Reviewing the literature: A comparison of traditional methods with meta-analysis[1]

Thomas D. Cook and Laura C. Leviton, *Northwestern University*

Abstract

Meta-analysis is the name given to a set of techniques for reviewing research in which the data from different studies are statistically combined. Meta-analysts have criticized the more traditional qualitative methods of review on three principal grounds: (1) that relevant information is ignored in favor of a simplistic box count of the number of studies in which a particular relationship is and is not statistically significant; (2) that the sample of studies for review often contains important biases; and (3) that box counts ignore statistical interactions. Our discussion suggests that these criticisms are not intrinsic to qualitative reviews, but rather represent poor practices by reviewers using traditional methods. Moreover, although meta-analysis has some advantages, it is not without its unique limitations. Our comparison of both methods is applied to the qualitative literature review of Zuckerman (1979) and the meta-analysis of Arkin, Cooper, and Kolditz (1980) which reached different conclusions about the "existence" of self-serving attributions in studies of interpersonal influence.

Literature reviews have long played a central role in scientific development. Science is a cumulative endeavor. In psychology, any one study is suspect, but because of the impossibly large array of validity threats that must be ruled out (Cook & Campbell, 1979), and also because strong tests of a theory require that predictions be made about multiple empirical relationships that compose a "nomological net" (Cronbach & Meehl, 1955). One purpose of literature reviews is to establish the "facts." These are the stubborn, dependable relationships that regularly occur despite any biases that may be present in particular studies because of the implicit theories behind the investigator's choice of measures, observation schedules, and the like (Stegmuller, 1978). At present, psychologists largely depend on the qualitative literature review to establish

1. Reprint requests should be sent to the first author, Department of Psychology, Northwestern University, Evanston, Ill. 60201. We would like to thank Drs. Albert Erlebacher, Kenneth Howard, and Daniel Romer for their assistance and suggestions. Partial support was received from the U.S. Department of Agriculture, Contract #53-3198-9-26.

Journal of Personality 48:4, December, 1980.

"facts." Such a review requires drawing up a list of theoretically relevant studies, examining each study for "methodological adequacy," and then counting the number of adequate and relevant studies which confirm and disconfirm a particular relationship. This last step is called the "box count," and typically the box in which the highest count falls is "voted" the winner (Light & Smith, 1971).

For reasons we discuss later, some researchers have expressed dissatisfaction with qualitative reviews. Glass (1978b; Glass & Smith, 1979; Smith & Glass, 1977) and others propose a set of alternatives called "meta-analysis" which brings standard data analytic techniques to bear in reviewing research. To do this, studies relevant to a conceptual issue are collected, summary statistics from each study (e.g., means or correlations) are treated as the units of analysis, and the aggregate data are then analyzed in quantitative tests of the proposition under examination.

Meta-analysis has created recent controversy because studies undertaken with this orientation have repeatedly reached less conservative conclusions about the presence and magnitude of particular conceptual relationships than have traditional literature reviews.[2] For instance, Glass and Smith (1979) concluded that class size affects the achievement of children, while qualitative reviews were inconsistent in their conclusions (Educational Research Service, 1978; Ryan & Greenfield, 1976). Also, Smith and Glass (1977) concluded that all the types of psychotherapy they examined were effective in improving outcomes by at least one-half of a standard deviation; whereas qualitative reviews have generally disagreed both on the effectiveness of specific therapies and the effectiveness of psychotherapy overall (Bandura, 1969; Bergin, 1971; Eysenck, 1952; Luborsky, Singer, & Luborsky, 1975; Rachman, 1971). Because meta-analysis appears to be less conservative (and less "subjective") than qualitative reviews, pressure may arise in the near future to make meta-analysis obligatory for many kinds of dissertation, grant proposal, and research report, as well as for reviews in journals such as *Psychological Bulletin*.

A controversy in the *Journal of Personality* reflects the need to examine the merits of meta-analysis more closely. Using qualitative techniques, Zuckerman (1979) concluded that studies involving subjects' influence over other persons did not show self-serving

2. It is probably the case that qualitative reviews are more conservative in their conclusions than meta-analyses. Cooper and Rosenthal (1980) tested this hypothesis by randomly assigning graduate students and faculty members to review a set of related studies, using either a meta-analytic technique or the traditional qualitative method. Reviewers who used meta-analysis believed that there was more support for the phenomenon under study than did qualitative reviewers.

biases in attribution. But employing meta-analytic methods, Arkin, Cooper, and Kolditz (1980) concluded that such studies generally did show the bias in question. We have been asked by the editors of *Journal of Personality* to comment upon the debate between Zuckerman and Arkin, Cooper, and Kolditz by relating it to issues about meta-analysis in general. To do this, we shall describe the major meta-analytic techniques, compare them to traditional methods of review, and then use this comparison to comment upon the work of Zuckerman and of Arkin, Cooper, and Kolditz.

Some Meta-Analytic Techniques

Space does not permit us to describe either the full range of meta-evaluative techniques used by scholars (Cook & Gruder, 1978) or all the available meta-analytic techniques using data analysis. Detailed technical descriptions can be found elsewhere (Cooper, 1979; Glass, 1976, 1978b; Light & Smith, 1971). Our purpose here is to supply only enough detail for the reader to gain an idea of the purposes that each method serves.

Glass and his colleagues have developed a method of meta-analysis that relies on effect sizes as the unit of analysis. Effect sizes are computed for each study by taking the difference between the means of experimental and control groups (or some proxy for this) and dividing by the standard deviation. Glass and his colleagues then take the average of the effect sizes as their summary statistic, which is expressed in standard deviation units. By converting the findings of studies to a common metric, the estimate of average impact can be based upon a wide range of conceptually related measures and hence a larger sample of studies.

Although Glass and his colleagues use the average of effect sizes as their summary statistic, they also examine some of the methodological and substantive characteristics of studies on which the size of the effect may depend. Thus, in their meta-analysis of psychotherapy, Smith and Glass (1977) divided therapies into behavioral and nonbehavioral types and discovered that the average effect size was similar for each type. Glass and his colleagues have also employed regression analysis to make sure that effect sizes were not highly correlated with irrelevant attributes of studies. For example, the psychotherapy meta-analysis showed that an index of methodological quality accounted for only one percent of the variance in effect sizes (Glass, 1978c). This implies that, *within the limits imposed by the validity of the index,* studies of higher and lower quality reached similar conclusions about impact.

Rosenthal (1969, 1976) developed the "package" of meta-analytic techniques used by Arkin, Cooper, and Kolditz to examine self-serving biases in attribution, also used by Cooper (1979) to ascer-

tain the conditions under which sex differences in conformity appear, and also by Rosenthal and Rubin (1978) to summarize evidence for the experimenter expectancy effect. This package consists of four elements. First, the probability values from each test of the relevant theoretical hypothesis are combined to reflect the probability, across all the studies, with which the null hypothesis can be rejected (Mosteller & Bush, 1954; Stouffer, 1949). From this same method of combining probabilities, Rosenthal and his colleagues are able to estimate the number of studies with "no effect" conclusions that would be necessary to change the obtained probability level to .05. This second element provides some assurance that, even if some studies are missed by the review, the results of the meta-analysis will not be seriously biased in favor of reaching conventional statistical significance levels.

Rosenthal's emphasis on aggregating probability values reflects his concern with testing specific theoretical hypotheses—in his case about experimenter and teacher expectancies. While combining p values is one reasonable procedure for postulating "whether an effect exists," it is not comprehensive. On the one hand, hypothesis testing does not describe the magnitude of a relationship, and on the other, it is overly dependent on sample size. For these reasons, Rosenthal now includes a third element in his package, the calculation of effect sizes, in much the same manner as Glass and his colleagues. Finally, Rosenthal and his colleagues employ a table presented by Cohen (1969) which shows the degree of overlap in the distributions of the experimental and control groups. A large average effect would indicate little overlap between the two distributions.

Light (1979; Light & Smith, 1971) has suggested a purpose for meta-analysis that is quite different from that of Rosenthal and Glass. The details of how to conduct such an analysis, and a critical examination of these details, are still lacking. While Glass and Rosenthal emphasize broad summary statements about areas of research (e.g., psychotherapy does have a beneficial effect), Light deliberately deemphasizes such statements. Instead, he advocates exploring the data from a set of similar studies to determine the populations or settings in which specific treatments do and do not have effects. To do this, he examines raw data from studies with identical measures rather than summary statistics from studies with different but conceptually related measures. Light's claim is that his procedures permit more flexibility in exploring the data, though they reduce the sample size of studies.

Because he wants to generate contingency-theoretic statements about treatments, settings, and populations, Light's approach holds

potential for social psychologists who are interested in investigating the variability of findings over time and settings (Gergen, 1973), the interaction between personality traits and situations (Mischel, 1968), and the use of interactions to help specify external validity and the construct validity of manipulations, measures, and relationships (Cook & Campbell, 1979).

Criticisms of Qualitative Review Methods

Meta-analysts justify their techniques, in part, by pointing to some failings of the typical "box count" as it is employed in qualitative reviews. They make three principal criticisms about traditional review techniques. First, some information is ignored; second, the sample of studies may be biased; and third, statistical interactions may not be detected. We will now elaborate on these criticisms, citing qualitative reviews that can justly be criticized on each of these grounds. For each criticism, we then ask whether it reflects an intrinsic limitation of the qualitative review or a common poor practice that new reviews could change. Finally, we ask whether meta-analysis is itself flawless in these three regards and point to some ways in which it is not.

Information ignored by qualitative review. Light and Smith (1971) argue that the typical review ignores information about the magnitude of relationships and about their direction in favor of simple counts of the frequency with which hypotheses are and are not confirmed at conventional levels of statistical significance. Counting the incidence of significant p values is overly conservative. This is because results which are in the right direction but fail to reach statistical significance will be counted as failures to corroborate the hypothesis. Yet the statistical power of the tests in question may be so low that unrealistically large effects would have to be obtained to produce significance with the particular sample size. In psychology—with its tradition of laboratory experiments using small samples—a reliance on counts of statistical significance will predispose qualitative reviews to a modal finding of "no difference." That the power of research in social psychology is often low was illustrated by Cohen (1962) in a review of studies published in the *Journal of Abnormal and Social Psychology.*

Effect sizes and other estimates of magnitude are less dependent on sample size than are significance tests. Moreover, even if effect sizes are small and not significant for individual studies, they add information if the results are in the same direction as other studies. Indeed, Cronbach and Snow (1976) have pointed out that many unbiased studies which individually show no significant differences can lead to rejection of the null hypothesis if most of them

have mean differences in the same direction. Yet in many box counts the direction of nonsignificant results is not even reported. Thus, in an analysis of a random sample of literature reviews from the behavioral sciences, Jackson (1978) found that only four of 28 reviews reporting the results of individual studies also included information about the direction of nonsignificant findings. If readers wanted to eliminate the box labeled "not statistically significant" and instead restrict the count to relationships with one sign and those with the opposite sign, they could seldom do this from the information in published reviews.

Reviews of psychotherapeutic outcomes illustrate the consequences of ignoring effect sizes and the direction of findings, since in these reviews the heavy reliance on box counts of significance tests may partly explain the discrepancy between the findings of qualitative reviewers and those of Smith and Glass (1977). For example, Bergin's (1971) influential review relied solely on statistical significance to assign studies to boxes. In some studies in Bergin's review, the sample of subjects was very small—as low as 10 and 12. While most studies had at least 30 subjects, even 30 is small when half the subjects constitute a control group. Moreover, studies of over 400 subjects appear to be weighted equally with small sample studies, and no cognizance appears to be taken of the issue of statistical power. Interestingly, magnitude estimates do appear in some reviews of psychotherapeutic outcomes, in the form of correlations (e.g., Luborsky et al., 1971). However, they are not the reviewers' primary source of information in reaching conclusions; and prior to Smith and Glass (1977) no one thought to aggregate the correlational studies because the measures were so heterogeneous (Howard, Note 1).

Is the overdependence on statistical hypothesis tests intrinsic to qualitative reviews, or does it merely reflect poor practices? The latter is the case. Some qualitative reviews make extensive use of effect sizes. For instance, Mischel (1968) used effect sizes in the form of correlations to argue for an interaction between personality traits and the situation in predicting behavior. Cartwright (1971) also used a form of magnitude estimates in showing that group decisions in the risky shift paradigm were of trivial magnitude—before he went on to show that an artifact was probably responsible for any effect there appeared to be. Qualitative reviews do use information about the direction of results (Cook et al., 1979) and sometimes differentially weight findings by sample size of the studies (Jencks et al., 1972). Criticisms of qualitative reviews based on misunderstandings of statistical significance are not criticisms of the review method. Indeed, many qualitative reviews share the

same assumptions about the role of statistical significance testing that lead some meta-analysts to ignore statistical significance in favor of estimates of effect size.

If reviewers were sensitized to the problems of significance tests, these bad practices might not be so pervasive. However, a strong tendency exists in research and training in psychology, with its traditional emphasis on testing theoretical hypotheses with small samples, to accept only significant findings and not even to report, let alone discuss, the direction of nonsignificant effects or effect sizes. In this respect psychology is markedly different from economics, where statistical significance is of trivial importance and effect size is dominant; or even from education, in which recognition of the importance of effect sizes is increasing (Cronbach & Snow, 1976).

Both meta-analysis and qualitative reviews can make proper use of effect sizes and the direction of findings. For reviews with a moderate number of studies, perusal of a box count of effect sizes could well lead to the same conclusion as a meta-analytic summary. Indeed, we hope to demonstrate that this is the case for the review by Arkin, Cooper, and Kolditz. For reviews with a great many studies, we are prepared to agree that meta-analysis can be more usefully employed, for convenience and to avoid cognitive overload (Glass, 1978b).

Meta-analysis has one potential disadvantage in its reliance on effect sizes and its claim for greater precision than qualitative reviews (Arkin, Cooper, & Kolditz, 1980; Jackson, 1978). In order to take the precision of an effect size seriously, one must assume that there is equal bias across studies. That is, irrelevancies that inflate a relationship in one direction in some studies are counterbalanced by equally potent irrelevancies inflating it in the opposite direction in other studies. Alternatively, meta-analysts assume that the degree of bias can be validly estimated, and that studies can be weighted to eliminate the bias (Cooper, 1979; Mosteller & Bush, 1954). Some methods that meta-analysts use to detect bias are given in the next section. For the moment, we note that if the assumption of counterbalanced biases is wrong, a misplaced specificity will result. While qualitative reviews may be equally prone to bias, the descriptive accuracy of a point estimate in meta-analysis can have mischievous consequences because of its apparent "objectivity," "precision," and "scientism." To naive readers, these lend a social credibility that may be built on procedural invalidity.

Bias in the sample of studies. Qualitative reviews are also criticized for introducing potential biases into the sample of studies reviewed. A reviewer may introduce bias in at least three distinct

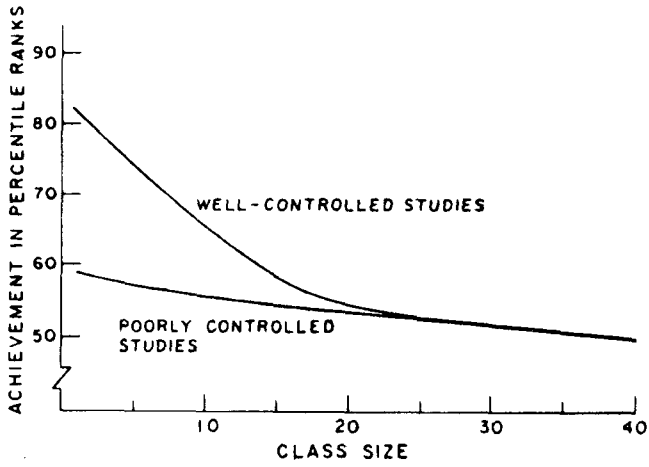

Figure 1. Effects of randomized and nonrandomized studies. Consistent regression lines for the regression of achievement (expressed in percentile ranks) onto class size for studies that were well controlled and poorly controlled in the assignment of pupils to classes.[3]

ways: the literature search may be so narrow as to omit relevant studies; some of the discovered studies may be excluded on methodological grounds; or studies may be excluded because the theoretical constructs are considered irrelevant. These three sources of bias are discussed below. Glass (1978b) notes that when studies are excluded on any grounds, it is not possible to ascertain whether the sample has become biased. His working assumption, which is shared by other meta-analysts, is that all the available published and unpublished studies of general substantive relevance should be included, and that the data should tell us whether different methods or different subsets of studies are associated with different magnitude estimates.

Meta-analysts like Rosenthal (1979) and Arkin, Cooper, and Kolditz (1980) stress the importance of a thorough literature search. Both published and unpublished studies (dissertations, for example) should be examined because editorial policies may bias published studies towards confirming replications or statistically significant studies. A recent meta-analysis by Smith (1980) supports

3. From G. V. Glass and M. L. Smith, Meta-analysis of research on class size and achievement, *Educational Evaluation and Policy Analysis*, 1979, **1**, 2–16. Reprinted by permission.

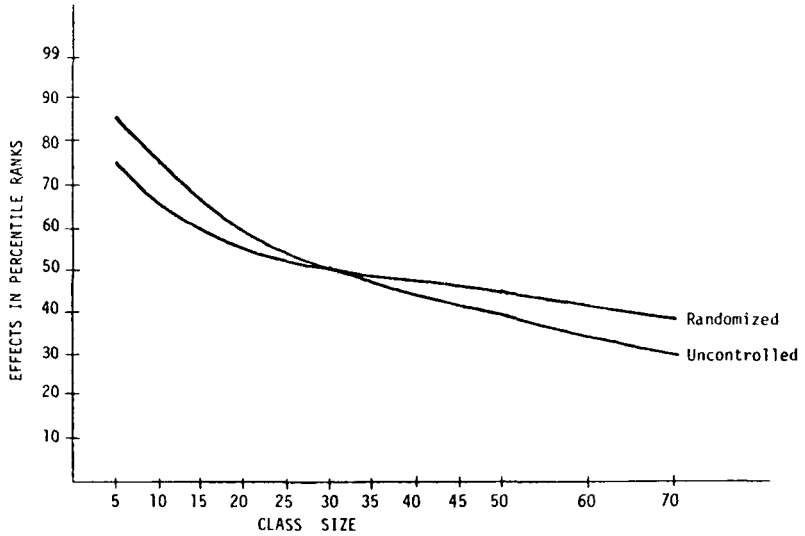

Figure 2. Effects of randomized and nonrandomized studies. Graph of the relationship of class size and effects on attitudes and instruction for studies using randomization versus uncontrolled studies (studies using matching or repeated measures produced intermediate effects and are not plotted).[4]

this contention. Published studies showed a .25 standard deviation tendency for counselors and therapists to stereotype women and view them more negatively than men, while unpublished studies showed an effect of equal magnitude for a bias against men.

Glass notes that in qualitative reviews studies are often excluded a priori for methodological reasons. This would perhaps be a reasonable procedure if all issues of methodology were resolved and no debates about methods could be heard. But such is not the case. Hence, Glass believes that if studies presumed to differ in methodological quality come to the same conclusions, then there is no point in excluding the studies of presumed lower quality. If different conclusions are reached, Glass is then prepared to give more weight to the higher quality studies. Consider Figures 1 and 2, which are taken from studies of the effects of class size on achievement (Glass & Smith, 1979) and nonachievement outcomes (Smith & Glass, 1979). Studies employing random assignment produced a pronounced effect of class size on achievement, while nonrandom

4. From Glass and Smith (1979). Reprinted by permission.

studies at best produced the effect very weakly. It seems, then, that different methods produced different conclusions, and that the results of the randomized studies should be treated as more valid. For nonachievement outcomes, on the other hand, both random and nonrandom studies showed similar effects of class size.

A bias in favor of narrowly defined constructs runs through many qualitative reviews. On the other hand, meta-analysts tend to prefer broader constructs (Arkin, Cooper, & Kolditz, 1980; Glass & Smith, 1979; Rosenthal & Rubin, 1978; Smith & Glass, 1977). Glass claims that broader constructs let the data decide whether relationships need respecifying in less global terms. Such respecification would be indicated, for example, if psychotherapy was beneficial as shown by questionnaires but not by therapists' ratings, or if "helping paradigms" resulted in attributions about performance that differed from the attributions found with other paradigms.

Letting the data specify cause and effect constructs leads to the criticism that meta-analysis aggregates "apples and oranges." In response, Glass notes that such aggregation is useful for the study of "fruit," and that it is often more useful to have knowledge about broader constructs than about subcategories. Thus he writes:

> In a field that lacks standard units of treatment and measurement, such as social psychology, all empiricism and reasoning is a problem at some level of coping with incommensurables. . . . It is a common bad habit of thought in most psychological fields to press more distinctions on investigators than can be usefully made—distinctions without real or important underlying differences. (1978a, p. 395)

According to Smith and Glass (1977), the psychotherapy literature is an excellent illustration of the consequences of excluding studies from review because they do not employ preferred methods and constructs. In their own work, Smith and Glass located over 475 studies. Yet they note that qualitative reviews typically employ fewer than 100, often fewer than 40 studies; and they reach conflicting conclusions about the effectiveness of psychotherapy.

Is bias in the sample of reviewed studies intrinsic to qualitative reviews, or does the bias merely reflect poor practices? Although meta-analysis has highlighted afresh the importance of comprehensive samples, many excellent qualitative reviews have exhaustive samples. Moreover, like quantitative reviews, qualitative reviews can be self-correcting with respect to sampling bias. Thus, Jeanne Block (1976) located some studies of sex differences that Maccoby and Jacklin (1974) omitted from their mammoth review of over 1600

studies. According to Block, including these studies changed some of the original conclusions.

In addition, it is not inevitable that qualitative reviews take the narrow view and include only studies that meet restricted methodological criteria. For example, in particular substantive domains within evaluation research Boruch (1977), Campbell and Erlebacher (1970), and Director (1979), among others, have assessed the relative importance of randomized experiments compared to other alternatives. They did this by including all types of experiment and quasi-experiment in their review, examining the results for each type in a qualitative manner after having disaggregated the studies by method type, much as Glass did in his more quantitative work on class size.

Finally, qualitative reviews can take issue with the breadth or narrowness of constructs. Campbell's (1950) review of the attitude literature used qualitative methods to draw attention to the narrow use of self-report measures of attitudes, and to suggest that multiple measures of attitude were feasible and necessary if conclusions were to be drawn about a general construct, "attitude," rather than about specific constructs like "paper and pencil measures of attitude." Moreover, qualitative reviews can use broad conceptions of constructs and then "let the data decide" whether narrower definitions are necessary. For example, Block (1976) concluded from data presented by Maccoby and Jacklin (1974) that sex differences in cognitive restructuring emerged differentially for two different methods, indicating that narrower constructs were necessary for explaining the obtained data.

Sensitizing traditional reviewers to sampling bias may help them to locate studies from a wider variety of sources. Also, it may help them realize that the methodological criteria by which they evaluate a single study need not be as stringent as those they now use for including studies in a review. The major criterion for trusting the results of a single study is the appropriateness of the chosen methodology to the problem and the care with which the methods have been implemented. A review, on the other hand, can admit less successful studies provided that the assumption of zero overall bias is accepted or the set of studies can be broken down into those employing better and worse methods.

Meta-analysis has a distinct advantage over qualitative reviews when a large number of studies exist. This creates a situation of cognitive overload, particularly when many cross-tabulations are required to test for differences among methods or constructs (Cooper, 1979). The problems that arise from the multiple classification

of variables is compounded when the variables under investigation have multiple levels. Consider the literature on class size. With class sizes varying from 1 to 50 it would be exceedingly difficult to make all possible comparisons of class size in a qualitative manner. For large samples, meta-analysis enhances the ease with which data can be manipulated and increases the ability to compute relationships that test whether particular irrelevancies of method or particular substantive variables influence the dependent variable in ways that aid interpretation.

If we turn now to the limitations of meta-analysis, it is worth repeating that meta-analysis cannot in and of itself detect bias that is preponderantly in one direction. The number of studies is a conceptual irrelevancy when the studies share sources of bias that predominantly operate in one direction. Indeed, the number of studies may even represent a psychological trap, for as the number of studies increases so too does the likelihood of invalid inductive leaps that create credible conclusions out of spurious findings. Campbell and Erlebacher (1970), and Director (1979) have illustrated this particular problem in their work on different substantive topics where "compensatory" treatments are assigned to persons who tend to have lower scores at the pretest when compared to no-treatment controls. This results in a constant bias to make the treatment look harmful. Wicker's (1969) review of the attitude-behavior relationship provides another example of a bias that prevailed across most of the studies in a review and produced an apparently credible consistency that in fact reflects bias. His conclusion that attitudes and behavior were only weakly related was based on the fact that in study after study their correlation seldom exceeded .30. However, errors of measurement attenuate the magnitude of relationships; and in most of the studies Wicker reviewed, reliabilities were probably closer to .5 than to 1.0. Had Wicker expressed the obtained correlations corrected for attenuation (where possible) or relative to the obtainable upper bound, his correlations would have been larger. Jack Block (1976) has made the same point about Mischel's (1968) review of the literature on consistency of individuals' behavior across situations.

An incidental negative side effect of the meta-analysts' concern with broad definitions of constructs may be a disregard for theoretical relevance and a vulnerability to misleading inductive inferences. To prevent this, meta-analysis should provide a breakdown by theoretically relevant variables. For example, if all past studies of the sleeper effect were included in a review without regard to how well the necessary conditions for a strong test of the effect were met, one would probably conclude from the many failures to

obtain the effect that "it is time to lay the sleeper effect to rest" (Gillig & Greenwald, 1974). And the number of failures to obtain the effect would be psychologically impressive. Yet the very few studies that demonstrably met the theory-derived conditions for a strong test of the effect all obtained it (Cook et al., 1979), and one can surmise that these studies should be assigned greatest inferential weight because of their demonstrably higher theoretical relevance. A similar point can be made for the dissonance literature on the effects of low payments. A meta-analysis prior to 1972 would probably have concluded that the evidence for the dissonance effect was weak or inconsistent, because it seemed that for every study with posttest means in favor of the dissonance hypothesis, there was a study with means in favor of the opposite (so called, incentive) hypothesis. However, once the importance of choice and personal responsibility were fully recognized (Collins & Hoyt, 1972), the dissonance phenomenon came under theoretical control. We would consider as poor practice any literature review, qualitative or quantitative, that failed to exclude studies that were irrelevant on theoretical grounds, or that did not test for possible irrelevancies by disaggregating studies into those that provide stronger and weaker tests of theory.

We believe that meta-analysis is inappropriate for very small samples of studies if the studies are heterogeneous with respect to methods and constructs. Such heterogeneity implies that the studies are best considered as constructual replications in the broadest sense (Lykken, 1968), and with constructual replications the likelihood of method and construct factors limiting the degree of obtained correspondence is high. Yet with small samples one cannot test directly whether or not this is the case. If, on the other hand, researchers strive for "exact" replication, more confidence can be placed in small samples because the studies, while not identical, will be "somewhat" similar. Indeed, Light and Smith (1971) would often be prepared to assume that with "exact" replications one has increased the sample size of a single study. The conundrum here is that constructual replications are more valuable than exact replications, for constructual replications make heterogeneous many more of the methodological and substantive irrelevancies that can condition a particular relationship. Yet when studies are heterogeneous, we must have larger samples to be confident that the studies are representative of an underlying distribution (Gilbert, McPeek, & Mosteller, 1977). Note that, for research on novel topics where only a handful of studies exist, the sample is rarely "representative" of how later studies will be—much as the first five subjects in a study need not be "representative" of the entire sample.

The central limit theorem applies to samples of studies as to samples of subjects.

Theory building and sensitivity to interactions. Light and Smith (1971) have claimed that the traditional box count is insensitive to detecting statistical interactions. If so, this would be a serious drawback to traditional review methods, because statistical interactions play a crucial role in the testing of psychological theories. Most such theories deal with intervening variables (in the MacCorqudale & Meehl, 1948, sense) that will never be directly measurable, such as "anxiety," or "dissonance." Inferences about such constructs depend on postulating complex patterns of data—the "nomological net" of Cronbach and Meehl (1955)—and then seeing if obtained data patterns match the expected pattern. A further need is to rule out alternative explanations of the relationships, which is more likely if multivariate predictions are made that seem unique. Within personality and social psychology, interpretation of the forced compliance literature depends on complex interactions (Calder, Ross, & Insko, 1973; Collins & Hoyt, 1972), as does research on leadership effectiveness (Fiedler, 1964), or the interaction of situational and personological variables (Block, 1976; Mischel, 1976).

In Tukey's (1969) language, reviewers can use interactions either to confirm or explore relationships. They confirm relationships through two strategies. The first is between-studies. Reviewers search for studies that did, and studies that did not, manipulate or measure the independent variables that are crucial to the theory. The outcomes of these two sets of studies are then contrasted. The second strategy is within-studies; the reviewer conducts a standard box count or meta-analysis of outcomes in the subset of studies in which the potential interacting variables were manipulated or measured.

In the exploratory mode, reviewers examine sets of studies and try to infer any relationships that seem worth further exploration. They become detectives who use obtained data patterns as clues for generating potential explanatory concepts that specify the conditions under which a positive, null, or negative relationship holds between two variables. Light (1979) stresses such an exploratory orientation when he describes Title I evaluations in which over 50 classrooms showed achievement gains and over 50 showed losses: "In fact, do such results indicate random outcomes? I think not. Even though an 'on the average' analysis may show zero change, a study of the distribution of outcomes shows that if outcomes were random, with no underlying program effect we would expect about five significant gains and five significant losses . . ." (p. 10).

For a number of reasons, it is difficult to estimate how often qualitative reviews ignore interactions of importance. First, many reviews set out to test simple main effect propositions, and if no conclusions emerge for interactions, this may be because the interactions are considered irrelevant to the major research question. Second, in theoretical work a recognition of the need for interaction predictions often evolves slowly as the guiding theory is modified to account for past explanatory failures. When this happens, little or no information of high quality may be available from past studies about potential contingency variables. The reviewer is left with only a small sample of more recent work that can be used for sensitively testing or detecting interactions.

Qualitative reviews can detect interactions when the guiding theory is specific enough and the past studies can be organized into box counts based on whether a particular interaction did or did not occur. Collins and Hoyt (1972) did this in their "confirmatory" review of past studies of forced compliance. Qualitative reviews can also handle the exploratory mode of reviewing, and in the same way that Light (1979) suggests. The best narrative reviews identify and explain contradictory and unexpected data patterns for which no specific boxes were initially set up. Thus, Zajonc (1965) was able to develop his theory of social facilitation by bringing order into a disparate group of studies whose common features were the performance of a behavior and variation in the number of others present. He first noted an apparent contradiction, enhancement of performance in some studies but a decrement in others. He then became a detective and resolved the puzzle by illustrating that increases in performance occurred when skills were overlearned, and that decreases in performance occurred when skills were not overlearned. Extant theories of "arousal" were then used to give a theoretical explanation of the interaction that Zajonc discovered.

Glass and Rosenthal focus primarily on global summary statements rather than interactions. The interactions they do examine are based on disaggregating studies. Such disaggregation can bias the analysis towards an examination of nominal, demographic, and gross empirical characteristics such as methodological characteristics of studies, procedural details, and the sex and age of respondents. Theoretical variables are less likely to be considered. This bias is probably the result of limitations in the data base of studies for review, because potential interaction variables of theoretical interest may not be measured in every study. This does not deny the utility of interactions involving demographic variables. They often offer clues to the underlying psychological relationships that

cause interactions. But they are only clues and are less direct tests of theoretical propositions than occur when a theoretical construct is directly measured.

The possibility of an inferential trap occurs when a meta-analysis across several procedural paradigms is statistically significant or reflects a large effect size, but the effect is specific to a subset of the paradigms. When this happens, some analysts tend to stress the main effect rather than the interaction that qualifies the main effect. For example, Cooper (1979) concluded that, over all paradigms, women conformed more than men. Yet a separate meta-analysis of each paradigm showed that the sex difference was confined to the set of 16 studies involving face-to-face interaction. No evidence was found with the 8 studies using fictitious group norms, and only weak trends in favor of greater female conformity were found in the 14 persuasive communication experiments. Making statements about a main effect could be interpreted to mean that greater conformity by females is transsituational, whereas it is situation-specific. Moreover, had there been fewer face-to-face studies, the obtained effect would have been smaller and perhaps not even statistically reliable. It is not comforting to think that conclusions about the generality of a main effect depend on the accidental rate at which face-to-face situations happen to have been chosen over other situations in past research. In Cooper's defense it should be noted, first, that he acknowledged the differences among procedural paradigms, and second, that Maccoby and Jacklin (1974) can be interpreted as having drawn a conclusion of no main effect across paradigms, even though they stressed the situations that predict sex-linked conformity. Our guess is that, with their stress on broad generalization, meta-analysts are even more prone than qualitative reviewers to overlook or to down play the importance of contingency-specifying interactions that in most situations have an inferential precedence over statements about main effects.

Application of These Observations to Zuckerman (1979) and Arkin, Cooper, and Kolditz (1980)

We now ask whether each of the major criticisms of qualitative reviews applies to Zuckerman's review, and whether the review of Arkin, Cooper, and Kolditz suffers from any of the disadvantages of meta-analysis. To illustrate our comparison, we refer the reader to Table 1 of Arkin, Cooper, and Kolditz (p. 438).

Zuckerman reviewed 13 studies concerning the attributions subjects made for their success or failure in influencing another person. Zuckerman divided these studies into three subcategories. The

first category included five studies that used a paradigm in which the subject was a teacher and the target person a student (teacher-learner). The second category had four studies in which the subject was a therapist and the target of influence a client (client-therapist). The remaining four studies did not share a common theme or procedure (miscellaneous). These 13 studies are indicated by an asterisk in Table 1. After reviewing them, Zuckerman concluded that there was little evidence overall for a bias toward attributions that enhanced self-esteem in studies of interpersonal influence.

Arkin, Cooper, and Kolditz performed a meta-analysis of these 13 studies, together with 10 others that were not reviewed by Zuckerman. From the two articles, anonymous reviews, Zuckerman's review of Arkin et al., and Arkin's response, we gather that four of these ten studies appeared after Zuckerman's review was completed; four were potentially available but were omitted by Zuckerman; and two were omitted because—unlike Arkin, Cooper, and Kolditz—Zuckerman judged that their subject matter, helping behavior, did not fit the category of interpersonal influence situations. Overall, 23 studies were included in the meta-analysis by Arkin, Cooper, and Kolditz: 8 as teacher-student studies; 7 in the client-therapist paradigm; and 8 in the miscellaneous group. (One publication had two studies.)

Arkin, Cooper, and Kolditz performed meta-analyses, both of their own sample of 23 studies and of Zuckerman's smaller sample. Restricting ourselves for the moment to the smaller sample, Arkin, Cooper, and Kolditz agree with Zuckerman's overall conclusion: the 13 studies show little evidence for a main effect statement that attributions were self-serving. Moreover, scrutiny of Table 1 of Arkin, Cooper, and Kolditz leads us to the conclusion that any qualitative reviewer would come to the same conclusion, irrespective of whether statistical significance, the direction of findings, or effect sizes were used as the classification criterion. Arkin, Cooper, and Kolditz are correct in stating that Zuckerman ignored effect sizes and the direction of nonsignificant findings by relying on a box count of statistically significant effects. However, *in this particular instance* (though not in general) the additional information to be gained from effect sizes is somewhat marginal.

Arkin, Cooper, and Kolditz's meta-analysis of their own larger sample revealed a highly reliable main effect conclusion that attributions did show the self-serving bias. Zuckerman was, of course, not able to perform a qualitative review of this larger sample. However, when one performs a qualitative box count, one comes to the same conclusion as Arkin, Cooper, and Kolditz, although with possibly less confidence. Consider the 23 studies in their Table 1.

Eleven of these show statistically significant effects in the direction of the hypothesis, while only one would be expected by chance at the .05 level of significance. Thus, even a box count with statistical significance as the criterion shows self-serving attributions beyond a chance level in the larger set of studies. Using the direction of findings as a criterion, the same conclusion is indicated. Finally, scrutiny of the frequently large effect sizes increases our confidence in the conclusion that self-serving bias "exists" in such studies.

The above reflections suggest that Zuckerman and Arkin, Cooper, and Kolditz reached different conclusions because of differences in sampling and not because of any fundamental differences in review methods. The four studies that were reported after Zuckerman's review contribute three large and reliable findings in the positive direction and one finding reported as zero. Assuming no constant methodological bias, these buttress the conclusion that self-serving attributions were prevalent in these studies. The four studies that were available to Zuckerman, but which were overlooked, contribute two large positive findings and two findings listed as zero. Finally, the two helping studies excluded by Zuckerman add the largest and most reliable positive findings to the subcategory of miscellaneous studies that previously contained few positive findings of any reasonable magnitude. Note that the issue involved here is not meta-analysis *versus* qualitative reviews. Although meta-analysts typically prefer broader definition of constructs and larger samples, the more crucial issues concern later publication dates, more rigorous literature searches, and personal judgments about the theoretical relevance of individual studies.

In his review of the meta-analysis of Arkin, Cooper, and Kolditz, Zuckerman (Note 2) questions the relevance of including the helping studies by Stephan (1975) and Wells et al. (1977). He would probably also question the inclusion of an additional helping study by Arkin, Appelman, and Burger (1980). In addition, Zuckerman objected to the exact probability values that were reported for two studies, on the grounds that inappropriate groups were compared. These studies were Johnson, Feigenbaum, and Weiby (1964) and Schopler and Layton (1972).[5] With small samples of studies, it is difficult to resolve the issue of theoretical relevance. The typical meta-analytic strategy would be to "let the data decide." However, in this instance if one eliminated all the studies

5. Arkin, Cooper, and Kolditz appear to have complied with Zuckerman's suggested changes for three additional studies: Beckman (1970), Beckman (1973), and Mynatt and Sherman (1975), for which zero effect sizes, or simple blanks, are given.

in dispute, one would be reduced to nearly the same number of uncontested relevant studies that Zuckerman used, and the even smaller sample of contested studies would have two in which the appropriate statistics were not reported. Hence, the meta-analytic strategy of disaggregating studies is well-nigh impossible in this instance because of the small sample. It is not our task to rule on the appropriateness of the sample of Arkin, Cooper, and Kolditz. The arguments for and against including particular studies both have merit. What is needed is a greater number of studies using methods and constructs about which the contestants agree.

Although Zuckerman presented three categories of studies, he did not report conclusions for each one separately. Arkin, Cooper, and Kolditz, on the other hand, performed meta-analyses of each category, both for their own sample and for that of Zuckerman. For both samples, they concluded that the teacher-student studies showed inconsistent results, and the client-therapist studies showed reliable self-serving effects. The miscellaneous category resulted in a self-serving bias for the larger sample and in a trend in the same direction for Zuckerman's sample.

These meta-analyses of categories touch on the analysis of very small samples, Zuckerman's being five, four, and four, respectively. Consider the client-therapist studies, where by the flawed statistical significance criterion two of the studies in Zuckerman's sample suggested the effect, one showed no effects, and the fourth was statistically significant but in the opposite direction. Arkin, Cooper, and Kolditz derived an effect size of .56 for this category, favoring self-serving attributions. However, with such a small sample of heterogeneous studies the estimate of the average effect size is unstable, and it would be difficult to judge whether to treat the one contradictory study as a provisional "outlier" or as a clue indicating that an interaction theory is needed. Adding to this category the studies that Arkin, Cooper, and Kolditz found helped sway the judgment towards the first option. But it is easy to see how the results of the additional studies could have been in the opposite direction—in which case it would have been more meaningful to think in terms of interactions than of magnitude estimates for a main effect.

The treatment of interactions in these two reviews relates to their purposes. Zuckerman's major point, in the section of his review dealing with interpersonal influence studies, was not to determine whether self-serving attributions "existed," although he did conclude that evidence in favor was weak and depended on a few studies that could not be easily interpreted. Rather, Zuckerman speculated about three theoretical forces that might countervail

against such attributions and prevent them from being made in the interpersonal influence situation. The clear implication is that Zuckerman would expect self-serving bias in the interpersonal influence situation at times when these countervailing forces were not operating. Arkin, Cooper, and Kolditz address a somewhat different question in their attempt to determine the "existence" of self-serving attributions, focusing on the main effect conclusion. They do consider interactions as qualifications of the main effect, and suggest that they be noted and if possible, explained. However, their own meta-analysis of interactions fails to do this. It is simply an aggregation of the heterogeneous set of interactions that past investigators happened to have reported. There is no guiding theory, and with the small sample of studies there is little hope of being able to examine *stable* interactions so as to detect data-based puzzles and begin to explain them.

As both Zuckerman and Arkin, Cooper, and Kolditz point out, the conflicting findings in the teacher-student category point to a data-based puzzle that needs resolution. Moreover, comparison of results from the interpersonal influence and other paradigms suggests that self-serving biases may be less prevalent in the interpersonal influence situation. This, too, is a possible puzzle worth speculating about. Such puzzles ultimately increase the predictive power and accuracy of theories, and are a boon to theoretical advancement. Arkin, Cooper, and Kolditz were somewhat less concerned than Zuckerman with summarizing in order to detect puzzles and speculate on novel theoretical attempts to solve the puzzles. Their principal aim was to summarize the evidence for the main effect proposition. What we have, then, is a difference in priorities about two types of questions, each of which has value. Science needs to know its stubborn, dependable, general "facts," and it also needs data-based, contingent puzzles that push ahead theory. Our impression is that meta-analysts stress the former over the latter, and that many qualitative reviewers stress the latter more than the former (or at least more than meta-analysts do). Of course, neither the meta-analyst nor the qualitative reviewer *needs* to make either prioritization. Each can do both; and any one reviewer can consciously use both qualitative and quantitative techniques in the same review. Indeed, s/he should.

Conclusions

Meta-analysis is a set of useful techniques for literature review, especially when the sample of studies for review is large. The smaller the sample, the more limitations are placed on qualitative and quantitative reviews alike. While meta-analysis requires fewer

judgments from reviewers, and in this sense may be more "objective," it is still replete with judgments about the definition of the area of investigation, the relevance of methodological and substantive characteristics of studies, and the appropriate meta-analytic tools to be used. Meta-analysis is not a single technique, but rather a flexible set of techniques that can be adapted to the question at hand, provided that enough information is provided in the reports of research.

It has made a major contribution by highlighting poor practices in qualitative reviews, and should therefore serve as a stimulus to improving qualitative reviews. It may also make a contribution through establishing the "facts"—dependable, stubborn relationships between concepts. This is useful since much psychological theory has been built on relationships that are not dependable. Finally, it can offer clues to the explanation of findings, within the limits of the studies available to it. These limits may be severe if new explanatory constructs are needed for which proxies were not measured in past research.

The major limitation of meta-analysis relates to its major advantage. With large samples of studies, meta-analysis is more convenient than qualitative reviewing, and one can explore more contingency relationships. However, the larger set of studies can lead to an unwarranted psychological sense of security if there is a consistent replication of a relationship. To accept the estimate of the relationship one has also to assume that the estimate is unbiased or that subsidiary analyses have been conducted which have *validly* shown that no bias of importance resulted from the principal forces from which bias would be expected.

We have outlined several dangers inherent in poor practice of meta-analysis. Our caution is not directed at sophisticated users of the technique, to whom little of what we have said may be novel. Rather, it is directed against the potential mindless use of meta-analysis if it becomes a fad. Like all social science methods, its use involves assumptions that need to be made explicit, and if these assumptions are not clear to the user, misleading results can easily occur. Finally, if our discussion has made the best qualitative reviews seem similar to the best quantitative reviews, it is because they are similar. As we have tried to show, the weaknesses of traditional reviews are not inherent, and most of the strengths of meta-analysis can be applied to qualitative reviews to improve them.

Reference Notes

1. Howard, K. Personal communication, 1980.
2. Zuckerman, M. Personal communication, 1980.

References

Arkin, R. M., Appelman, A. J., & Burger, J. M. Social anxiety, self-presentation, and the self-serving bias in causal attribution. *Journal of Personality and Social Psychology*, 1980, in press.

Arkin, R., Cooper, H., & Kolditz, T. A statistical review of the literature concerning the self-serving bias in interpersonal influence situations. *Journal of Personality*, 1980, **48**, 435–448.

Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.

Beckman, L. Effects of students' performance on teachers' and observers' attributions of causality. *Journal of Educational Psychology*, 1970, **61**, 76–82.

Beckman, L. Teachers' and observers' perceptions of causality for a child's performance. *Journal of Educational Psychology*, 1973, **65**, 198–204.

Bergin, A. E. The evaluation of therapeutic outcomes. In A. E. Bergin and S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: John Wiley & Sons, 1971.

Block, J. Advancing the psychology of personality: Paradigmatic shift or improving the quality of research? In D. Endler and N. S. Magnusson (Eds.), *Interaction psychology and personality*. New York: John Wiley & Sons, 1976.

Block, J. H. Issues, problems and pitfalls in assessing sex differences: A critical review of *The psychology of sex differences*. *Merrill-Palmer Quarterly*, 1976, **22**, 283–308.

Boruch, R. F. On common contentions about randomized field experiments. In R. F. Boruch and H. W. Riecken (Eds.), *Experimental tests of public policy*. Boulder: Westview, 1977.

Calder, B. J., Ross, M., & Insko, C. A. Attitude attribution: Effects of incentives, choice, and consequences. *Journal of Personality and Social Psychology*, 1973, **25**, 84–99.

Campbell, D. T. The indirect assessment of social attitudes. *Psychological Bulletin*, 1950, **47**, 15–38.

Campbell, D. T., & Erlebacher, A. E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: A national debate*, Vol. 3, *Disadvantaged child*. New York: Brunner/Mazel, 1970.

Cartwright, D. Risk taking by individuals and groups: An assessment of research employing choice dilemmas. *Journal of Personality and Social Psychology*, 1971, **20**, 361–378.

Cohen, J. W. The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 1962, **65**, 145–153.

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

Collins, B. E., & Hoyt, M. F. Personal responsibility-for-consequences: An integration and extension of the "forced compliance" literature. *Journal of Experimental Social Psychology*, 1972, **8**, 448–593.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.

Cook, T. D., & Gruder, C. L. Meta-evaluation research. *Evaluation Quarterly*, 1978, **2**, 5–51.

Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin*, 1979, **86**, 662–679.

Cooper, H. M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 1979, **37**, 131–146.

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

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, **52**, 281–302.

Cronbach, L. J., & Snow R. E. *Attitudes and instructional methods*. New York: Irvington Publishers, 1976.
Director, S. M. Underadjustment bias in the evaluation of manpower training. *Evaluation Quarterly*, 1979, **3**, 190–218.
Educational Research Service. *Class size: A summary of research*. Arlington, Va.: Author, 1978.
Eysenck, H. The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, 1952, **16**, 319–324.
Fiedler, F. E. A contingency model of leadership effectiveness. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. New York: Academic Press, 1964.
Gergen, K. J. Social psychology as history. *Journal of Personality and Social Psychology*, 1973, **26**, 309–320.
Gilbert, J. P., McPeek, B., & Mosteller, P. Statistics and ethics in surgery and anesthesia. *Science*, 1977, **198**, 684–689.
Gillig, P. M., & Greenwald, A. G. Is it time to lay the sleeper effect to rest? *Journal of Personality and Social Psychology*, 1974, **29**, 132–139.
Glass, G. V. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 1976, **5**, 3–8.
Glass, G. V. Commentary on "Interpersonal expectancy effects: The first 345 studies" *The Brain and Behavioral Sciences*, 1978, **3**, 394–395. (a)
Glass, G. V. Integrating findings: The meta-analysis of research. *Review of Research in Education*, 1978, **5**, 351–379. (b)
Glass, G. V. Reply to Mansfield and Busse. *Educational Researcher*, 1978, **7**, 3. (c)
Glass, G. V., & Smith, M. L. Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, 1979, **1**, 2–16.
Jackson, G. B. *Methods for reviewing and integrating research in the social sciences* (DIS 76-20398). Washington, D.C.: George Washington University, Social Research Group, May 1978 (NTIS No. PB-283 747).
Jencks, C. S., et al. *Inequality: A reassessment of the effects of family and schooling in America*. New York: Basic Books, 1972.
Johnson, T. J., Feigenbaum, R., & Weiby, M. Some determinants and consequences of the teacher's perception of causation. *Journal of Educational Psychology*, 1964, **55**, 237–246.
Light, R. J. Capitalizing on variation: How conflicting research findings can be helpful for policy. *Educational Researcher*, Oct. 1979, **8**, 3–11.
Light, R. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 1971, **41**, 429–471.
Luborsky, L., Chandler, M., Auerbach, A. H., Cohen, J., & Bachrach, H. M. Factors influencing the outcome of psychotherapy: A review of quantitative research. *Psychological Bulletin*, 1971, **75**, 145–185.
Luborsky, L., Singer, B., & Luborsky, L. Comparative studies of psychotherapies. *Archive of General Psychiatry*, 1975, **32**, 995–1008.
Lykken, D. T. Statistical significance in psychological research. *Psychological Bulletin*, 1968, **70**, 151–159.
Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Stanford, Calif.: Stanford University Press, 1974.
MacCorqudale, K., & Meehl, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 1948, **55**, 95–107.
Mischel, W. *Personality and assessment*. New York: John Wiley, 1968.
Mischel, W. The interaction of person and situation. In D. Endler and N. S. Magnusson (Eds.), *Interaction psychology and personality*. New York: John Wiley, 1976.
Mynatt, C., & Sherman, S. J. Responsibility attribution in groups and individuals: A direct test of the diffusion of responsibility hypothesis. *Journal of Personality and Social Psychology*, 1975, **32**, 1111–1118.
Mosteller, R. L., & Bush, R. R. Selected quantitative techniques. In G. Lindzey

(Ed.), *Handbook of social psychology,* Vol. 1. Cambridge: Addison-Wesley, 1954.

Rachman, S. *The effects of psychotherapy.* Oxford: Pergamon Press, 1971.

Rosenthal, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal and R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press, 1969.

Rosenthal, R. *Experimenter effects in behavioral research,* Enlarged Edition. New York: Irvington Press, 1976.

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

Rosenthal, R., & Rubin, B. D. Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences,* 1978, **3**, 377–386.

Ryan, D. W., & Greenfield, T. B. *Clarifying the class size question: Evaluation and synthesis of studies related to the effects of class size, pupil-adult, and pupil-teacher ratios.* Toronto, Ont.: Ontario Ministry of Education, 1976.

Schopler, J., & Layton, B. Determinants of the self-attribution of having influenced another person. *Journal of Personality and Social Psychology,* 1972, **22**, 326–332.

Smith, M. L. Sex bias in counselling and psychotherapy. *Psychological Bulletin,* 1980, **87**, 392–407.

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

Smith, M. L., & Glass, G. V. *Class-size and its relationship to attitude and instruction.* San Francisco: Far West Laboratory, 1979.

Stegmuller, W. *The structure and dynamics of theories.* New York: Springer-Verlag, 1978.

Stephan, W. G. Actor vs. observer: Attributions to behavior with positive or negative outcomes and empathy for the other role. *Journal of Experimental Social Psychology,* 1975, **11**, 205–214.

Stouffer, S. A. *The American soldier: Vol. 1. Adjustment during army life.* Princeton: Princeton University Press, 1949.

Tukey, J. W. Analyzing data: Sanctification or detective work? *American Psychologist,* 1969, **24**, 83–91.

Wells, G. L., Petty, R. E., Harkins, S. G., Kagehiro, D., & Harvey, J. H. Anticipated discussion of interpretation eliminates actor-observer differences in the attribution of causality. *Sociometry,* 1977, **40**, 247–253.

Wicker, A. W. Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues,* 1969, **25**, 41–78.

Zajonc, R. B. Social facilitation. *Science,* 1965, **149**, 269–274.

Zuckerman, M. Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality,* 1979, **47**, 245–287.

Manuscript received March 15, 1980.