 9, pages 407-442.

ISBN: 0-89232-636-0

Agreed-upon rules for conducting "scientific" studies of human behavior, and for interpreting the results of such studies, have been evolving to higher and higher levels of sophistication. The questions researchers ask before accepting the data-based conclusions of a single study reflect awareness of the many flaws that can diminish the validity of a study. This awareness, usually developed through formal education, cautions against the acceptance of research that does not sufficiently meet scientific standards. Thus, our profession teaches us to eschew "dustbowl empiricism" in favor of theory-driven hypothesis testing, to insist that constructs be clearly and precisely defined, to question results based upon measures with unknown reliability and validity, to scrutinize a research design in search of confounding factors that weaken internal validity, and to demand acknowledgment of the boundaries beyond which generalizations of a study's results are inappropriate.

In contrast to the clarity of standards that exist for single studies is the ambiguity of criteria for assessing conclusions drawn from multiple studies dealing with a particular issue. Consider the standards against which a traditional narrative literature review is evaluated, standards that are imprecise and judgmental: Is the review comprehensive? Well organized? Unbiased? Clearly written? Insightful? Probably because there are no clear standards in this area, the art of doing a review receives little attention in formal treatments of social science methodology (see Jackson, 1980).

The traditional narrative review is vulnerable on grounds of subjectivity. It is not unheard of for two reviewers to arrive at rather different conclusions from the same general body of literature; For example, compare reviews of Sashkin (1984) Locke and Schweiger (1978) regarding the effects of participation in organizations. To researchers trained to value and always maximize objectivity in the pursuit of knowledge, the subjective quality of the traditional narrative review is its major shortcoming. Truth, like beauty, may in these instances be in the eye of the beholder.

This chapter examines a new method of literature review, *meta-analysis* (Glass, 1976; Glass, McGaw, & Smith, 1981; Rosenthal, 1984; Hunter, Schmidt, & Jackson, 1982). Meta-analysis, as we shall see, has been put forth as a more objective method for conducting literature reviews than alternative methods. We focus on the decisions a meta-analyst makes in conducting such a review, giving particular attention to how these decisions affect the validity of conclusions reached through meta-analysis. By examining the choices made in conducting a meta-analysis, we hope to inform others of the ways in which biases and subjectivity can affect the conclusions drawn on the basis of meta-analytic data. Where possible, we offer prescriptions and suggestions for dealing with bias and subjectivity in meta-analysis. Further, we discuss generally the costs and benefits of quantification in literature review.

Before beginning our analysis, we describe briefly the purposes and the range of methods of literature reviews and the promises and procedures of meta-analysis.

PURPOSES OF LITERATURE REVIEWS

Literature reviews are acknowledged as valuable contributions to the advancement of science. When done well, they help us sum up where we have been and where we need to go next. They identify gaps in our knowledge and suggest new directions for future research. That such endeavors are considered useful is reflected in the number published each year; that they are considered of a different kind than original research studies is reflected in the segregation of such articles in journals such as the *Academy of Management Review* and *Psychological Bulletin.*

Reviews of research literature are carried out for several reasons. One reason, of course, is to assess the state of current knowledge. Relatedly, literature reviews assess what is *not* known-the gaps in knowledge. It is quite common in published literature reviews to find a discussion of what directions ought to be taken by future research on a topic. Such discussions are of value to the extent that they define productive lines of research and promote the integration of future findings with current knowledge. A third reason for conducting literature reviews is to advance theory. Good literature reviews can make strong statements about the validity of theories and can stimulate new theoretical development. A fourth reason is to answer the "so what" question so often put to behavioral science research. That is, literature reviews can provide statements about the policy implications of research findings, the practices that can be justified on the basis of research.

CONTINUUM OF QUANTIFICATION

Literature reviews conducted for the reasons previously cited can be carried out through a variety of methods. In fact, it is possible to depict forms of literature review as varying along *a continuum of quantification.* At one end of the continuum is the traditional narrative review. In this form, verbal descriptions of research findings are presented along with conclusions drawn from those findings. This type of review is perhaps the most familiar. A form of literature review that introduces a bit more quantification is the "box score" review. Here, a tally is made of the frequency with which existing research findings support a particular proposition or reveal an interpretable pattern.

Forms of literature review with an even greater degree of quantification

include (a) cumulating reported significance levels and (b) calculating the number of "file drawer" studies necessary to disconfirm the findings of published literature (e.g., Rosenthal, 1984). In the former method, levels of statistical significance reported in original research are combined to estimate the overall probability of the pattern of findings presented by that research. This procedure provides a more precise way of integrating statistically significant with nonsignificant findings than does simply tallying the number of significant and nonsignificant findings.

The calculation of the number of file drawer studies needed to disconfirm a conclusion based on reported literature makes use of an estimate of the overall probability of a pattern of findings to answer the question, How many studies with nonsignificant findings-studies stored in file drawers-must exist in order to disconfirm the findings that appear in published literature? The larger the number of such studies, the more confidence consumers of research can have in the conclusions derived from them.

The most extreme form of quantification in literature review, the newest method on the scene, is meta-analysis. Meta-analysis refers to a family of related, quantitative procedures for reviewing a body of literature. Meta-analytic procedures all share certain properties, including calculation of *effect size* estimates and statistical assessment of the strength of relationships between effect sizes and variables such as population characteristics or aspects of the research designs of reported studies.'

THE STATUS OF META-ANALYSIS

Accompanying the pride taken by many behavioral scientists in valuing objectivity over subjectivity in original research or in literature reviews is an appetite for quantification, abetted by ever-more sophisticated statistical technology. The availability of such methods seem to assure us that we are indeed scientific. Thus, new statistical procedures are quickly accepted and, thanks to computers, sometimes used before they are fully understood.

Given a continuing desire to be more scientific, and a tendency to equate sophisticated statistical technology with science, it is not surprising that researchers quickly adopt the new statistical method that promises to transform literature reviews into precise, quantitative, objective-scientific-endeavors.

Meta-analysis has enjoyed skyrocketing use. The bar chart in Figure 1 shows, by year, the number of journal articles and dissertations reported in *Psychological Abstracts* (1967-1985) that used or commented on, *and* were key-worded as, *meta-analysis* (no articles were key-worded *meta-analysis* prior to 1977). The rapid increase in the use of meta-analysis has led some to speculate, and others to perhaps fear, that meta-analysis may

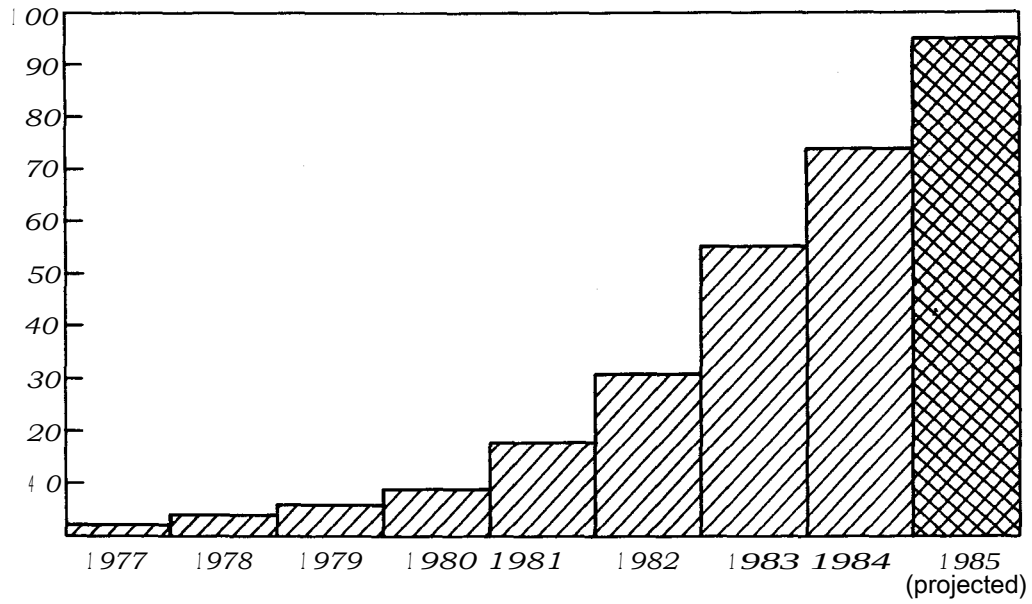

Figure 1. Frequency of appearances of meta-analyses in *Psychological Abstracts* by year for 1977 through 1985. The projected value for 1985 is based upon data from the first half of the year.

become the obligatory method for comprehensive literature reviews (Cook & Leviton, 1980; Eysenck, 1978).

Despite this skyrocketing use, however, not all behavioral science researchers are familiar with the specifics of meta-analysis. The scope of researchers' understanding of meta-analysis is suggested by the results of a survey conducted by Jackson (1984). To learn more about how meta-analysis was viewed by authors of review papers, Jackson sent a short questionnaire to the primary authors of 150 reviews published in the *Psychological Bulletin* between January 1980 and May 1984. One question asked, "How familiar are you with meta-analysis *at this time?" Of* the 132 respondents, 15% indicated meta-analysis procedures were "very familiar; I have used it." Another 15% indicated meta-analysis was "fairly familiar; I could use it without much further study." The remainder of the respondents indicated that meta-analysis was only "somewhat familiar; I know the basic principles" (37%), or was "not very familiar; I could not state the basic principles" (43%).[2] One interpretation *of* these data is that most social scientists do not have a deep enough understanding *of* meta-analysis to be critical consumers *of* its results.

It appears that literature reviews *of* organizational research increasingly will rely on meta-analysis. Meta-analysis results thus will be the basis from which conclusions are drawn about what is known and what to study in the future. It is therefore important that users *of* meta-analysis-both "doers" and "consumers"-know how to critically evaluate a meta-an-

alytic review, in order to decide whether to accept the author's statistical and conceptual conclusions. For just as is true of original empirical studies, the research design used to generate empirical meta-analysis results determines the inferences that can validly be drawn.

OVERVIEW OF META-ANALYSIS

Promises

Quantitative methods for conducting literature reviews are intended to be aids that minimize the extent to which conclusions from a review reflect subjective interpretations that are not justified by data. As Cook and Leviton (1980) note, literature reviews seek "to establish the `facts' . . . the stubborn, dependable relationships that regularly occur despite any biases that may be present in particular studies" (p. 449). Meta-analysis aims to identify these facts more accurately than can be done in a traditional narrative review. By using it, we are told, we can discover the "patterns of relatively invariant underlying relations and causalities, the establishment of which will constitute general principles and cumulative knowledge" (Hunter, et al., 1982, p. 26).

The ability to identify general principles is important because it marks where we stand as a science and, hopefully, it encourages the making of policy decisions-at both an organizational and societal level-that are consistent with accepted facts (Hunter et al., 1982; Schmidt, 1984). As Masters (1984) states, policy makers voice a need for "one-armed" social scientists who give straight, unqualified answers when asked for advice. Instead, they too often encounter a response such as, "Well, on the one hand thus and so is the case, but on the other hand this and that may hold."

Unlike scientists, who often prefer to dwell on the contingencies that make global generalizations impossible, policy makers must seek to find the most justifiable generalizations. An example of a conclusion based on meta-analysis that could dramatically reshape legal guidelines for both employment testing and personnel selection practices is: "Research has shown that most of the observed variability from one validity study to another for a particular test is due to statistical artifacts. The main one is sampling error. That is, different companies find different validity coefficients for the same cognitive ability test because their sample sizes are small" (Cordes, 1985, p. 20). Acceptance of this conclusion by federal regulators could lead to a resurgence in the use of paper-and-pencil employment selection tests, tests that many people believe may be used in ways that discriminate against minority groups in some situations.

Meta-analysts promise to deliver the "facts" with objectivity. And, as

Jackson (1984) found, social scientists consider objectivity to be one of the primary advantages of meta-analysis. After reading a narrative review, one seldom feels the "facts" are indisputable. Proponents of meta-analysis point to the many flaws that can lead to biased narrative reviews and argue that meta-analytic reviews avoid these flaws. Some of these flaws are as follows:

1. Selecting only a subset of studies for inclusion in the review. This practice allows the reviewer's biases to operate at the initial stage of defining the relevant population of studies to review.
2. Ignoring information about the magnitude of relationships between variables and focusing instead on whether an hypothesis is or is not supported by tests of statistical significance. This practice fails to acknowledge the limited power of significance tests applied to small samples.
3. Ignoring random sampling error when interpreting the meaning of observed variation in results found across studies. This practice leads reviewers to see complexities in the data that are more illusory than real. Biases in human information processing, such as selective perception, illusory correlation, and the assumed representativeness of small samples of data (see Hogarth, 1980, Table 9.2, for a description of these biases), may be sources of this shortcoming of narrative reviews. Hunter et al. (1982, ch. 1) illustrate the operation of such biases through an example of a narrative review of hypothetical research findings.
4. Ignoring the possible effects of study characteristics on the relationships observed among variables. That is, narrative reviewers are thought to fail to detect systematic relationships between research results and features of research design (such as the nature of comparison or control groups or the length of time between measures of variables).
5. Using poorly specified, subjective procedures for conducting the review. This practice limits the replicability of the conclusions.

To avoid the first flaw, proponents of meta-analysis emphasize the importance of doing a comprehensive review of the literature to locate *all* extant studies relevant to the topic of review. The second flaw is avoided because effect sizes (rather than the mere presence or absence of statistical significance in research reports) are the primary information used to draw conclusions about the nature of relationships between variables. Depending on the particular meta-analysis procedures used, the third flaw can be avoided through application of procedures that purportedly allow the reviewer to determine how much of the variation in results found across studies is "real" rather than artifactual, thus reducing the role of human judgment in interpreting variation in findings. The fourth flaw can be avoided because meta-analysis procedures include methods for quantifying

the value of moderator variables-including characteristics of research design-as explanations for variations in results across studies. One of the most exciting promises of meta-analysis is that it allows the reviewer to determine the effects of moderators that have never been examined in an original empirical study. For example, although organization size may never have been investigated in any one study as a moderator of a relationship between, say, formalization and absenteeism, the cumulative results of several studies that happen to have been conducted in organizations of varying sizes permits the investigation of organization size as a moderator of the relationship between these two variables. Finally, the standardization of meta-analysis procedures helps overcome the last of the previously cited flaws.

Procedures

As mentioned earlier, meta-analysis refers to a family of procedures for quantitatively accumulating effect sizes across studies. The procedures used most often are those developed by Glass and his colleagues (Glass et al., 1981) and by Hunter and Schmidt (1981) and their colleagues (Hunter et al., 1982), or some variation of the latter (see Burke, 1984, for a review of six variations of meta-analysis procedures). For the reader unfamiliar with the basic steps of a meta-analysis, we briefly outline first the sequence of steps prescribed by Glass et al. (1981) and then those prescribed by Hunter et al., 1982).

To conduct a meta-analysis according to Glass et al., one goes through the following steps:

1. Select the independent and dependent variables of interest (note that Glass et al. emphasize studies of causal relationships among variables, although their procedures apply to the study of noncausal relationships).
2. Locate all relevant and usable studies containing information about the effect of interest.
3. Code the characteristics of each study that might relate to the size of effects obtained in different studies. For example, one might code for the presence or absence of a control group, for characteristics of the participants, and/or for characteristics of the organization in which the study was conducted.
4. Calculate effect size estimates for independent-dependent variable pairs of interest. An effect size, called a d-value, is obtained by subtracting the control (or comparison) group's mean on the dependent variable from the experimental group's mean on the de-

pendent variable and dividing by the standard deviation for the control group. Specifically:

$$d = \frac{(\bar{X}_e - \bar{X}_c)}{SD_c}$$

5. Calculate the mean effect size(s) across studies.
6. Regress effect size estimates on coded study characteristics to assess the relationship between study results and study characteristics.

The method developed by Hunter and Schmidt and their colleagues is more complex in that it includes corrections for the effects of several artifacts, principally including sampling error, unreliability of measurement, and range restriction. The full procedure is as follows:

1. Define the effect or relationship of interest.
   Locate all relevant and usable studies containing information about the effect of interest.
3. Code each study for characteristics that may be related to the size of the effect obtained in the study.
4. Calculate effect size estimates for each independent-dependent variable pair of interest. Effect sizes can be expressed in analogous forms as correlation coefficients or as d-values.
5. Calculate the mean effect size across studies, weighting the effect size obtained in each study by the size of the sample employed and correcting for unreliability of measurement.
6. Calculate the variance in effect sizes across studies.
7. Determine the amount of variance in effect sizes that can be attributed to the artifacts of sampling error, unreliability of measurement, and range restriction (and other artifacts, if possible).
8. If a sufficiently large percentage of variance across studies can be attributed to the preceding artifacts, the meta-analysis ends. Conclude that the average, corrected effect size accurately reflects the true relationship between the two variables of interest.
9. If a large percentage of variance across studies is unaccounted for by statistical artifacts, determine whether study characteristics can account for the residual variance.

The reader is reminded that the preceding lists of steps involved in conducting a meta-analysis are not intended as recommendations for how one *should* do meta-analysis, but simply as brief descriptions of two of the more widely-used techniques. Recently, a number of statisticians have argued that modifications in these procedures are needed in order to reach valid conclusions (e.g., Callender & Osburn, 1980; Raju & Burke, 1983:

Hedges, 1982; James, Damaree, & Mulaik, 1986). Although a discussion of the disagreements among statisticians regarding the appropriateness of certain meta-analytic procedures is beyond the scope of this chapter, we feel that consumers of meta-analysis (both doers and readers) should be alert to the possibility that, upon resolution of the currently ongoing statistical debates about meta-analysis, reassessments of the validity of published meta-analyses may be warranted. At present, however, the extent to which the proposed revisions would affect meta-analytic findings is not clear.

Throughout the remainder of this chapter, we address several aspects of meta-analytic review procedures that threaten their presumed objectivity and precision. As the chapter reveals, there are many judgment calls that must be made when executing the steps of Glass et al.'s (1981) or Hunter et al.'s (1982) procedures. These judgment calls are elaborated as the chapter progresses; they include the identification and selection of studies for review, the calculation of effect size estimates, and the use of meta-analysis to detect moderated relationships. The judgment calls introduce subjectivity into the review process because there are no agreed-upon standards for the reviewer to use when making them. Awareness of the points at which a meta-analyst must make judgment calls is necessary, we feel, before a consumer of meta-analysis results can evaluate the validity of the conclusions yielded by the review.

The objectivity of a meta-analysis is also threatened by the constraints meta-analytic procedures place upon the literature that can be included in a review. Unlike a traditional narrative review, a meta-analytic review does not allow the reviewer to easily integrate the results of certain types of studies. In particular, the specific statistics reported for a study will determine whether or not it can be incorporated into a meta-analytic review. Consequently, for some literatures, meta-analytic conclusions may be based on a relatively small portion of the total available research.

THE EXECUTION OF META-ANALYSIS

In this section we bring attention to the judgment calls, problems, and ambiguities of meta-analysis that render it vulnerable to bias and error. *Judgment calls* are those decisions big and small that must be made in the absence of guidance from an objective rule or standard. In doing any type of research, many judgment calls are made, the effects of which "often determine the outcome of research" (McGrath, 1982, p. 13). Meta-analysis, which can be thought of as one type of research design, requires the researcher to make a variety of crucial judgment calls.

Some judgment calls arise because meta-analysis is a relatively new statistical technique for which agreed-upon procedures are yet to be developed. These unresolved arguments concerning statistical issues inject ambiguity into the execution of a meta-analysis. Statistical debates generally assume as a point of departure the logic of quantification in literature review and their focus is on fine-tuning, not reforming, the approach. In contrast, the focus of this chapter is on judgment calls reviewers must make prior to pushing one's numbers through formulae. These judgment calls are important to understand irrespective of the particular formulae one employs in a meta-analysis. Future resolution of the statistical debates will reduce a bit of the uncertainty surrounding technical issues, but it will not completely remove the subjective aspects of conducting a literature review.

Judgment calls and ambiguities inherent to meta-analysis are discussed in the next section roughly in the sequence in which they arise when conducting a meta-analysis. Each issue is addressed with two audiences in mind: researchers conducting meta-analyses, whose primary concerns are about the proper application of the technique, and consumers of meta-analytic review articles, whose primary concerns are about how to interpret the results. When possible, we offer prescriptions and suggestions for dealing with these difficulties.

Identifying Studies for Review

Defining the population of relevant studies. The studies appearing in any literature review reflect three types of choices made by the reviewer. The first set of choices, often made implicitly, are choices about how to define the population of potentially relevant studies. The next set of choices revolve around the operational issues of locating those studies that belong to the population identified as relevant. Having located potentially relevant studies, a third set of choices arises concerning which of the located studies actually are incorporated into the review.

Glass et al. (1981) state that traditional narrative reviews of research are seriously flawed because studies are excluded from review because they fail to meet various arbitrary criteria imposed by a reviewer (p. 22). In contrast, meta-analytic reviews are held to be free from this flaw because meta-analysts emphasize that *all* studies conducted on a topic should be included in the review. Is meta-analysis free of the arbitrary inclusion or exclusion of studies for review?

The research covered by any review, quantitative or otherwise, reflects decisions made about how to define the population of relevant studies for the review. Such decisions are judgment calls. There are no unambiguous rules to apply when establishing the boundaries of the set of studies to

be reviewed. Questions one might consider include: Should unpublished studies be included? Are technical reports in or out? What about masters theses and doctoral dissertations? If they are included, which disciplines are likely to produce relevant dissertations? Does the age of the study matter? Does the language in which the report is written matter? Which journals will be searched for studies to include in the review? These and many other questions can be raised in order to determine the rules one will use to define the population of studies relevant to a review.

All literature reviews deal with a bounded domain of studies. Ideally, any review should be exhaustive in its inclusion of studies within that domain. But because there are no standard rules for how reviewers should answer the preceding questions (and thus fix the boundary and methods of including studies), both meta-analytic and traditional narrative reviewers must exercise judgment in the determination of the population of relevant studies. Thus, both forms of literature review can be fairly characterized as "arbitrary" in Glass et al.'s terms. As long as a clear statement is made in a review about how the reviewer defined the population of relevant studies for the review, consumers of the review can make their own judgment about the appropriateness of the sample of studies on which the review and its conclusions are based.

Searching for relevant studies. Having defined a population of relevant studies, the next step is to identify and locate those studies. A recommendation frequently made by advocates of meta-analysis is to use computers for searching the literature. Computer searches offer great convenience, are often assumed to be more exhaustive that manual searches, and have the potential to be highly reliable. Although computer searches are automatically more convenient than manual searches, they are not automatically more exhaustive or more accurate than manual searches.

As anyone who has ever conducted a computer search of the literature is aware, the results of one's search are completely dependent on the keywords used for the search for studies. Given the importance of the keywords one chooses for the results of an electronic search, it is interesting that descriptions of literature search methods seldom include a list of keywords used. The importance of the keywords one chooses are illustrated by an experience of the second author. In preparing a talk about meta-analysis, an electronic search of *Psychological Abstracts* was conducted using the keyword *meta-analysis.* Surprisingly, this search located none of the many meta-analytic studies published by Schmidt and his colleagues during the past several years. Why? Because until recently. Schmidt and colleagues used the term *validity generalization* rather than *meta-analysis* in their publications. Now, validity generalization is regarded as a special application of meta-analysis.

It is virtually impossible to estimate how many studies containing data

relevant to a review are likely to be passed over by an electronic search because they are not keyworded in a way that matches the reviewer's judgment calls about which keyword(s) to use. We believe the number to be substantial. As another illustration, consider the number of studies included in two meta-analytic reviews of research on role conflict and role ambiguity. One review (Fisher & Gitelson, 1983) included 43 studies. Another review of the same literature (Jackson & Schuler, 1985) included 105 studies. Both reviews relied on computer searches of almost the identical time span. In addition, Jackson and Schuler conducted a manual search of selected journals. This limited manual-search process appears to be the primary explanation for the large difference in the studies located by the two research teams.

The differences in the results yielded by these two meta-analyses are substantial. One difference that can be directly linked to the differences in search procedures used are the conclusions the two sets of authors lead the reader to make about which correlates of role conflict and role ambiguity have been studied. Both teams of researchers reported meta-analysis results for all correlates for which they could locate at least three studies. Fisher and Gitelson's review located 18 minimally studied correlates. Jackson and Schuler's review located 29. Of the additional correlates reported on by Jackson and Schuler, two are of particular interest: leader initiating structure and leader consideration. The Jackson and Schuler search yielded 31 studies of initiating structure and role ambiguity and 25 studies of leader consideration and role ambiguity. In contrast, Fisher and Gitelson apparently located fewer than three for each of these correlates. Undoubtedly, a third team of researchers could locate an even larger set of studies relevant to the topic of role ambiguity and conflict.

Keywording problems are not the only shortcoming of electronic searches. Another major drawback is the limited time span covered by such searches. For example, *PsychINFO,* an electronic data base for the psychological literature, covers only the years 1967 to the present. *Sociological Abstracts* can be electronically searched as far back as 1963. For *Management Contents,* the beginning date is 1974; for the *American Business Index,* the beginning date for electronic searches is 1971. Yet another problem of electronic search involves the failure to locate a relevant study in an unlikely source. Finally, just as manual searches are limited to a finite number of sources, so too are electronic searches-no search is likely to be truly "exhaustive."

Our purpose in reviewing some of the limitations of both electronic and manual search techniques is not to suggest that these techniques should not be used; they should. But when used, their limitations should be explicitly acknowledged. Finding the relevant literature "is of utmost importance for the quantitative review" (Green & Hall, 1984, p. 46). Estimates of effect size, even of the true direction of a relationship, can be

faulty if a biased portion of the literature is retrieved. Unless the consumer of meta-analytic results is informed of the procedures used to locate the reviewed studies, the reader cannot decide how much confidence he or she should have in the review's results, because no search procedure is conducted independent of a reviewer's judgment calls.

Surviving the final cut. As noted earlier, meta-analysts appear to be particularly suspicious of the final round of study selection choices made by narrative reviewers. Presumably, even after a subset of studies is retrieved and evaluated by the narrative reviewer, some studies are excluded from further consideration. So, for example, several studies may be discarded as unworthy of attention because they are methodologically flawed.

Although some reviewers defend the practice of screening studies for scientific rigor, meta-analysts argue that this screening process allows the reviewer to introduce irrelevant biases into the decision process. To prevent such biases from contaminating one's results, meta-analysts favor including all studies, regardless of their methodological weaknesses. The logic behind this practice is that the results of any single study are biased by the methodological imperfections of the study, but the effects of these biases should differ from one study to the next. Thus, by averaging across all studies, one averages out the effects of different types of methodological flaws.

One retort to this logic is "garbage in, garbage out" (see Eysenck, 1978). The effects on one's conclusions of including methodologically flawed studies in a meta-analytic review has been debated at length. Fortunately, the answer to the question can be addressed empirically. Using meta-analysis, one can sort out the effects of methodological rigor on the results. Thus, Guzzo, Jette, and Katzell (1985) found that less-rigorous studies of productivity improvement showed stronger effects than did more rigorous studies.

In light of the emphasis meta-analysts give to conducting comprehensive reviews of the literature, one might assume that meta-analysis reviews are likely to be more comprehensive than other reviews. Yet, meta analytic reviews of the research literature are necessarily highly selective: most obviously, they cannot include nonquantitative research results. Furthermore, quantitative research results may not be included if the particular statistics required by meta-analysis formulas are not reported (e.g., sample size, variance, reliability).

Nonquantitative studies are ignored completely in meta-analytic procedures because there is no means for expressing findings in terms of effect sizes. In some areas of organizational research, the distinction between qualitative and quantitative research is not important for meta-analysis because quantitative reports may constitute nearly the entire

population of studies. Research on the relationship between ability and performance is primarily quantitative, for example. Other areas of organizational research rely heavily on qualitative data, such as research on socialization or career development. However, in most areas of organizational research a mix of quantitative and qualitative findings exists. Conflict resolution, group and intergroup dynamics, organizational communication, strategy formulation, and quality of work life are all examples of topical areas in which a mix of qualitative and quantitative findings exist. For any of these topics meta-analytic reviews must necessarily focus only on quantitative reports. Although such reviews have considerable value, they are based on a selected portion of the literature that underlies the constructs and theories found in a research domain. Conclusions based on meta-analyses of data relevant to those constructs and theories should be accepted with awareness of the selective nature of meta-analysis as a review technique.

Whereas the inability of meta-analysis to incorporate qualitative research findings is obvious, the inability of meta-analysis to make use of all available quantitative data is less obvious, until one conducts his or her first meta-analytic review. Meta-analysis demands very specific statistical information about a study, sometimes more information than is reported. The calculation of effect-size estimates requires information about sample size, a commonly reported datum, and information about variances, which are often not reported. The Hunter et al. (1982) approach requires substantial additional information, such as estimates of the reliabilities of measures and of the amount of range restriction. According to Hunter et al. (1982, p. 154), data necessary for meta-analysis are usually missing from at least some of the studies located for review, particularly older studies.

An example of the magnitude of data loss that may occur due to the requirements of meta-analysis comes from a report by Guzzo et al. (1985). The research studies available for inclusion in that meta-analysis were the 207 studies of productivity improvement programs reviewed in narrative form by Katzell, Bienstock, and Faerstein (1977) and Guzzo and Bondy (1983). Each of those studies contained at least one quantitative index of the impact on productivity of an organizational intervention. Of those 207 studies, 98 could be included in the meta-analysis. Over half of the available studies had to be excluded, principally for reasons of insufficient reporting of data in the original research reports. This is not an isolated example. Stone (1986) reported deleting 166 out of 219 studies located for his meta-analytic review of research on the job-scope literature. Of the 166 deleted, about two thirds of those were deleted because adequate data were not reported (Stone, 1985).

Which are preferable: conclusions based on a narrative review of a larger

set of studies (e.g., 207 in Katzell & Guzzo, 1983) or those based on a quantitative review of fewer studies (e.g., 98 in Guzzo et al., 1985)? In the case of the reviews of productivity experiments, the findings yielded by narrative (Katzell & Guzzo, 1983) and quantitative (Guzzo et al., 1985) reviews were fairly similar, thus the question did not prove vital. In some domains, though, the question *is* vital and must be answered by the reviewer.

In attempting to answer this question, two concerns arise. One is the *amount* of literature lost to a reviewer who chooses to conduct a meta-analysis. When few available studies are lost, meta-analysis may be useful for summarizing the "facts" yielded by research on a topic. But when a large portion of the available studies are lost, the costs associated with meta-analysis may be large. To the extent fewer studies are included in the review, confidence in the conclusions of the review will decrease simply because there will appear to be little data upon which conclusions can be based.

A second, related concern is whether the exclusion of available studies leads to systematic biases in the literature used for the review. For example, suppose a reviewer excludes 45 out of 100 available studies on a topic because they were qualitative reports. Does it not seem likely that the 45 excluded studies might have been carried out on the basis of different theoretical assumptions?

To illustrate, consider the case of literature on the nature of managerial jobs. Mintzberg's (1973) analysis of managerial jobs is representative of research based on qualitative observation of managers at work. Tornow and Pinto's (1976) report, on the other hand, is representative of highly quantified analyses of managerial jobs. Two very different streams of research on the nature of managerial work exist, and these two streams appear to reflect fundamental differences in assumptions regarding the way in which managerial work is best understood. Any literature review that were to include only one stream of this research would provide a biased account of the nature of managerial work.

The point here is that the usefulness of meta-analysis for summarizing and integrating research findings depends on the *representativeness* of the studies included in the meta-analysis. Meta-analysis can deal only with quantitative studies; narrative reviews, on the other hand, can deal with both quantitative and nonquantitative studies. When the two types of studies in a topical area are not similar in their methods of data collection and interpretation, reviews based on meta-analysis will yield biased summaries of what is known.[3]

Thus, the use of meta-analysis leads to a gain in precision at the cost of selectivity. There is no easy solution to the meta-analyst's problem of data loss because the technique is inherently selective: If the necessary

quantitative data are not available, a study cannot be included in a meta-analysis. One long-term solution to the problem of data loss would be for journal editors to impose strict reporting requirements for quantitative studies to ensure that those studies contain the information necessary for them to be included in future meta-analyses (Hunter et al., 1982). Short-term solutions to the problem include locating the authors of past reports and requesting the needed information. For a variety of reasons, this approach is likely to have limited success. Another short-term solution is simply to make a best guess, as needed, about sample size or reliability. This approach may be safe in some cases but faulty in others. For some time to come, meta-analysts will necessarily continue to exclude existing quantitative research from their reviews as a result of current and past reporting practices.

Although not a solution to the problem, a statement of the number of studies excluded from a review relative to the number of studies located should be included in meta-analytic reports. Such statements would provide relevant information about the degree to which the meta-analysis is based on a selective sample of studies. Just as the representativeness of the sample of subjects included in any one research study is an important determinant of the interpretation of its findings so, too, is the representativeness of the sample of studies included in a meta-analysis an important determinant of the interpretation of its findings.

Calculating Effect Size Estimates

A crucial datum of meta-analysis is an estimate of the magnitude of relationship between two variables, referred to as *effect size estimates.* The meta-analysis literature contains several proposals for how best to calculate effect size estimates but no one method is universally adopted. Little is known about the consequences of calculating effect size estimates one way or another and a judgment call must be made by the meta-analyst concerning how to calculate this crucial datum.

Ambiguity concerning how best to calculate effect size estimates exists for several reasons. The formula proposed by Glass et al. (1981) and presented earlier in this chapter, is an adaptation of Cohen's (1969) d. A point of disagreement prominent in the literature concerns the proper denominator for this formula. Glass et al. (1981) argue that the standard deviation of the control group is the proper term, whereas Schmidt et al. (1982) argue that the denominator should be a variance estimate based on pooling the variance estimates of treatment and control groups. Other estimates of the magnitude of a relationship, such as Rosenthal's (1984) r, exist that, theoretically, can be used interchangeably with the former. Further, estimates of the magnitude of relationship can be inferred in the absence of

direct reports of means and standard deviations or correlations. Glass et al. (1981) provide formulas for estimating effect sizes on the basis of reported t-test values, analyses of variance results, time-series data, and on the basis of data expressed as proportions. These formulas give the meta-analyst considerable power to make use of many forms of quantitative data reported in a literature. Although conceptually similar, little is known about the actual empirical differences that might appear in effect size estimates calculated through the various procedures.

Complexity in research designs also poses difficulties for the calculation of effect size estimates. For example, consider a study containing elements of both true experiments and time series designs, such as one in which groups, randomly assigned to treatment and control conditions, measured on a dependent variable on repeated occasions, some occurring before and some after the treatment. How should effect sizes be calculated in this instance? Should several effect size estimates be calculated, one for each time of measurement of the dependent variable, or should only one effect size be calculated, reflecting an average of the multiple measures of the dependent variable? Is the between-group comparison more relevant than the comparison between pre- and post-intervention measures? On what basis should an estimate of the variance in the dependent variable be calculated? Although answering these questions is beyond the scope of this paper, their answers do have implications for the value of the effect size estimate, the primary datum of meta-analysis. The point of raising such questions is to illustrate how judgment calls permeate the process of data analysis in meta-analysis.

Detecting Moderators

Ideally, the final stage of a meta-analytic review involves analyses to detect important moderator variables that might explain variance in observed effect sizes. Depending upon which meta-analytic procedures one is following, the decision of *whether* to examine the data for possible moderator effects is either a judgment call made by the investigator (when using the Glass et al., 1981 techniques) or a decision based on the results of statistical analyses (when using the Hunter et al., 1982 techniques).

According to Glass et al. (1981), any meta-analysis should end with a search for moderator effects. To date, Glassian meta-analyses have typically focused on examining methodological differences across studies to determine whether these methodological differences can account for differences in effect sizes. Thus, Glass et al. encourage researchers to code studies for variables such as whether the study included a true control group, the study's sample size, year of study, and the amount of time that elapsed between the time of the intervention and assessment of outcomes.

This approach to meta-analysis can be characterized as having as its goal an exhaustive consideration of potentially important methodological moderators.

When using the Glassian procedures, judgment calls relate primarily to which moderators to consider and how to code studies on these moderators. These judgment calls will be addressed in a later section. First, however, it is useful to describe briefly Hunter et al.'s (1982) prescriptions concerning when to search for possible moderator effects.

The Decision to search for moderators. In contrast to Glass et al. (1981), Hunter et al. (1982) are more conservative in their treatment of moderators. Whereas a search for moderator effects is almost automatic when using Glassian procedures, moderator effects are the explanation of last resort when using the Hunter et al. procedures. A key feature of the Hunter et al. approach to meta-analysis is the assumption that variance in observed effect sizes should be attributed first to statistical artifacts and other sources of "error." A search for substantive explanations for variance found across studies is therefore conducted only after variance attributable to nonsubstantive artifacts is accounted for. Thus, the Hunter et al. procedures for conducting a meta-analysis include methods for determining how much of the variance in effect sizes can be explained by the "artifacts" of restriction of range, unreliability of measurement, sampling error, and certain other sources of error (e.g., reporting errors and typographical errors).

Hunter et al. have suggested that 25% of the variance found across studies should be assumed to be caused by unquantifiable errors (this estimate is subject to debate; James et al., 1986). This leaves 75% of the variance in effect sizes to be accounted for by statistical artifacts and substantive moderators. In order to determine how much variance is due to substantive moderators, the meta-analyst first determines error attributed to artifacts. Then, any variance still unaccounted for can be assumed to be caused by moderator variables. At this point, the judgment calls a meta-analyst makes in order to proceed with a search for moderators are the same for the Hunter and Glass procedures.

Although, in principle, correcting estimates of the magnitude of relationship for statistical artifacts is desirable, in practice problems may arise (James et al., 1986). Some problems concern the mis-estimation of the amount of variance in relationships that can be justly attributed to sampling error and other statistical artifacts. James et al. also argue that certain assumptions upon which corrections are predicated are false (e.g., the assumption of independence of true scores and error scores in measures). Another problem identified by James et al. is the inability to distinguish between variance in estimates of effect size due to error from variance clue to meaningful situational differences among the studies reviewed.

Selecting potential moderators. Regardless of whether a reviewer selects the Glass et al. (1981) or the Hunter et al. (1982) approach to the detection of moderator variables, once an affirmative decision is made the next step is to decide which particular moderators to examine. The particular moderators that a meta-analyst might examine can be suggested by theoretical hypotheses about potentially important moderators, by methodological concerns, or by intuition and curiosity. From a statistical standpoint, the conservative approach is to adopt an hypothesis-testing perspective. From this perspective, studies are coded only for those variables predicted, with conceptual justification, to be potential moderators. Although this strategy minimizes the likelihood of accepting a moderator effect that appears only by chance, it has the drawback of reducing the possibility of new knowledge being created through the detection of unforeseen moderators.

A liberal approach to selecting variables to be examined as potential moderators is to code studies on all possible variables and test whether effect sizes found across studies differ as a function of any of these variables. This strategy increases the likelihood of making Type I errors, but it also maximizes the chance of discovering something new by following up a hunch. Indeed, the promised power of meta-analysis to test hunches about moderator variables is one source of its appeal to those interested in theory development (Jackson, 1984). However, the power of meta-analysis to test moderators depends on the number of studies available in the literature. It is when the number of studies (hence the number of effect size estimates) is large that meta-analysis is most powerful in this regard.

Ultimately, the choice between a conservative or liberal approach in the detection of moderators is up to the meta-analyst. For now, at least, there are no standard guidelines for making this choice or for evaluating the risks associated with either option. Whatever choice is made, though, should be made clear to the consumer.

Coding studies for moderators of interest. Although the search for moderator effects might be one of the most exciting aspects of conducting a meta-analytic review, it is also the most frustrating, for what sounds easy in principle is extremely difficult in practice. A successful search for important moderators presumes the ability to accurately code individual studies (or subgroups of participants within a study) for the moderator variables of interest. For example, consider Locke's (1986) attempt to determine whether testing a theory in the laboratory versus in the field leads to different conclusions. Meta-analysis can be used to address this general question for a variety of empirical literatures, as follows: First, identify the effects of interest. Second, locate all empirical studies that have reported data for the effect of interest. Third, calculate the effect sizes for these studies. Fourth, code all studies on the moderator of in-

terest, which in this case is "laboratory" or "field." Fifth, compare the effect sizes found in the two types of studies. This is in fact the procedure followed by some reviewers interested in the question of whether different effects have been found in the laboratory versus in the field (e.g., Stone, 1986). Such studies are predicated on the assumption that studies in organizational behavior can be reliably and validly coded as having occurred in either the laboratory or the field. Yet, Campbell (1986) makes the case that the laboratory-versus-field distinction is not without ambiguity. Thus, judgment calls must be made. For example, how should one classify a study in which 200 people are employed for two weeks to work in an employment setting created by a researcher? How one makes this particular judgment call will be a function of how one defines the critical features of laboratory and field research designs. Other examples of ambiguity that could be encountered in coding potential moderator variables include distinguishing between types of workers (e.g., technical vs. professional) and organization type (e.g., batch versus process production).

The meaningfulness of any conclusions drawn about the role of any moderator variable depends on the meta-analyst's ability to distinguish in a way considered acceptable by readers between values or levels of a variable and to demonstrate that the distinction can be reliably operationalized-that is, to demonstrate inter-rater reliability for the coding of the moderator variable. In effect, conducting a search for moderators using meta-analysis requires that the researcher address all of the issues that arise when doing any form of content analysis: dimensions of interest must be clearly defined, coding schemes must be developed, and multiple raters must be trained to reliably code the variables of interest (Bullock & Svyantek, 1985).

Because information about samples, context, and certain features of research design is often reported in incomplete and imprecise ways, coding studies for moderators may represent the bulk of the work in conducting a meta-analysis. Uncertainty arises in the process of translating prose descriptions of study characteristics into quantitative form; the more uncertainty, the more likely that errors exist. Under such conditions-which, indeed, are typical-the coding in meta-analysis should be subject to the same standards of reliability and validity that would be demanded of any study whose findings rely heavily on content analysis.

Evidence of reliability for coding procedures is not common in existing meta-analytic reports, but an interesting study by Stock and colleagues suggests such coding may have low reliability unless adequate precautions are taken to insure high reliability (Stock et al., 1982). Two variables that coders could-not rate reliably were the quality of a study and the number of different subsamples included in a study. With sufficient training and a well-designed coding scheme, though, acceptable reliability was produced by Stock et al.'s coders for many coding categories.

THE TECHNIQUE OF META-ANALYSIS: A SUMMING UP

Meta-analysis is a literature-review technique with great face validity. Prior to its appearance, literature reviews were restricted to narrative accounts or accounts that made use of minimal levels of quantification. Through its quantification, meta-analysis is a significant alternative to these traditional methods.

A number of variants of quantitative literature review exist under the rubric of meta-analysis. Although each variant has unique properties, they share much, including the claim of being objective. All things considered, how well does meta-analysis keep its promise of objectivity?

As we have discussed, meta-analysis demands many judgment calls. It lacks uniform standards and rules regarding what studies to include, how to calculate effect sizes, whether or not to correct estimates for variance due to error, and other critical choices. Further, meta-analysis is a selective technique for literature review because of its inability to deal with nonquantitative research and quantitative research lacking certain information. The selectivity, judgment calls, and ambiguities render meta-analysis no more objective than narrative review. However, if meta-analytic procedures become uniform and standardized, this quantitative review procedure may well become more objective than narrative forms of review *when applied to identical bodies of literature.*

Although meta-analysis does not deliver the hoped-for objective literature review, it does make several contributions not made by traditional forms of review. A significant contribution is its emphasis on the calculation of effect sizes. Effect sizes provide unique information. Not only is information about an average effect size valuable, information about variance of effect sizes also is valuable. Effect size estimates are also of value in planning future research: By knowing the typical strength of relationship between variables (as estimated through meta-analysis), one can make reasonable estimates of sample sizes for future research. Further, meta-analysis capably integrates studies with nonsignificant findings with those with significant findings because of its use of a common metric (effect size estimates) for representing any study's findings.

META-ANALYSIS AND THEORY

As pointed out earlier, literature reviews are conducted for several reasons. Some of their purposes relate to stimulating the development of new theories and testing the adequacy of existing theories. In the sections that follow we discuss the contribution of meta-analysis to theory generation and theory testing.

Theory Generation

To what extent is meta-analysis useful for generating theory? One way to answer this question is to determine the extent to which existing meta-analyses have been used for theory development. A few examples of using meta-analysis for theory development have been published (e.g., Tanford & Penrod, 1984; Winkler & Winett, 1983), but these are exceptions to the typical use of the technique. If meta-analysis *can* be used for theory building, why is it not typically used this way? Cook and Leviton (1980) suggest one reason is that, at least to date, people using meta-analysis have tended toward main effect conclusions due to the topics of interest to them and the arena in which those topics exist. For example, Schmidt and Hunter's (1978) work is in part a response to pressures that exist for personnel psychologists who have traditionally had to live with the "problem" of situational specificity of test validities. Smith and Glass's (1977) work is in part a response to criticisms of the effectiveness of psychotherapy.

The characteristic of meta-analysis that it is well suited to detecting main effects and simple relationships indicates that the technique is useful for building "main effect" theories-that is, explanations without contingencies. However, meta-analysis appears less well suited for generating theories in which contingencies exist.

A problem in the use of meta-analysis, identified earlier, is the technique's difficulty in establishing the existence of moderated relationships. Establishing the existence of moderated relationships is a good way of generating *new* theories that involve contingent explanations. Because meta-analysis encounters difficulty in this regard, it is limited as a technique for theory generation.

Other reasons limiting the use of meta-analysis as a tool for theory generation concern the nature of research on behavior in organizations. These issues have received little attention to date but are likely to be encountered by anyone attempting to review a research domain that consists primarily of large-sample, correlational field studies.

Using meta-analysis for theory development often implies empirically examining the usefulness of moderator variables in accounting for differences in results across studies or across subgroups within studies. As an illustration of how the nature of our empirical research makes this a difficult task, consider the literature on role conflict and role ambiguity. In this literature, results for homogeneous subgroups (e.g., males vs. females or supervisors vs. subordinates) are only infrequently published. Availability of such data would only come about if (a) researchers conducted large survey studies on "pure" groups (e.g., all males), or (b) researchers reported separate correlation matrices for "pure" subgroups comprising larger samples. And, when pure subgroups are used for a study, the dimensions along which the samples are homogeneous are more likely to

reflect constraints imposed by host organizations than constraints imposed by researchers in the interest of theory.

Given that correlational field studies of the type found in the literature on role conflict and role ambiguity tend to rely on heterogenous samples, meta-analysis of results for subgroups within such samples is possible only if separate correlation matrices are reported for the subsamples. Limited journal space and the absence of editorial requirements for full reporting of even a single correlation matrix make it unlikely that reporting practices will change sufficiently to eliminate this problem. For many research areas in organizational behavior, the inability to obtain effect sizes for disaggregated subgroups may be a serious problem that diminishes the value of meta-analysis for theory building.

Another obstacle to the use of meta-analysis for theory generation, especially when contingencies are involved, concerns the state of literature reports in our field. Over a period of years, a large amount of data on homogeneous subsamples should accumulate for topics where there is strong consensus about which variables have theoretical significance. Examples of such topics include personnel selection, job design (Loher, Noe, Moeller, & Fitzgerald, 1985), and Fiedler's (1978) contingency theory of leadership (e.g., see Strube & Garcia, 1981; Peters, Hartke, & Pohlmann, 1985). For topics such as these, journal reviewers and editors are likely to insist on the presentation of disaggregated results. Contrast this situation with the one faced by a meta-analytic reviewer who, after several years of reading studies on a topic, develops a new insight about a moderator variable that could help explain inconsistent findings and wishes to test it using meta-analysis. For example, suppose one wished to test the hypothesis that the relationship between role conflict and performance is moderated by job experience. Because job experience has not been considered a relevant moderator variable in previous studies of role conflict and performance, the necessary data would not be available. Consequently, meta-analytic reviews are more likely to be informative of main effect than of contingent relationships.

A final reason that meta-analysis may not stimulate creative theoretical insights concerns the technique's classical perspective on measurement issues, a perspective that excludes consideration of certain issues. For example, in Hunter et al.'s (1982) approach, variability in relationships between two variables studied in different organizational settings is treated first as a "problem" (of unreliable measurement, of range restriction, etc.) to be "corrected" statistically. As Hunter and Schmidt (1981) state: "Reliability of measurement is a matter of feasibility and practicality independent of the theoretical and psychological meaning of the variables measured" (p. 20). Could it not be possible that variability in observed relationships could be due to *meaningful* organizational differences (e.g., James et al., 1986).

Imagine a meta-analysis of absenteeism. By applying the correction for-

mula for range restriction to one's data, one is likely to increase the effect sizes found for the correlates of absenteeism examined. And by considering the variation in range restriction in absenteeism across studies, one is likely to account for some of the difference in the results of studies under review. Yet this approach to reviewing the absenteeism literature seems foolish if one considers that variance (or lack thereof) in absenteeism and other measured variables may be due to organizational policies and procedures, culture, economic well-being, or other moderating factors.

As meta-analysis becomes accessible via computer software, the application of correction formulae could easily become so "automatic" that questions about the organizational conditions and psychological processes that create the phenomena labeled "unreliability" or "range restriction" are never raised. Whereas range restriction may be a methodological artifact in studies of personnel selection, in the context of other types of organizational research range restriction may be an indicator that should be attended to rather than simply corrected. Corrections advocated by Hunter et al. (1982) are certainly appropriate under some circumstances, but their automatic use is to be avoided.

Theory Testing

The preceding discussion suggests that meta-analysis cannot be expected to play a significant role in the generation of complex contingency theories in organizational behavior. However, meta-analysis can play a role in generating theories that emphasize noncontingent effects. Relatedly, meta-analysis appears to be a highly applicable method of conducting a literature review for the purpose of testing theory. Hypotheses and propositions deduced from existing theories often are empirically testable and an existing body of empirical tests may be usefully reviewed through meta-analysis.

Meta-analysis can advance the field in other ways, also. For example, meta-analytic reviews might highlight results that no longer require replication. Fifty-six samples including over 10,000 respondents indicate that the correlation between role ambiguity and overall job satisfaction is about -.46 (Jackson & Schuler, 1985). Perhaps such information can be used by journal reviewers to legitimately reject manuscripts that once again replicate this result without appreciably expanding our knowledge.

PURPOSES FULFILLED?

Literature reviews are carried out for several purposes and in differing degrees of quantification. In this section we examine how well the purposes are fulfilled by meta-analysis in comparison to less-quantitative forms of literature review.

Summarizing Knowledge

How do various forms of literature review, differing in their quantification of findings, compare for the purpose of summarizing research findings? Although the question cannot be answered definitively, insight can be gained by comparing (1) the results of meta-analytic with non-meta-analytic reviews of comparable literature and (2) the results of different meta-analyses of comparable literature.

Consider the literature on the relationship between job` satisfaction and performance. Vroom (1964) provides a narrative review of 20 studies (reporting a total of 23 correlations) investigating this relationship. The only quantification Vroom undertook was to figure a median correlation between satisfaction (and similar attitudes) and performance. The median correlation was .14, the range was + .86 to - .31, and the correlation was reported to be slightly lower when the performance criterion was a rating than when it was more objective.

Iaffaldano and Muchinsky (1985) conducted a meta-analysis of the relationship between job satisfaction and performance. The meta-analysis was based on 74 empirical studies reporting a total of 217 correlations. Employing corrections advocated by Hunter et al. (1982), the reviewers report an estimated true correlation between satisfaction and performance of .17, a figure quite similar to Vroom's (1964). Further, Iaffaldano and Muchinsky (1985) found little evidence that variations in correlations were moderated by such things as type of worker (e.g., blue vs. white collar) or aspects of research design (e.g., longitudinal vs. cross-sectional). Thus, a meta-analytic and a less-quantitative form of review yielded almost identical conclusions about the relationship between satisfaction and performance.

Another area of research that has been summarized by various forms of literature review is that pertaining to Fiedler's (1978) contingency theory of leadership. Fiedler examined the frequency with which reported findings were in the direction predicted by the theory and found support for the theory. Strube and Garcia (1981) performed a more quantitative review of essentially the same literature, examining the cumulative probabilities that observed findings were due to chance. Their review also found strong support for the theory. More recently, Peters, Hartke, and Pohlman (1985) conducted a meta-analysis of relevant literature and also reported that considerable evidence exists in support of the theory. Thus, three different procedures of literature review, differing in degrees of quantification, all yielded essentially identical conclusions about the validity of the contingency theory of leadership, suggesting that quantification is not an issue in determining the results of a review in this area of research.

Another area of research in which narrative and meta-analytic reviews yielded highly similar conclusions is research on the effect of the similarity of rater and ratee race in performance evaluation. Landy and Farr (1980)

conclude, on the basis of a narrative review, that ratees tend to receive higher rating from raters of the same race than from raters of a different race. Kraiger and Ford (1985) reach the same conclusion through meta-analysis and detect no powerful moderators that effect.

Although the general conclusions reached through a narrative review and a meta-analysis were similar in reviews of productivity improvement programs (Katzell & Guzzo, 1983; Guzzo et al., 1985), some findings emerged from the meta-analysis review that did not result from the narrative review. In particular, the meta-analysis added information about the magnitude of effect of intervention programs-an important datum-not obtainable through the narrative review. Further, the meta-analysis revealed evidence that the impact of productivity improvement programs was moderated by factors such as the type of workers involved and type and size of organization. Thus, in this case, meta-analysis provided information above and beyond that provided by the narrative review. However, the meta-analysis was based on a much smaller sample of studies than the narrative review, and it is not clear whether this difference accounts for the differing conclusions of the two reviews.

There are at least two instances of recently published meta-analyses research on the same topic in organizational behavior yielding different conclusions. One, alluded to earlier, involved the reviews of role conflict and ambiguity conducted by Jackson and Schuler (1985) and Fisher and Gitelson (1983). The former review, based on a larger number of studies, found that the potential causes and consequences of role conflict and ambiguity are likely to be moderated by other variables; the latter review reached the opposite conclusion. A second instance of differing conclusions reached through meta-analyses of similar bodies of literature concerns the relationship between job satisfaction and job performance. Iaffaldano and Muchinsky (1985), as noted earlier, found a modest relationship between satisfaction and performance (average correlation of 17). In contrast, Petty, McGee, and Cavender (1984), using methods similar to laffaldano and Muchinsky (1985), concluded that a much stronger relationship (average correlation of .31) exists. The former meta-analysis contained more studies than the latter (74 vs. 16), was based on a greater number of observed correlations (217 vs. 20), and included in the analysis a greater variety of measures of job satisfaction. All of these factors may help explain differences in the conclusions of the two reviews.

Although the number of cases on which to judge are small, overall it appears that meta-analytic and non-meta-analytic reviews often coincide in their conclusions. A number of interpretations can be made of this. If one accepts quantitative reviews as a standard of objectivity, it appears that narrative reviews are not biased, because they tend to yield the same conclusions as meta-analyses. Alternatively, the similarity of results of the various review procedures may suggest that the costs of meta-analysis

may often outweigh its benefits, because similar conclusions could be reached more efficiently through less-quantitative means of review. Other possible interpretations exist as well. All things considered, the similarity of conclusions reached through quantitative and nonquantitative reviews is encouraging.

More perplexing are those instances in which different meta-analyses of similar literature reach different conclusions. Were meta-analysis a truly objective method of literature review, such differences in conclusions should not have obtained. These differences also speak to the need to create uniform practices for meta-analyses.

Identifying Gaps in Knowledge

All forms of literature review have the potential to make a clear statement about what is *not* known and thus to direct future research and theorizing. Meta-analysis, because of its quantitative emphasis, may have special strengths and weaknesses in this regard. Its strengths lie in its ability to be precise in summarizing relationships demonstrated in a body of literature and, by virtue of this precision, be clear about what research questions have not yet been answered. However, meta-analysis may also fail to detect promising areas of investigation because of its reliance on statistical inference. When statistical power is low, promising "leads" about moderators may be missed. Thus, for example, unexpected relationships between effect sizes and, say, organizational setting may go undetected because of low statistical power.

Advancing Theory

As argued earlier, meta-analysis appears to be a useful method for testing some theoretical propositions and for generating "main effect" theories. Meta-analysis may provide a more powerful means of *testing* existing theory than less-quantitative forms of literature review, especially in testing the existence or strength of theorized main effects. However, meta-analysis appears to be a comparatively weak method of elaborating existing theory by introducing contingencies in a theory's explanations. This is because meta-analysis often is not well suited to detecting moderated relationships among variables. Because of this weakness, less-quantitative forms of literature review may have an advantage over meta-analysis in the generation of new theory.

Policy Implications

Meta-analysis, like other forms of literature review, is valuable becausee it can make clear what actions can be justified on the basis of research results. Because it estimates the effect sizes (or strength of relationship

between variables), meta-analysis can provide potentially useful information not obtainable through less-quantitative review procedures. For example, estimates of the strength of effect of various productivity improvment programs (Guzzo et al. 1985) could be used to estimate the dollar value of such programs, a datum potentially of interest to organizations seeking to raise productivity through human resource management programs. Meta-analysis is thus recommended in those cases where policy decisions will be aided by estimates of effect sizes.

REPRISE: THE STATUS OF META-ANALYSIS

At this writing, meta-analysis is "hot." It is rapidly gaining acceptance as the way to go when reviewing a body of literature, as evidenced by the data of Figure 1. Most of the citations on which Figure 1 is based are studies from clinical, educational, and social/personality psychology. Research in organizational behavior has lagged behind these areas of psychology in the use of meta-analysis, though judging by recent presentations at professional meetings and publications in the field, we see that the use of meta-analysis is on a swift rise.

At several places in this chapter we have offered cautionary notes about the use of meta-analysis and suggestions for coping with the judgment calls and problems it entails. We wish to offer one more cautionary note, embedded in a prediction. It concerns meta-analysis and publication practices in the organizational sciences.

More and more meta-analyses will no doubt appear in print, and their appearance probably will not be restricted to those journals that exist for the purpose of publishing reviews (e.g., *Academy of Management Review, Psychological Bulletin).* Rather, journals that historically have published original research reports (e.g., *Journal of Applied Psychology) will* publish increasing numbers of review papers *when those papers are based on meta-analysis.* Perhaps because of its quantitative form, meta-analysis "looks like" original research and will increasingly find space in journals that have historically published original research. In reality, meta-analysis is but a method of literature review, a quantitative analog of traditional methods of review. We caution against allowing the mere use of a statistical technique in a review paper to influence our perception of the type of contribution made by such a paper and against allowing misperceptions to guide publication practices.

SUMMARY AND CONCLUSIONS

In this final section a summary of conclusions about meta-analysis is offered. Here we compare meta-analysis to traditional-especially narra-

tive-reviews on six key parameters commonly used to assess research practices.

1. Internal Validity

How credible are the results of reviews bearing on the effects of interest? To answer that question, several relevant facets of the review process need to be analyzed.

a. Objectivity. To the extent that the reviews are open to bias or error in interpretation, objectivity is diminished. Among the chief promises of meta-analysis is its alleged objectivity stemming from the perception that, in analyzing the findings, the reviewer need only calculate effect sizes and the role of possible moderators. But subjective judgment inevitably permeates these operations, as the preceding discussion has shown. The interpretations introduce error and bias at numerous points, as does the vagueness of scholarly standards for conducting such reviews. Thus, the potential adverse effect of subjectivity on internal validity is great for both meta-analytic and other forms of review.

b. Precision. Because meta-analysis employs quantitative indices and analyzes them statistically, it promises greater precision than the traditional "eye-balling" of nonquantitative reviews. In practice, though, that precision may not be as high as promised, because of disagreements about the best ways to calculate and analyze effects and a lack of standards of precision. Although a modicum of quantification can appear in other forms of review, meta-analysis has the highest potential for precision in summarizing findings. Kaplan's (1964) point regarding precision in behavioral science research is worth noting: More is not always better. The value of precision depends on the issue at hand, and researchers need not be swayed by "the mystique of quantity" (Kaplan, 1964, p. 204).

c. Control over confounds. A major threat to internal validity is the failure to account for the possible influences of other variables on the relationships of interest. A precaution that can be taken by both meta-analysts and narrative reviewers is a critical reading of the literature to detect this possible source of invalidity. In general, though, it may be more difficult for meta-analysts to handle suspected confounds, compared with narrative reviewers, who need only weigh such factors in their interpretations. Meta-analysts try to handle confounds by considering them as moderator variables, which requires both a reliable way of coding the moderator and a sufficiently large sample of studies to permit an adequate statistical test. The systematic literature review method described by Salipante, Notz, and Bigelow (1982) demonstrates how threats to internal validity can be carefully analyzed in a narrative review.

2. External Validity

Meta-analysis prescribes inclusion of all studies that assess the effect of interest. If that prescription could be followed, its findings would be completely projectable to the population of studies, that is, have high external validity. However, in practice that prescription ordinarily cannot be realized because not all studies report data needed to calculate effect sizes. The result is substantial attrition of the sample. Selective attrition, resulting in a biased sample of studies for review, is particularly troublesome in meta-analyses, especially due to the exclusion of qualitative studies. Narrative reviews need not suffer from this limitation, although reviewers in practice usually are selective in their choice of studies to review. Both kinds of review are likely to lose cases because of deficiencies or choices made in the search process, with corresponding adverse effects on external validity.

3. Statistical Validity

Are the findings and conclusions of reviews clear and dependable, or are they too weak or variable to be accorded much credence? This is an issue of what has been termed *statistical validity*. Meta-analysis tries to deal with this question in at least two ways. One way is to calculate average effect sizes across numerous studies to determine whether average results are sufficiently consistent to be significant. Another is to correct the findings for statistical artifacts.

Narrative reviewers generally have no precise ways of dealing with such problems, although knowledgeable ones undoubtedly try to consider them in drawing their conclusions. There is one respect in which narrative reviews can have greater statistical validity than meta-analyses of the same literature: to the extent that the latter must exclude studies lacking requisite quantitative data, the number of studies on which their findings are based will be smaller and hence the review will have lower power or persuasiveness. A comparison may be made between the two forms of review in terms analogous to Type I and Type II errors of inference. Overall, narrative reviews appear prone to Type I error (i.e., inferring relationships that do not really exist). Hunter et al. (1982) illustrates this tendency in their description of narrative reviews of data concerning the relationship between job satisfaction, job commitment, and circumstantial factors: Narrative reviewers tended to see patterns of relationships in randomly generated data. Meta-analysis, by comparison, appears prone to Type II errors (i.e., failing to detect existing relationships). This is because it relies on stringent rules of statistical inference when statistical power is low. Generally speaking, then, narrative reviewers may detect more in the data

than is really there, whereas meta-analysts may fail to detect what in fact is there.

4. Construct Validity

How firm is the understanding of the reviewers of what is being tapped by the variables under study? Neither meta-analysis nor narrative review procedures have any special approach to dealing with this issue. In this regard, the contribution of any form of literature review depends on the capabilities of the reviewer.

5. Usefulness

Here we consider the review as a basis for theory advancement and for policy decisions. As has been discussed, meta-analysis is hampered by problems in identifying moderated relationships among variables that could contribute to the generation of new theory. Unencumbered by statistical requirements, narrative reviews may be in a better position to detect such relationships, even though these may often be more conjectural than confirmed. Although narrative reviews are often in better position to develop theory, meta-analyses can be useful to test theories.

As regards policy decisions, meta-analysis has the distinct advantage of being able to cite the magnitude of effects to be expected by interventions, thereby helping policy makers evaluate the desirability of potential changes. However, not all policy decisions are rooted in degrees of effect; sometimes, if the effect is at all appreciable (or appears to be negligible), there may be sufficient basis for policy (e.g., where matters of life and death are involved). In such matters, narrative reviews may be adequate.

6. Investment

Most decisions entail assessments not only of benefits but of costs. In deciding between narrative and meta-analytic reviews, how do the two compare? In our experience, the initial, if not total investment in time, effort, and even money is likely to be greater for a meta-analysis. Although both methods require the same initial search for literature, meta-analysis requires the mastery and execution of new methods of analysis, acquiring the necessary techniques, preparing the computer program, translating study findings into effect sizes, coding moderators, and entering and processing the data via computers. Such efforts may have little utility in sonic instances. For example, Latham and Lee (1986) suggest that meta-analysis of studies of the impact of goal setting are not worth the effort because about half of the known studies of goal setting are lost when reviewed

through meta-analysis and existing reviews provide sufficient accounts of the effects of goals (e.g., Locke, Shaw, Saari, & Latham, 1981). On the other hand, as one becomes experienced with meta-analysis, those operations should become more efficient and cost-effective, narrowing the gap between meta-analysis and narrative reviewing. Furthermore, narrative reviews may often require more time in thought, discussion, and writing than does meta-analysis, thereby further narrowing that gap. Generalization is therefore not practical, and cost-benefit analyses should be made in each case.

Overall, then, there is no clear advantage of meta-analysis over narrative reviews or vice versa. They differ in various relevant parameters, and reviewers must decide which approach best meets their objectives and resources, taking account also of the available body of literature. Each approach is likely to continue to occupy a legitimate niche in the field of organizational behavior.

ACKNOWLEDGMENTS

We wish to thank Richard D. Jette for his thoughtful review and discussion of many points in this paper.

NOTES

1. The term *effect size* is used by Glass et al. (1981), Hunter et al. (1982), and Rosenthal (1984) to denote a measure of shared variance or magnitude of relationship between variables, whether the variables measured are truly causes and effects or simply covariates. We use the term *effect size* in this paper to be consistent with prior use, cognizant that effect size estimates often do not refer to causal relationships among variables.

2. The data presented here update the results first presented by Jackson (1984) on the basis of an increased sample of respondents.

3. Note, too, that meta-analysis necessarily excludes data from studies that mix qualitative and quantitative methods, since it can only make use of the quantitative portion of data from such studies.

REFERENCES

Bullock, R.J., & Svyantek, D.J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluational criteria. *Journal of Applied Psychology, 70,* 108-115.

Burke, M.J. (1984). Validity generalization: A review and critique of the correlation model. *Personnel Psychology, 37,* 93-116.

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

Campbell, J.P. (1986). Labs, fields, and straw persons. In E.A. Locke (Ed.), *Generalizing from laboratory to field settings: Research findings from industrial-organizational psychology, organizational behavior, and human resources management (pp.* 269-279). Lexington, MA: Heath-Lexington Books.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cook, T.D., & Leviton, L.C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality, 48,* 449-472.

Cordes, C. (1985). Review may relax testing rules. *APA Monitor, 16 (5),* pp. 1 and 20.

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

Fielder, F.E. (1978). The contingency model and the dynamics of the leadership process. In L. Berkowitz (Ed.), *Advances in experimental social psychology, (pp.* 59-112). New York: Academic Press.

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

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

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

Green, B.F., & Hall, J.A. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology, 35,* 37-53.

Guzzo, R.A., & Bondy, J.S. (1983). *A guide to worker productivity experiments in the United States, 1976-81.* New York: Pergamon.

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

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490-499.

Hogarth, R.M. (1980). *Judgment and choice.* New York: Wiley.

Hunter, J.E., & Schmidt, F.L. (1981). *Cumulating results across studies: Correction for sampling error, a proposed moratorium on the significance test, and a critique of current multivariate reporting practices.* Unpublished manuscript, Department of Psychology, Michigan State University, East Lansing, MI.

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

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

Jackson, G.B. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 438-460.

Jackson, S.E. (1984). *Can meta-analysis be used for theory development in organizational psychology?* Paper presented at the annual meetings of the American Psychological Association, August, Toronto, Ontario.

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 Behavior and Humnan Decision Processes, 36,* 16-78.

James, L.R., Demaree, R.G., & Mulaik, S.A. (1985). *A note on validity generalization procedures. Journal of Applied Psychology, 71,* 440-450.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* Scranton, PA: Chandler.

Katzell, R.A., Bienstock, P., & Faerstein, P.H. (1977). *A guide to worker productivity e ' periments in the United States 1971-75.* New York: New York University Press.

Katzell, R.A., & Guzzo, R.A. (1983). Psychological approaches to productivity improvenncnt. *American Psychologist, 38,* 468-472.

Kraiger, K., & Ford, J.K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology, 70,* 56-65.

Landy, F.J., & Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107.

Latham, G.P., & Lee, T.W. (1986). Goal setting. In E.A. Locke (Ed.), ***Generalizing from laboratory to field settings: Research findings from industrial-organizational psychology, organizational behavior, and human resources management (pp.*** 101-117). Lexington, MA: Heath-Lexington Books.

Locke, E.A. (Ed.) (1986). ***Generalizing from laboratory to field settings: Research findings from industrial-organizational psychology, organizational behavior, and human resources management.*** Lexington, MA: Heath-Lexington Books.

Locke, E.A., & Schweiger, D.M. (1978). Participation in decision making: One more look. In B.M. Staw & L.L. Cummings (Eds.), *Research in Organizational Behavior (Vol. 1)* (pp. 265-339). Greenwich, CT: JAI Press.

Locke, E.A., Shaw, K.N., Saari, L.M., & Latham, G.P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin, 90,* 125-152.

Loher, B.T., Noe, R.A., Moeller, N.L., & Fitzgerald, M.P. (1985). A meta-analysis of the relation of job characteristics to job satisfaction. *Journal of Applied Psychology, 70,* 280-289.

Masters, J.C. (1984). Psychology, research, and social policy. *American Psychologist, 39,* 851-862.

McGrath, J.E. (1982). Introduction. In J.E. McGrath, J. Martin, & R.A. Kukla (Eds.), *Judgment calls in research.* Beverly Hills, CA: Sage.

Mintzberg, H. (1973). *The nature of managerial work.* New York: Harper & Row.

Peters, L.H., Hartke, D.D., & Pohlmann, J.T. (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.

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.

Salipante, P., Notz, W., & Bigelow, J. (1982). A matrix approach to literature reviews. In B.M. Staw & L.L. Cummings (Eds.), *Research in Organizational Behavior (Vol.* 4) (pp. 321-348). Greenwich, CT: JAI Press.

Sashkin, M'. (1984, Spring). Participative management is an ethical imperative. *Organization Dynamics,* 5-22.

Schmidt, F.L. (1984). *Meta-analysis: Implications for cumulative knowledge in the behavioral and social sciences.* Paper presented at the annual meetings of the American Psychological Association, August, Toronto, Ontario.

Schmidt, F.L., & Hunter, J.E. (1978). Moderator research and the law of small numbers. *Personnel Psychology, 31,* 215-232.

Smith, M.L., & Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.

Stock, W.A., Okun, M.A., Haring, M.J., Miller, W., Kinney, C., & Seurvorst, R.W. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. *Educational Researcher, 11,* 10-14.

Stone, E.F. (1986). Job-scope satisfaction and job-scope performance relationships. In E.A. Locke (Ed.), *Generalizing from lab to field settings: Research findings from industrial-organizational psychology, organizational behavior, and human resource management* (pp. 189-206). Lexington, MA: Heath-Lexington Books.

Stone, E.F. (1985b). Personal communication.

Strube, M.J., & Garcia, J.E. (1981). A meta-analytic investigation of Fiedler's contingency model of leadership effectiveness. *Psychological Bulletin, 90,* 307-321.

Tanford, S., & Penrod, S. (1984). Social influence model: A formal integration of research on majority and minority influence processes. *Psychological Bulletin, 95,* 189-225.

Tornow, W.W., & Pinto, P.R. (1976). The development of a managerial job taxonomy: A system for describing, classifying, and evaluating executive positions. *Journal of Applied Psychology, 61,* 410-418.

Vroom, V.H. (1964). *Work and motivation.* New York: Wiley.

Winkler, R.C., & Winett, R.A. (1982). Behavioral economics. *American Psychologist, 37,* 421-435.