Clinical Psychology Review, Vol. 7, pp. 695-714, 1987
Printed in the USA.

0272-7358/87 $3.00 + .00

RESEARCH SYNTHESIS REVIEWS: AN ILLUSTRATED CRITIQUE OF "HIDDEN" JUDGMENTS, CHOICES, AND COMPROMISES

Paula S. Nurius

School of Social Work
University of Washington

William H. Yeaton

Institute for Social Research
University of Michigan

ABSTRACT. *This paper takes a pragmatic view of the steps involved in conducting a quantitative literature review. It emphasizes the multitude of judgments, choices, and compromises commonly encountered. Examination of a meta-analysis study of implosion therapy outcome research is provided as a structured means of illustrating and constructively considering questions, decisions, and issues likely to arise at each stage of the review process. The discussion should prove useful to prospective meta-analysis researchers and consumers. An extensive set of references that characterize contemporary meta-analysis research is also included.*

Over the past decade and particularly within the past 5 years, remarkable advances have been made in the research methodology underlying integrative reviews of the empirical literature. Special sections and complete issues in professional journals have been dedicated to the topic (Light, 1983; Yeaton & Wortman, 1984; *Journal of Consulting and Clinical Psychology*, 1983) and book length treatments of the topic have become increasingly available (Cooper, 1984; Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Light & Pillemer, 1984; Rosenthal, 1984; Smith, Glass, & Miller, 1980; Wolf, 1986).

The authors would like to thank Steven Schinke and two anonymous reviewers for their helpful comments and suggestions on an earlier version of this paper.

Various labels have been applied to these emerging methods for synthesizing and analyzing information contained within the existing research literature. Among these are data synthesis (Pillemer & Light, 1980), evaluation synthesis (U.S. General Accounting Office, 1983), research integration (Walberg & Haertel, 1980), quantitative assessment of research domains (Rosenthal, 1980), meta-analysis (Glass, 1976; Glass et al., 1981), and meta-evaluation (Cook & Gruder, 1978). All of these labels refer to formalized, systematic strategies for quantitatively combining outcomes of various sorts (proportions, correlations, means) *across* independent empirical studies and, when possible, applying statistical analysis techniques to the aggregated results.

Early discussions of meta-analysis were soon followed by articles applying and reporting use of research synthesis methods in integrative literature reviews. As an illustration, one major review journal, *Psychological Bulletin*, was examined over the past 10 years to determine the extent to which formal quantitative research synthesis methods were being applied in review articles. This examination indicated that several writers have simply compared study results using box count and voting methods. However, a steadily increasing incidence of literature reviews have incorporated formal research synthesis procedures shortly after introduction of these methods in the latter half of the 1970s.

The evolution of research synthesis appears to be following a pattern similar to that observed among many innovative medical technologies. In this pattern of development, a new technology is introduced into the research literature and rapidly becomes the focus of considerable attention, often generating unbridled enthusiasm. Soon, an enormous influx of articles appears in the literature that report initial experiences with the technology, leading to widespread dissemination and endorsement of the technology. Unfortunately, since these initial studies are typically observational and not based on controlled research design, their conclusions are often invalid. Wortman (1981) has documented this sequence of events among health care innovations such as gastric freezing of duodenal ulcers, electronic fetal monitoring, and coronary bypass surgery, and emphasizes diminishing favorability of outcomes during this sequence as well as shifts in investigator enthusiasm.

Two important sets of activities typically follow the initial pattern. One involves stepping back and taking a second look at the technology and better defining its problems, gaps, weaknesses, and limitations, as well as its strengths and potential benefits. The second involves modification of the technology's original form to remedy its deficits and to clarify its range of applicability. This is the stage of development where research synthesis methods now stand. While there seems little doubt as to its potential value, current concern lies more with practical questions regarding the appropriate avenues of application.

The primary purpose of this paper is to highlight many of the judgments, choices, and compromises germane to the emerging technology, especially those dilemmas particularly evident to reviewers of the applied social and behavioral science literature. The potential value of meta-analytic techniques in rendering relatively subtle and "hidden" judgments endemic to any review more explicit and thus more "mindful" and accountable is also explored. General points will be illustrated through examination of the steps taken in one research synthesis study, a meta-analysis of implosion therapy outcome research. This example is intended

to provide a structured means of considering the questions, decisions, and issues that are likely to occur at each stage of the review process. A secondary purpose of this paper is to provide pertinent references for the interested reader who wishes to conduct a research synthesis and seeks detailed discussion of one or more steps in the process.

POTENTIAL BENEFITS

The focus of this paper will be on the potential problems that result from the many arbitrary decision points in a synthesis review. To ensure a balanced viewpoint in this paper, several major benefits of formal synthesis techniques should be noted. Two benefits of particular value to the applied scientist are: (a) the increase in statistical power afforded by pooling comparable data from several different studies; and (b) the increase in generalizability since description, evaluation, and inference are based on a *body* of research as opposed to a single study.

Also of benefit is the extent to which research synthesis methodologies provide explicit, systematic, and, thus, more readily replicable procedures for conducting an integrative review of the literature. That is, while interpretations of individuals may vary, reviewers who employ formal quantitative synthesis procedures are more likely to achieve comparable outcome results and conclusions than those who rely solely on more traditional narrative-discursive approaches (Glass et al., 1981; Light & Pillemer, 1984).

In addition to increasing the reliability of integrative review *outcomes*, quantitative synthesis methods systematize and make more explicit the review *process* as well. The general steps involved in this process include (Jackson, 1980):

1. selecting the questions or hypotheses for review;
2. sampling the research studies that are to be reviewed;
3. representing the characteristics of the studies and their findings;
4. analyzing the findings;
5. interpreting the results; and
6. reporting the review.

This framework will be used to discuss the process of conducting a quantitative synthesis study. Within this framework the authors' research synthesis of implosion therapy (IT) will be used to illustrate specific activities and decisions involved in quantitative approaches to integrative literature reviews.

Implosion therapy research is itself a good example of the evolutionary pattern of a new technology. That is, it has a relatively clear beginning point in the literature, made a notable impact, spread quickly, underwent a variety of revisions in its treatment protocol, entered a more critical phase of research wherein mixed and negative outcomes were reported and caveats and limitations were identified, and has been expanded and adapted to accommodate varying needs, constraints, and circumstances.

SELECTING THE QUESTIONS OR HYPOTHESES FOR REVIEW

The task of establishing the central questions or hypotheses upon which a review will be based is of fundamental importance. This is the point at which key theoretical and substantive dimensions are carefully considered, and likely link-

ages between review findings and policy, practice, or future research are proposed. Decisions made at this step will serve to structure the review and to act as a conceptual guide in the multitude of choice points inevitably encountered. While the review process has clearly become more formalized and systematized, it is still replete with judgment calls (Leviton & Cook, 1981; Light, 1980). Thus, it is critical that any theoretical filters be made explicit since they will account for much of the variation between individuals in the focus, course, and interpretation of the literature synthesis.

The extent to which one can specify key hypotheses in advance will influence considerably the credibility and interpretability of statistical analyses (see Hunter, Schmidt, & Jackson, 1982, for examples). This, of course, is also true for primary research. However, the risk of capitalization on chance and of undercorrection for artifacts warrants additional attention in data synthesis research. If the review is primarily exploratory, the reviewer may opt to summarize research characteristics and to compile descriptive information that does not require inferential statistical tests. That is, some investigators may be interested primarily in the descriptive state of a particular body of research (e.g., the relative frequency of research designs, measures, client/subject groups, sample sizes, and so on). Data synthesis techniques can be used to develop an actuarial base on key variables of interest.

In addition to determining the most central and important questions, the reviewer must also decide which questions are most "answerable." In many instances, research syntheses will allow questions to be addressed that are not answerable within single studies (Light, 1984). What the quantitative reviewer must be keenly aware of is how previous questions have been asked *and* how answers have been reported. What generalizations can one make in the context of incomplete outcome reporting? For example, our original motivating question was concerned with the level of effectiveness of implosion therapy compared to other treatments, yet the literature revealed that many studies had been conducted using no-treatment control groups. In instances such as this a reviewer would need to rethink ways in which the review question could be reformulated to accommodate the limits and strengths of the existing data base. Logistically, if one is not already familiar with these aspects of the target literature, a representative sample would need to be coded and preliminary results noted.

IT Study Illustration

To avoid formulating questions and hypotheses about implosion therapy that proved unanswerable, small subsets of clinical and analogue research studies were sampled at three time points to get a more complete sense of the typical study questions, designs, measures, and outcomes reported. A point was also made of reviewing early theoretical papers which described in detail both the procedures and protocol of the original treatment paradigm as well as the treatment's theoretical underpinnings (e.g., Levis & Hare, 1977; Marshall, Gauthier, & Gordon, 1979; Stampfl & Levis, 1967, 1973).

Through this preliminary examination, it became evident that the manner in which the treatment was implemented changed over time as did the study samples and the types of problems presented for treatment (see Yeaton & Wortman, 1984). The resulting plan, therefore, was to track treatment and study attributes and outcomes over time and to examine differences in average treatment effect size as

a function of differences in methodological, treatment, and client/subject attributes (e.g., design, type of sample, problem type, treatment paradigm). This agenda was in keeping with the authors' concern about treatment strength and integrity issues (cf. Sechrest & Yeaton, 1981; Yeaton & Sechrest, 1981) as well as their interests in research methodology.

SAMPLING THE RESEARCH STUDIES

Sampling issues in reviews generally fall into categories regarding representativeness, bias, relevance, and appropriateness. With respect to representativeness, the potential consequences of drawing upon a sample of the available literature as opposed to reviewing the entire population of reports must be considered when the sample of studies is large. The problem of generalizability will be critical when a sampling approach is chosen, and various forms of bias are likely to be introduced. Bias attributable to source of the report (journal, book, dissertation or thesis, paper presentation, unpublished work) has been documented (Glass et al., 1981; Johnson, Maruyama, & Johnson, 1982; Light, 1983; Smith, 1980; White, 1982). Additionally, bias may be associated with time (year or year span) of reporting (Baum et al., 1981; Smith & Glass, 1980; Wortman, 1981) and with study design and quality (Chalmers, 1982; Dersimonian & Laird, 1983; Glass & Smith, 1979; Wortman & Yeaton, 1983). Judgments regarding lack of revelance will be based largely on the degree to which treatments and outcome measures reflect the hypotheses being tested or the exploratory questions being pursued. Judgments regarding appropriateness will dictate specific inclusion and exclusion criteria and influence the methodological quality of the studies included in the review.

Decisions pertaining to search methods will naturally be closely allied with those pertaining to sampling methods. Strong argument has been made for a comprehensive, multi-method search approach (Glass et al., 1981; Light & Pillemer, 1984; Rothman, 1980). Thus, use would be made not only of journal abstract and review indexes and of computerized search and retrieval systems but also of prior review articles, dissertations, governmental and unpublished reports, and correspondence with experts in the field.

IT Study Illustration

There is little doubt concerning the benefits of employing thorough search procedures to capture virtually the entire population of existing empirical reports on the subject in question. Yet, there can equally be little doubt as to the practical limitations of time, resources, and accessibility, particularly for those reviews being conducted in nonresearch settings (e.g., schools, human service agencies, health care services). Similarly, one may eventually reach a point of diminishing returns wherein the cost involved in retrieving a small proportion of the total number of studies exceeds their expected value and their likelihood of significantly influencing the quantitative outcomes (see Light, 1980, for an overview of selection issues).

Such constraints, however, are the bane of all types of research and are certainly not unique to research synthesis. What is needed are reliable indicators of the robustness of various search and sampling methods. That is, to what extent can optimal procedures be abbreviated without significant risk of substantial error?

Rosenthal (1979) has developed a simple calculation that enables one to determine how many null results would have to exist in order for the combined probability of the entire set, retrieved and unretrieved, to exceed a specified probability level. Orwin (1983) has provided a similar statistic for use with meta-analysis results.

The strategy chosen for the IT study was to gather a comprehensive set of published studies reported in English that met a priori selection and exclusion criteria. Exceptions included dissertations, government documents (of which none were uncovered in the search), and project reports not contained within the professional literature. Search methods employed were: (a) examination of an existing review (Levis & Hare, 1977) of all IT research from its introduction in 1967 through 1975; (b) an exhaustive review of *Psychological Abstracts* from 1967 through 1981; and (c) a computerized search of the literature from 1957 through August 1982.

To be included in the meta-analysis, the following criteria had to be met: (a) use of two or more treatment groups, one of which utilized implosion therapy; (b) sample size of at least 10; 5 per group being viewed as the smallest allowable sample; (c) use of a quasi-experimental or true experimental design; (d) group posttest means and standard deviations were reported; and (e) no use of drugs as part of treatment. This latter criterion was initially tentative. It was unclear at the onset how large a role drugs would play in IT research. Consequently, those studies which reported use of some type of medication were collected and then excluded after examination indicated considerable variability in type of drug and in their manner of administration. This decision was consistent with our intent to study the effects of IT unconfounded by supplemental interventions.

Several noteworthy decisions were required at this stage, some that were later revised. The choice to gather a reasonably comprehensive sample was facilitated by the relatively short time span during which IT research had been conducted and the specific nature of the treatment itself. The scope of the present study was modest compared to such general treatments and enormous sample sizes as Miller's (1977) study of the effects of drug therapy on psychological disorders (N = 2,963) or even Smith and Glass' (1977) study of the effectiveness of psychotherapy (500 studies were selected for inclusion in the study, and 375 were fully analyzed).

The decision not to include dissertations presents an unknown biasing factor. Effect sizes are frequently lower for dissertation research, which suggests the possibility of an inflated average effect size with their exclusion. However, several factors convinced the authors that the risk was within acceptable bounds and that the cost of retrieval was in excess of the probable benefits. In addition to the time and cost issues associated with obtaining full dissertation documents, these factors included the relatively small number of IT dissertations revealed by computer search (roughly 20% of located studies were dissertations) and the frequent finding of published articles being based on prior dissertation research.

A decision was made to include alternative treatment groups in addition to control (both no-treatment and placebo) groups. This decision reflects the authors' interest in making separate determinations of the average effect size of implosion therapy relative to other types of treatment as well as to the more conventional placebo and no-treatment conditions. This approach also allows one to gather descriptive data regarding the relative prevalence of control conditions and comparative treatments over time. Furthermore, in clinical settings, the use

of no-treatment and placebo groups is sometimes deemed less ethically supportable than is the use of alternative treatments. Thus, to avoid losing these valuable comparisons, the type of group means and standard deviations used in calculating effect sizes relative to implosion theory was expanded to include alternative treatment groups in addition to placebo and no-treatment control groups. The conventional formula for calculating effect size, $ES = (\bar{X}_t + \bar{X}_c)/SD_c$ (Cohen, 1977), was used. In words, this states: subtract either the control group mean or the comparative treatment group mean ($\bar{X}_c$) from the implosion therapy group mean ($\bar{X}_t$) and divide this difference by the standard deviation of the same control or comparative group (SD_c). Issues associated with different approaches to calculating effect sizes will be addressed in a later section.

The need for two additional revisions in a priori thinking also became evident during the search and sampling activities. One was the selection criterion that only those studies in which administration of implosion therapy appeared to be reasonably consonant with the original treatment paradigm would be included. It soon became evident that part of the evolution of implosion therapy research included periodic revision in the sequence and character of the treatment resulting in similarly labeled treatments with numerous dissimilarities. The authors' solution was to create an "implosion therapy derivative" category to encompass those treatments that did not conform to the original paradigm yet were conceptually related.

The second change in thinking involved the original intent to include only those studies reporting group posttest scores, thus allowing calculation of an effect size. A choice was later made to include all studies that met the remaining inclusion criteria noted earlier even if group posttest scores were not reported. This decision allowed us to compile information on all studies utilizing the types of design, groups, and sample size prespecified as appropriate even if outcome data were not available. By casting this larger net it was possible to provide descriptive data pertaining to clients, treatments, and studies using a far greater number of IT studies (cf., Light & Pillemer, 1984), but to address questions of effectiveness using only those studies for which effect sizes could be calculated.

REPRESENTING THE CHARACTERISTICS OF STUDIES AND THEIR FINDINGS

As was the case with each of the prior steps, the codification and classification process is an iterative one. The previewing of the literature that has taken place to this point should make it easier to design a coding form to record the information desired, to allow for variations in the form reflecting what is reported, and to represent information based on judgment or interpretation as reliably and usefully as possible. The initial time spent in carefully considering the substantive and methodological issues should later prove to be a valuable investment in minimizing the need to recode studies. One important consideration is whether the coded results will be entered into a computer for later analysis. If so, the coding form should be designed to facilitate entry directly from the form. Glass et al. (1981, pp. 223–237) provide a useful example of the coding form used for computer data entry in their psychotherapy meta-analysis.

A likely scenario with applied social science data is that there will be many studies from which one can glean some of the desired data, a moderate number from which one can retrieve most of the data, and only a few from which one can

obtain virtually all of the data. It is thereby important to maximize use of whatever information is available, to consider various levels of information which bear on the study questions, and to document decisions made as well as their rationales.

During this data collection process, the reviewer will be called upon to make a great many judgments and decisions. Social and behavioral research reports are notorious for their deficient reporting of procedural and methodological detail (Orwin & Cordray, 1983; Smith et al., 1980). In addition to developing explicit decision rules that make clear these judgments, the reviewer would be well advised to employ independent coders to assess the reliability as well as the clarity and completeness of information retrieval.

IT Study Illustration

As was noted previously, theoretical papers were used as guides in determining important variables regarding implosion therapy administration and evaluation. A sampling of the research reports at different points in time was then undertaken to assess how many of these variables tended to be reported and in what form. The coding form was revised three times in the process as some variables were dropped, others added, and finer distinctions established.

Coding of two treatment strength and integrity variables provides a useful example of needed revisions in coding procedures. Two assumptions central to implosion therapy are: (a) that a graduated sequence of anxiety eliciting cues/stimulus (an Avoidance Serial Cue Hierarchy) are employed in such a manner that, (b) the subsequent conditioned cue is not presented until there is no anxiety associated with the cue currently being presented. The latter assumption is predicated on the requirement that response extinction be accomplished at each cue level.

Coding for each of these variables was initially based upon simple statements of: (a) yes, the condition was met; (b) no, the condition was not met; and (c) it was not discernable whether the condition was met. Since most reports were ambiguous, the majority of responses were categorized as "not discernable" rendering the variables virtually useless as potential moderating variables of treatment effectiveness. A decision was made to recode the two variables, expanding the "yes" category to include not only completely unambiguous cases but also those in which the conditions appeared to have been met or in which the authors discussed the importance of the conditions. In retrospect, still finer gradations might have been made. Such coding schemes must, however, be quite explicitly defined to yield valid, reliable data.

This situation can be readily generalized to a host of key variables in applied social and behavioral research. And, in addition to being ambiguously reported, some variables of interest are simply not routinely reported. In this study clinical expertise of the individual administering the treatment was seldom included. Nor was descriptive information (e.g., race, gender, age) of these individuals routinely provided. While specific clinician/experimenter characteristics and client/subject characteristics were sought, they simply were not generally available. Probably the best one can do in such situations is to identify potentially important variables from theory and previous studies, to gather data on proxy variables, and to report the dearth of the more direct indicators. In the IT study, for example, categories to denote authors' professional affiliation (e.g., academic institution, clinical set-

ting) as well as their training credentials (e.g., certified clinician, doctoral candidate, doctorate holder) were included in addition to categories more explicitly representing the nature and extent of clinical experience. Parenthetically, variables identified as important may involve commonly raised factors such as demographic characteristics or may involve even more elusive factors such as those germane to ecological validity (cf. Berkowitz & Donnerstein, 1982).

Even when attributes and procedures are reported in sufficient detail, it may become apparent after coding has begun that more varied categories or more fine-grained distinctions may be necessary to ensure coding reliability. Various means of employing independent coders and of determining their interrater reliability have been described by several authors in an edited volume by Donald Hartman (1982). Directly relevant also are the guidelines by Orwin and Cordray (1983) and Rosenthal (1984) indicating that different types of indicators are needed for establishing interrater reliability and confidence judgments for different types of data. For categorical data, for example, a combination of Kappa (unweighted for nominal, weighted for ordinal level measures) plus percentage of agreement are generally regarded as appropriate (e.g., Cohen, 1960; Light, 1971). For continuous variables, agreement and Pearson's correlation coefficient (r) are more appropriate; including, perhaps, calculation of r in the form of phi (ϕ) for dichotomous variables (Hartmann, 1977).

ANALYZING THE FINDINGS

It is at the data analysis stage that quantitative approaches to integrative literature review will diverge most from qualitative approaches. The specific course of data analysis will be dictated both by the nature of the questions asked and by the availability of desired data. It is quite likely that a variety of research synthesis techniques will be necessary to accommodate differences between studies and to remediate deficits in data reporting. An overview of many of the currently available research synthesis techniques follows.

If the review question deals with the overall magnitude of effect or central tendency of outcomes from a collection of studies, one might consider: (a) calculating a standardized difference (d) statistic between treatment and either contrast or control group means (Glass, 1976; Glass et al., 1981; Hedges, 1984) using either the control group standard deviation or the pooled, within group standard deviation in the denominator (Miller & Berman, 1983); (b) examining the percentage of overlap between treatment and control or contrast group distributions (Cohen, 1977); (c) determining the proportion of variance accounted for in dependent variables (Hays, 1973); or (d) combining correlations, frequency counts, and percentages across studies (Baum et al., 1981; Hunter et al., 1982).

On the other hand, if one is more concerned with or has available only the probability levels across studies or if a single, overall significance test of intervention impact or association is desired, alternatives would include: (a) summing probability levels; (b) summing z scores and weighted z scores; (c) summing t scores; (d) summing logs, or testing the mean probability level and the mean z score (see Rosenthal, 1978, 1980, for more details). More fine-grain approaches could be pursued that would allow one to investigate interactions between study, intervention, or participant attributes and outcomes. One could also examine a variety of differences in outcome by subgroup: (a) by regressing outcomes on

selected predictor variables of interest (e.g., see Glass, 1977); or (b) by use of blocking techniques such as a one-way analysis of variance with, for example, studies as the blocking variable, or a two-way analysis of variance with a treatment by studies design (e.g., see Cochran & Cox, 1957; Rosenthal, 1978).

Finally, when the goal centers more around comparison of similarly labeled interventions, addressing questions of conceptual validity or of differences among derivations of a given intervention, one may consider use of: (a) cluster approaches wherein subgroups are clustered based on specific statistical or substantive criteria (e.g., homogeneity); or use of (b) statistical contrasts of subgroups using either tests of mean outcome differences or calculating correlations between variants of similarly labeled interventions and their outcomes (see Light & Smith, 1971; Pillemer & Light, 1980).

Considerable attention has been devoted to the limitations associated with the analysis stage of integrative reviews. Wortman (1983) and Bryant and Wortman (1984), for example, have used Campbell and Stanley's (1966) research paradigm to discuss the potential threats to construct validity, internal validity, external validity, and statistical conclusion validity. Critics have questioned several aspects of research synthesis including study selection and the quantification and statistical analysis of outcomes (e.g., Bandura, 1978; Crown, 1981; Eysenck, 1978; Kazrin, Durac, & Agteros, 1979; Presby, 1978; Rachman & Wilson, 1980; Wilson & Rachman, 1983). Others have acknowledged potential pitfalls but have offered suggestions for minimizing them (e.g., Cooper, 1979, 1981, 1982; Glass et al., 1981; Glass & Kliegl, 1983; Hunter et al., 1982; Light & Pillemer, 1984; Shapiro & Shapiro, 1982; Strube, 1981; Strube & Garcia, 1981; Strube & Hartmann, 1982, 1983). Increasingly, revisions and refinements of aggregation, summarization, univariate and multivariate analysis, and interpretation methods have become available (e.g., Cohen, 1977; Cooper, 1982; Glass et al., 1981; Hedges, 1982, 1984; Hedges & Olkin, 1980; Kulik & Kulik, 1982; Mitchell & Hartmann, 1981; Orwin & Cordray, 1983; Rosenthal, 1978, 1979, 1980, 1983; Rosenthal & Rubin, 1982a, 1982b, 1982c; Walberg, 1983).

It is not within the scope of this paper to attempt to provide a complete overview of all major and emerging meta-analysis issues and methods. Instead, references for more extensive reading have been provided for the interested reader. Discussion of those aspects of research synthesis that involve the greatest degree of subjectivity will be elaborated upon in the context of the IT study. These include judgments regarding data quality, inclusion criteria, assumptions of statistical techniques (e.g., normality, independence), conceptual comparability (e.g., of treatments, measures, samples, effect size, contrasts), and interpretability of aggregated outcomes.

IT Study Illustration

Selective reading of the implosion therapy studies indicated that a great deal of information would be lost if only those studies were included that reported group posttest means and standard deviations and adhered to the original implosion therapy protocol. Thus, implosion therapy treatments were coded if they appeared to be in accordance with the original protocol or represented a derivation of that protocol. Subsequent analyses and summaries of background characteristics were stratified by this treatment variation. Several lines of thinking prompted

this decision. For one, the calculation of effect size is predicated upon both the mean and the standard deviation of the contrast group. Not only might one expect differences in mean levels and in variance among these groups, but evidence suggests that other differences are also likely. Landman and Dawes (1982), for example, in their reanalysis of the now classic Smith and Glass (1977) study found an effect size of .56 when comparing treatment and no-treatment control groups and a notably lower effect size of .38 when comparing treatment with placebo control groups. Similarly, Yeaton and Sechrest (1981) argue that strength and integrity dimensions of treatment should be taken into account. In the IT study case case, variations of implosion therapy (i.e., derivations) were expected to sufficiently differ from the original version in strength and protocol to warrant separate consideration.

Stratification was also imposed according to the type of contrast. Six possible contrasts were allowed: (a) IT versus no treatment control; (b) IT versus placebo treatment control; (c) IT versus alternative forms of treatment; (d) IT versus an IT derivation; (e) IT derivations versus no treatment controls; and (f) IT derivations versus alternative forms of treatment. No IT derivation versus placebo treatment contrasts were reported.

This approach represents a compromise in that several important conceptual and methodological distinctions were respected whereas others were not (e.g., different types of treatment such as systematic desensitization, use of modeling/instructional techniques, "traditional" psychotherapy, and so on were combined). This decision was based on the authors' interest in fairly general substantive questions (e.g., Is IT superior to IT derivatives?) as opposed to highly specific ones (e.g., Is IT superior to systematic desensitization?) and on the practical constraints posed by making conclusions based on a very small number of studies represented within any one type of alternative treatment.

With respect to data quality, a decision was made to include all true experimental and quasi-experimental designs, to code for study design, and to examine the relationship of study quality, based on design, to background characteristics and to treatment outcomes. Treatment results were initially coded separately by the type of outcome measure reported. However, the authors later judged that stratification by this variable in addition to type of implosion therapy condition and type of meta-analytic contrast resulted in numbers of cases too small to justify its use. Consequently, a decision was made to combine different types of reported measures in the calculation of average effect sizes, rendering the study rather than the measure as the unit of analysis. This decision was also strengthened by the fact that the measures to be used in effect size calculation did meet a conceptual validity requirement imposed by the authors of measuring some aspect of anxiety associated with the target stimulus. While the decision reflects a practical concern regarding small sample size, it underscores the importance throughout each stage of integrative reviews of establishing and reporting all decision rules.

In those cases in which desired study outcomes are not reported, one may elect either to exclude those studies or to attempt to derive the desired effect size measure from diverse outcome statistics. Several pros and cons need to be considered here. As an illustration, we considered deriving effect size estimates from between-group ANOVA scores and pooled estimates of variance. However, this approach could be utilized in only one study. While Grass et al. (1981) strongly advocate the use of a wide spectrum of indirect methods to derive effect size

estimates from related statistics, concern has also been voiced that this strategy frequently runs the risk of severely violating underlying assumptions (e.g., of normality), rendering approximations considerably more controversial than those made from more direct methods (Cordray & Orwin, 1983; Jackson, 1980; Orwin & Cordray, 1983).

A final note concerns the use of computers in quantitative integrative reviews. The present authors found computerization of *study* level data to be enormously useful for summarization and analysis. Extensive difficulties were encountered, however, in devising methods for computer analysis of group by treatment level information within the same data matrix as the study level data. In essence, this constituted a mixing of units of analysis. To illustrate, one must picture the data file as a matrix with variables constituting the columns and cases constituting the rows. In any one matrix, the cases can represent only one unit of analysis—either individual studies *or* individual groups within each study. Representing data on both levels within the same data file proved to be extremely difficult using available computer software.

Accordingly, a decision was made to calculate, by hand calculator, the effect sizes for the various group contrasts and to enter these outcomes as study level data. Unique group level information such as length of session, stimulus exposure time, proportion and timing of dropouts, and gender composition was summarized, but no tests of the relationship between study outcomes and these group level variables were conducted. As an alternative, however, the interested reader is referred to Glass et al.'s (1981, pp. 233–237) computer based coding from and to McDaniel's (1983) Meta-Analysis Computer Program, a software package written in SAS macro language to aid in computing a variety of statistics associated with meta-analysis.

INTERPRETING RESULTS

The interpretation stage is one in which the complementary nature of qualitative and quantitative approaches to integrative review becomes particularly obvious. Guidelines for interpreting the relative significance of average effect sizes for social and behavioral science data have been offered (e.g., Cohen, 1977; Cooper, 1981, Gallo, 1978; Kendall & Norton-Ford, 1982; Rimland, 1979; Rosenthal & Rubin, 1982b; Sohn, 1980). Rosenthal and Rubin (1982a) have proposed the use of a binomial effect size display (BESD) given its ease of understanding and its ready translation to success rates, a metric particularly familiar to clinicians. Jacobson, Follette, and Revenstorf (1984) have suggested use of a "reliable change index" which produces a standardized score with a clear-cut decision criterion for improvement, and is recommended in conjunction with additional, substantive-based criteria of clinical significance of change. In the final analysis, interpretation of practical significance and importance of effects relies heavily on theory and on qualitative judgments of outcome magnitude by experts in each field for which syntheses are conducted (Sechrest & Yeaton, 1981).

Questions with respect to inference and to generalizability must also be considered. The lack of a priori hypotheses, of nonbiased sampling, and of assurance that the synthesized data reasonably conform to assumptions underlying statistical testing all serve to diminish the credibility of one's interpretations. Similarly, the extent to which a given set of synthesized outcomes and interpretations can be

generalized depends on the nature of the samples, settings, and interventions, and the extent to which there exists variation in outcomes within and between studies (Cook & Campbell, 1979).

Conflicts and variation in review outcomes are actually quite common (Jackson, 1980), which underscores the importance of qualifying and explaining or resolving such differences. Examples of qualitative information within a study include narrative characterizations of the intervention and the manner in which it was implemented, descriptions of potential mediating variables, and discussion of pertinent conceptual and substantive topics. Examples of valuable sources of qualitative information outside those studies within a quantitative review include: (a) case studies; (b) nonquantitative studies; (c) nonquantitative information in quantitative studies; (d) expert judgment; and (e) narrative reviews of collections of research studies (Light & Pillemer, 1984). As many have argued (e.g., Cook & Leviton, 1980; Light & Pillemer, 1982, 1984; Nurius, 1984), the richest and most reliable summarizations of "what we know" in a given area can best be achieved through an alliance between qualitative and quantitative information and methods of investigation. In short, emphasis should be on supplementing not supplanting one approach with the other.

IT Study Illustration

The intention of this section is not to fully report the findings of the IT quantitative review. Rather, illustrative issues with respect to interpretation of results encountered in this review are offered. One point raised earlier is the importance of taking into consideration the type of comparison being made. For example, substantial differences in effect size were observed among the different types of group comparisons in the IT study.

Notable differences were evident in comparisons of implosion therapy to no-treatment control groups, to placebo control groups, and to groups receiving some other form of treatment (e.g., systematic desensitization, traditional psychotherapy) for anxiety and phobia related problems. The average effect size contrasting IT to no-treatment controls was $-.35$, whereas the effect size relative to placebo controls was $-.16$. (In both cases, the negative sign indicates that, overall, IT intervention produced less effective outcomes than did either of the comparison groups.)

Averages, while generally informative, may be somewhat misleading. And herein lies an important potential advantage of quantitative synthesis procedures such as meta-analysis. That is, when results do not come out as expected, both the data and the review process are available for closer examination. Thus, quantitative synthesis first aids in detecting contradictions and then allows one to backtrack and to determine factors leading to the unexpected results.

An obvious first step when counterintuitive results emerge is to check one's calculations. Following this, closer examination of a variety of substantive or methodological causes may be indicated. For example, are sampling variability or the small number of studies likely sources of distortion? These two factors may have partially accounted for the unexpected IT study results. In the case of no-treatment controls, most individual effect sizes were very small, indicating almost no difference between the groups. Two relatively large negative effect sizes greatly influenced the mean. For comparisons with placebo groups, while both positive

and negative effect sizes were evident, their combination served to neutralize observed differences, resulting in a low average effect size.

By contrast, comparisons of IT to some other form of treatment resulted in an average effect size of .46. This finding is particularly surprising given the negative outcomes with no-treatment and placebo control groups. It is also curious in that the majority of studies comprising this comparison were based on true experimental designs. Therefore, this apparent contradiction is not consistent with the common finding of larger effect sizes associated with weaker quasi-experimental designs (Sacks, Chalmers, & Smith, 1982; Yeaton & Wortman, 1985).

What these differences do point out is the importance of differentiating types of contrasts with respect to types of groups as well as types of measures. What is also indicated is the need for further investigation. The range of outcomes in this case was considerable for each type of comparison, and the number of studies reporting outcomes that permitted aggregation was relatively small. Under these conditions, it is difficult to assess how representative the average effect size is. This ambiguity would, of course, also exist for the reviewer relying solely on qualitative, narrative methods. The benefit of quantitative methods is that they render the ambiguity more evident, requiring the reviewer to explicitly address them in some fashion.

At this point in the review, the author of the research synthesis is in the best position to qualify the quantitative results and, indeed, has a responsibility to do so. The reader needs to be alerted to notable biases, trends, or problems with the findings. For example, a temporal trend was observable in IT study treatments, subjects, and target problems. Early studies adhered more closely to the classic protocol for implosion therapy; later studies experimented with varying dimensions of the treatment procedure. Early studies had a greater proportion of subjects in treatment as opposed to educational settings and a higher proportion of more clearly apparent clinically significant problems to which the treatment was being applied.

REPORTING THE RESULTS

As with primary research, it is important to provide sufficient information about one's review to enable critical examination of the evidence as well as to facilitate replication of the review. Jackson (1980) has emphasized the importance of carefully reporting literature search and selection procedures accompanied by a rationale for one's decisions. Cooper (1982) has elaborated this point in speaking of potential threats to validity common to synthesis studies. And, as the preceding sections have illustrated, numerous junctures are encountered where judgments and compromises must be made. It is the reviewer's responsibility to be explicit in reporting these.

In an effort to establish general guidelines regarding information important in clarifying the conclusions of a review, Light and Pillemer (1984) suggest the following checklist of questions:

1. What is the precise purpose of the review?
2. How were studies selected?
3. Is there publication bias?
4. Are treatments similar enough to meaningfully combine their outcomes?
5. Are control groups similar enough to meaningfully combine their results?

6. What is the distribution of study outcomes?
7. Are outcomes related to research design?
8. Are outcomes related to characteristics of programs, participants, and settings?
9. Is the unit of analysis similar across studies?
10. What are the recommended guidelines for future research, practice, and policy?

IT Study Illustration

It soon became apparent in reviewing the implosion therapy literature that different interests and uses of results by different audiences (e.g., practitioners and researchers) corresponded with differences in both interpreting and reporting results. In many cases, of course, the investigator is both a researcher and a clinician. However, implosion therapy is more amenable to application by nonpractitioner researchers in nonclinical analogue studies than are some other forms of treatment. For example, numerous IT studies were conducted in educational settings with individuals whose aversion to a given stimulus was relatively low in severity. A sampling of some of the general differences in reporting according to the orientation of the author and audience are noted below.

Regarding the type of data reported, researchers are likely to be interested in descriptive data such as types of designs, measures, and sampling techniques as well as the relation of these factors to aggregated outcomes. Also of likely interest will be evidence of interactions (e.g., different results for different measures) and the plausibility of systematic bias (e.g., nonblindness with respect to subjective measures). Practitioners as a group will more likely be interested in the effectiveness of IT for specific target populations or with specific types of problems. Results that speak to differences in treatment effects attributable to specific IT components, parameters, and combinations with other treatment factors (e.g., medication) are also likely to be of particular interest.

Validity emphasis is another dimension of interpretation to consider. Using Cook and Campbell's (1979) typology, one would predict researchers to place greater emphasis on interval validity (i.e., were randomized clinical trials employed) as well as on statistical conclusion validity (i.e., is there reasonable evidence from which to infer covariation between the presumed cause and effect). Practitioners, on the other hand, are likely to be more concerned with external validity (i.e., to what extent do results generalize to other client or problem types) and with construct validity (i.e., what makes IT work).

Finally, the two groups will be inclined to approach the review with somewhat different goals. By and large, the practitioner is primarily interested in improving his or her practice or in ascertaining whether the use of IT is empirically supported. The researcher, while obviously concerned with the empirical support (or lack of) for IT, is also seeking guidance in targeting future research.

These differences may have produced uneven reporting along several dimensions. The following generalizations illustrate some of these and the difficulties these deficits pose for the reviewer. In a great many cases, statistics which lend themselves to quantitative synthesis (e.g., means, standard deviations, correlations, proportions, pooled F scores) were not reported. The reviewer must then

either exclude the study or, if supportable, work to derive useful results from those statistics which are reported.

Another pervasive limitation was the extent to which treatment procedures were described. It was difficult and sometimes impossible in many studies to ascertain either the degree of strength or the extent to which the formal tenets of IT were adhered. Information important to generalization was also frequently inadequate. This included descriptive information on clients/subjects, on therapists/experimenters, on the treatment setting, and on the magnitude or severity of the target problem.

These deficiencies may be remedied if journal review boards require that authors provide such information. Without thorough reporting, one is left with such global questions as "Do interventions labeled as implosion therapy generally appear to be more effective than no treatment?" rather than more interesting and useful questions such as "With what populations and what problems treated under what conditions does implosion therapy appear to be most effective?"

CONCLUSION

This paper has presented an overview of the steps involved in a quantitative literature review with particular attention to the judgments, choices, and compromises commonly encountered. Broadly applicable points were illustrated by specific examples from one research synthesis study. Extensive references have been provided throughout as a resource to the reader interested in conducting a quantitative synthesis study or in tracking recent methodological developments.

The field is clearly at a point of moving beyond the novelty appeal of a new technology. Second stage questions regarding potential pitfalls, tradeoffs, and limitations as well as benefits are coming to the fore. Early fervent and, at times, adversarial positions pitting quantitative against qualitative approaches are giving way to a dialogue of how best to synthesize the strengths of "new" and "old" technologies.

One of the major caveats regarding use of quantitative synthesis techniques is also an important benefit. The caveat involves the threat to construct validity and, subsequently, to meaningful interpretations when findings across a heterogenous sample of studies (and, consequently, of measure, population, and intervention characteristics) are combined. The benefit is the ability to begin to assess the difference these differences make through such means as stratifying analyses by key outcome or mediating variables. Clearly, quantitative synthesis techniques cannot remedy methodological or substantive flaws or weaknesses in our knowledge base. What they may be useful in accomplishing is taking better advantage of our heterogenous and complex literatures filled with discrepancies and variation. Rather than inflate an aura of scientism, the present authors have attempted: (a) to emphasize the critical importance of a firm conceptual grounding by investigators engaged in quantitative synthesis of any given substantive domain, (b) to highlight the multitude of judgments and compromises involved, and (c) to point out the reality that these dilemmas and choice points exist in *any* literature review—that quantitative synthesis methods merely make them more evident and explicit.

Certainly, continued refinement and revision are needed. Clearer norms are required for resolving the many dilemmas and decisions commonly encountered.

Additional procedures of aggregating data reported in varied forms would prove extremely useful. The application of data synthesis methods to a broader spectrum of designs and comparisons is beginning to emerge. Lambert and his colleagues (Lambert, Hatch, Kingston, & Edwards, 1986) have used effect size measures to compare common rating scales for depression while Berman and Norton (1985) have moved one step beyond the issue of therapy effectiveness by asking if professional therapists produce patient outcomes bigger than paraprofessional therapists. Gingerich (1984) and Corcoran (1985) have begun to consider how meta-analytic procedures may be used to synthesize applied ($N = 1$) time-series data. These efforts are especially important given the growing literature single-subject, clinical research.

Among the benefits contributed by the developing quantitative synthesis technology have been an increased reliability in review outcomes and a greater awareness of the importance of providing important study detail. Additional byproducts have included a more thorough exploitation of our existing knowledge base and the subsequent guidance in asking more fruitful questions. Finally, new technologies make possible the study of new problems. As our burgeoning knowledge base continues to swell and the questions it raises become increasingly complex, a creative and balanced use of both known and innovative methods appears to be the most promising course.

REFERENCES

Bandura, A. (1978). On paradigms and recycled ideologies. *Cognitive Therapy and Research*, 2, 79-103.

Baum, M. L., Anish, D. S., Chalmers, T. C., Sacks, H. S., Smith, H., & Fagerstrom, R. M. (1981). A survey of clinical trials of antibiotic prophylaxis in colon surgery: Evidence against further use of no treatment controls. *New England Journal of Medicine*, **305**, 795-799.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep. *American Psychologist*, **37**, 245-257.

Berman, J. S., & Norton, N. C. (1985). Does professional training make a therapist more effective? *Psychological Bulletin*, **98**, 401-407.

Bryant, F. B., & Wortman, P. M. (1984). Methodological issues in the meta-analysis of quasi-experiments. In W. H. Yeaton & P. M. Wortman (Eds.), *Issues in data synthesis*. San Francisco, CA: Jossey-Bass.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Chalmers, T. C. (1982). The randomized controlled trial as a basis for therapeutic decisions. In J. M. Lachin, N. Tygstrup, & E. Juhl (Eds.), *The randomized clinical trial and therapeutic decisions*. New York: Marcel Dekker.

Cochran, W. J., & Cox, G. M. (1957). *Experimental designs* (2nd ed.). New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.

Cohen, J, (1977). *Statistical power analysis for the behavioral sciences* (rev. ed). New York: Academic Press.

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

Cook, T. D., & Gruder, C. L. (1978). Metaevaluation research. *Evaluation Quarterly*, **2**, 5-51.

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

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

Cooper, H. M. (1981). On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology*, **41**, 1013-1018.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, **52**, 291-302.

Cooper, H. M. (1984). *The integrative research review: A systematic approach*. Beverly Hills, CA: Sage.
Corcoran, K. (1985). Aggregating idiographic data: A meta-analytic statistic for single-subject research. *Social Work Research and Abstracts*, **21**, 9-12.
Cordray, D. S., & Orwin, R. C. (1983). Improving the quality of evidence: Interconnections among primary evaluation, secondary analysis, and quantitative syntheses. In R. J. Light (Ed.), *Evaluation studies review annual* (Vol. 8). Beverly Hills, CA: Sage.
Crown, S. (1981). Review of the benefits of psychotherapy. *British Journal of Psychiatry*, **139**, 199.
Dersimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review*, **53**, 1-19.
Eysenk, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, **33**, 517.
Gallo, P. S. (1978). Meta-analysis: A mixed meta-phor. *American Psychologist*, **33**, 515-517.
Garfield, S. L. (Ed.). (1983). Meta-analysis and psychotherapy. (Special issue). *Journal of Consulting and Clinical Psychology*, **51**(1).
Gingerich, W. (1984). Meta-analysis of applied time-series data. *Journal of Applied Behavioral Science*, **20**, 71-79.
Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, **5**, 3-8.
Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, **5**, 351-379.
Glass, G. V., & Kliegl, R. M. (1983). An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology*, **51**, 28-41.
Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Education Evaluation and Policy Analysis*, **1**, 2-16.
Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, **10**, 103-116.
Hartmann, D. P. (Ed.). (1982). Using observers to study behavior. *New directions for methodology of social and behavioral sciences*. San Francisco: Jossey-Bass.
Hays, W. L. (1973). *Statistics and the social sciences*. New York: Holt, Rinehart, Winston.
Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, **92**, 490-499.
Hedges, L. V. (1984). Advances in statistical methods for meta-analysis. *New Directions for Program Evaluation*, **24**, 25-42.
Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, **88**, 359-369.
Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research*, **50**, 438-460.
Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, **15**, 336-352.
Johnson, D. W., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. (1981). Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin*, **89**, 47-67.
Johnson, D. W., Maruyama, G., & Johnson, R. T. (1982). Separating ideology from currently available data: A reply to Cotton and Cook and McGlynn. *Psychological Bulletin*, **92**, 186-192.
Kazrin, A., Durac, J., & Agteros, T. (1979). Meta-meta analysis: A new method for evaluating therapy outcomes. *Behavior Research and Therapy*, **17**, 397-399.
Kendall, P. C., & Norton-Ford, J. D. (1982). Therapy outcome research methods. In P. C. Kendall & J. N. Butcher (Eds.), *Research methods in clinical psychology*. New York: Wiley.
Kulik, C. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal*, **19**, 415-428.
Lambert, M. J., Hatch, D. R., Kingston, M. D., & Edwards, B. C. (1986). Zung, Beck, and Hamilton rating scales as measures of treatment outcome: A meta-analytic comparison. *Journal of Consulting and Clinical Psychology*, **54**, 54-59.
Landman, J. T., & Dawes, R. M. (1982). Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist*, **37**, 504-516.

Leviton, L. C., & Cook, T. D. (1981). What differentiates meta-analysis from other forms of review? *Journal of Personality*, **49**, 231-236.

Levis, D. J., & Hare, N. (1977). A review of the theoretical rationale and empirical support for the extinction approach of implosion (flooding) therapy. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 4). New York: Academic Press.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, **76**, 365-377.

Light, R. J. (1980). Synthesis methods: Some judgment calls that must be made. *Evaluation in Education: An International Review Series*, **4**, 3-17.

Light, R. J. (Ed.). (1983). *Evaluation Studies Review Annual* (Vol. 8). Beverly Hill, CA: Sage.

Light, R. J. (1984). Six evaluation issues that synthesis can resolve better than single studies. *New Directions for Program Evaluation*, **24**, 57-73.

Light, R. J., & Pillemer, D. B. (1982). Numbers and narrative: Combining their strengths in research review. *Harvard Educational Review*, **52**, 1-26.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions around different studies. *Harvard Education Review*, **41**, 429-471.

Marshall, W. L., Gauthier, J., & Gordon, A. (1979). The current status of flooding therapy. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 7). New York: Academic Press.

McDaniel, M. A. (1983). *The MAME meta-analysis computer program*. P.O. Box 227, Arlington, VA 22210.

Miller, T. I. (1977). *The effects of drug therapy on psychological disorders*. Unpublished dissertation, University of Colorado.

Miller, R. C., & Berman, J. S. (1983). The efficacy of cognitive behavior therapies: A quantitative review of the research evidence. *Psychological Bulletin*, **94**, 39-53.

Mitchell, C., & Hartmann, D. P. (1981). A cautionary note on the use of omega squared to evaluate the effectiveness behavioral treatments. *Behavioral Assessment*, **3**, 93-100.

Nurius, P. S. (1984). Utility of data synthesis for social work. *Social Work Research and Abstracts*, **20**, 23-32.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, **8**, 157-159.

Orwin, R. G., & Cordray, D. S. (1983). *The effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis*. Unpublished manuscript, Northwestern University.

Pillemer, D. B., & Light, R. J. (1980). Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review*, **50**, 176-195.

Presby, S. (1978). Overly broad categories observe important differences between therapies. *American Psychologist* **33**, 514-515.

Rachman, S. J., & Wilson, G. T. (1980). *The effects of psychological therapy* (2nd ed.). New York: Pergamon Press.

Rimland, G. (1979). Death knell for psychotherapy? *American Psychologist*, **34**, 192.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, **85**. 185-193.

Rosenthal, R. (1979). The 'filedrawer problem' and tolerance for null results. *Psychological Bulletin*, **86**, 638-641.

Rosenthal, R. (1980). Quantitative assessment of research domains. In R. Rosenthal (Ed.), *New directions for methodology of social and behavioral science* (No. 5). San Francisco: Jossey-Bass.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology*, **51**, 4-13.

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

Rosenthal, R., & Rubin, D. B. (1982a). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, **74**, 166-169.

Rosenthal, R., & Rubin, D. B. (1982b). Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500-504.

Rosenthal, R., & Rubin, D. B. (1982c). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology*, **74**, 708-712.

Rothman, J. (1980). *Social R & D: Research development in the human services*. Englewood Cliffs, NJ: Prentice-Hall.

Sacks, H., Chalmers, T. C., & Smith, H. (1982). Randomized versus historical controls in clinical trials. *The American Journal of Medicine*, **72**, 233-240.

Sechrest, L., & Yeaton, W. H. (1981). Assessing the effectiveness of social programs: Methodological and conceptual issues. *New Directions for Program Evaluation*, **9**, 41-56.

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome research: A critical appraisal. *Behavioral Psychotherapy*, **10**, 4-25.

Smith, M. L. (1980). Publication bias and meta-analysis. *Evaluation in Education: An International Review Series*, **4**, 22-24.

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

Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and is relationship to attitudes and instruction. *American Educational Research Journal*, **17**, 419-433.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *Benefits of psychotherapy*. Baltimore: Johns Hopkins University Press.

Sohn, D. (1980). Critique of Cooper's meta-analytic assessment of the findings on sex differences in conformity behavior. *Journal of Personality and Social Psychology*, **39**, 1215-1221.

Stampfl, T. G., & Levis, D. J. (1967). Essentials of implosive therapy: A learning-theory-based psychodynamic behavioral therapy. *Journal of Abnormal Psychology*, **72**, 496-503.

Stampfl, T. G., & Levis, D. J. (1973). *Implosive therapy: Theory and technique*. Morristown, NJ: General Learning Press.

Strube, M. J. (1981). Meta-analysis and cross-cultural comparison: Sex differences in child competitiveness. *Journal of Cross-Cultural Psychology*, **12**, 3-20.

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

Strube, M. J., & Hartmann, D. P. (1982). A critical appraisal of meta-analysis. *British Journal of Clinical Psychology*, **21**, 129-139.

Strube, M. J., & Hartmann, D. P. (1983). Meta-analysis: Techniques, applications, and functions. *Journal of Consulting and Clinical Psychology*, **51**, 14-27.

U.S. General Accounting Office (1983). *The evaluation synthesis*. Institute for Program Evaluation: Methods, Paper 1, Washington, DC.

Walberg, H. J. (1983). Synthesis of research on teaching. In M. C. Wittrock (Ed.), *The third handbook of research on teaching*. Washington, DC: American Educational Research Association.

Walberg, H. J., & Haertel, E. H. (Eds.). (1980). Research integration. The state of the art. *Evaluation in Education: An International Review Series*, [Special issue], **4**, 1-142.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, **91**, 461-481.

Wilson, G. T., & Rachman, S. J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology*, **51**, 54-64.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills: Sage.

Wortman, P. M. (1981). Randomized clinical trials. In P. M. Wortman (Ed.), *Methods for evaluating health services*. Beverly Hills, CA: Sage.

Wortman, P. M., & Yeaton, W. H. (1983). Syntheses of results in controlled trials of coronary artery bypass graft surgery. In Richard J. Light (Ed.), *Evaluation Studies Review Annual* (Vol. 8). Beverly Hills, CA: Sage.

Wortman, P. M. (1983). Evaluation research: A methodological perspective. *Annual Review of Psychology*, **34**, 223-260.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, **49**, 156-167.

Yeaton, W. H., & Wortman, P. M. (Eds.). (1984). *Issues in data synthesis*. Beverly Hills: Sage.

Yeaton, W. H., & Wortman, P. M. (1984). Evaluation issues in medical research synthesis. In W. H. Yeaton & P. M. Wortman (Eds.), *Issues in data synthesis*. Beverly Hills: Sage.

Yeaton, W. H., & Wortman, P. M. (1985). Medical technology assessment: The evaluation of coronary artery bypass graft surgery using data synthesis techniques. *International Journal of Technology Assessment in Health Care*, **1**, 125-146.