

1 OF 2

United States General Accounting Office

GAO

**Program Evaluation and Methodology
Division**

**Revised
March 1992**

The Evaluation Synthesis

GAO/PEMD-10.1.2

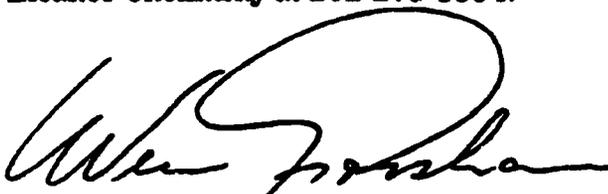
Preface

GAO assists congressional decisionmakers in their deliberative process by furnishing analytical information on issues and options under consideration. Many diverse methodologies are needed to develop sound and timely answers to the questions that are posed by the Congress. To provide GAO evaluators with basic information about the more commonly used methodologies, GAO's policy guidance includes documents such as methodology transfer papers and technical guidelines.

The Evaluation Synthesis presents techniques by which questions about a federal program are developed collaboratively with congressional committee staff, existing studies addressing those questions are identified and collected, and the studies are assessed in terms of their quality and, based on the strength of the evidence supporting the findings, used as a data base for answering the questions. The end-product is information about the state of knowledge in relation to the particular questions at a particular point in time.

The evaluation synthesis seeks to address the needs of a client for the rapid production of information relevant to a specific program and the analysis of large amounts of sometimes conflicting information on the topic. Conflicts cannot always be readily resolved, of course, but sometimes they can be when it turns out, for example, that one study has been soundly designed, implemented, and reported, whereas another has been inappropriately designed for the questions it seeks to answer. In addition to meeting these needs, the evaluation synthesis develops an agenda showing clearly where the gaps in needed information are that call for new agency research, and it also lays the groundwork for further evaluation or audit work. This reissued version supersedes the April 1983 edition.

We look forward to receiving comments from the readers of this paper. They should be addressed to Eleanor Chelimsky at 202-275-1854.



Werner Grosshans
Assistant Comptroller General
Office of Policy



Eleanor Chelimsky
Assistant Comptroller General
for Program Evaluation and
Methodology

BLANK PAGE

Contents

Preface		1
<hr/>		
Chapter 1		6
Defining Evaluation Synthesis	What Exactly Is Evaluation Synthesis?	6
	Steps in an Evaluation Synthesis	9
<hr/>		
Chapter 2		12
Why Do Evaluation Synthesis?	Why Evaluation Synthesis Is Important	12
	The Strengths and Limitations of Synthesis	14
	Evaluation Synthesis Can Guide Future Research	16
<hr/>		
Chapter 3		18
Developing the Synthesis	Specifying the Questions	18
	Gathering the Studies	23
	Developing Criteria for Choosing Studies	27
	Organizing and Implementing a Reviewing Strategy	30
	Redetermining the Appropriateness of the Synthesis Method	33
	An Example	34
<hr/>		
Chapter 4		37
Performing the Synthesis	Quantitative Approaches for Evaluation Synthesis	38
	Special Problems of Quantitative Synthesis	42
	Nonquantitative Approaches in Evaluation Synthesis	48
	Presenting the Findings	56
<hr/>		
Chapter 5		59
Evaluation Synthesis Can Answer Questions a Single Study Cannot	Why Interaction Effects Are Important	59
	Why Synthesis Is Useful in Identifying Interactions	60
	Summary	76

Contents

Chapter 6		78
Special Topics in Evaluation Synthesis	Comparing and Contrasting Studies and Their Findings	78
	Merging the Quantitative and Nonquantitative Approaches	80
	Exploiting Differences in Study Findings	90
	Anticipating Problems That Might Emerge	93
<hr/>		
Bibliography		100
<hr/>		
Glossary		125
<hr/>		
Contributors		128
<hr/>		
Papers in This Series		129
<hr/>		
Tables	Table 4.1: Mean Birthweight Quantitative Summary	46
<hr/>		
Figures	Figure 1.1: Sequence of Steps in Evaluation Synthesis	10

Abbreviations

CETA	Comprehensive Employment and Training Act
GAO	General Accounting Office
OEO	Office of Economic Opportunity
WIC	Special Supplemental Food Program for Women, Infants, and Children

Defining Evaluation Synthesis

To provide timely yet comprehensive and integrated information to a client on how a program is working, one approach that the General Accounting Office (GAO) applies is a cluster of techniques known collectively as the evaluation synthesis. This approach addresses the problem of timeliness by making use of existing evaluations. The evaluation synthesis is a methodology for addressing questions that can be satisfactorily answered without conducting primary data collection; it is not a replacement for original data collection.

The evaluation synthesis has two major benefits. First, the ability to draw on a number of soundly designed and executed studies adds great strength to the knowledge base when findings are consistent across different studies conducted by different analysts using different methods. No single study, no matter how good, can have this kind of power. Second, when studies are not well designed and executed, the knowledge that there exists no firm basis for action is also an important benefit: the size of the risk being taken is clarified, necessary caution is introduced into the debate, and over the long term, the number of failed shots in the dark is likely to be diminished.

What Exactly Is Evaluation Synthesis?

An evaluation synthesis is a systematic procedure for organizing findings from several disparate evaluation studies. It enables the evaluator to gather results from different evaluation reports, performed by different people at different places and at different times, and to ask several questions about this group of reports. Some of the questions are broad; others are quite specific and narrow.

An evaluation synthesis can answer several different kinds of questions—about overall program effectiveness, about specific versions of the program that are working especially well or especially poorly, and about how to organize future evaluation studies to

provide even more useful information about a program.

GAO has used the evaluation synthesis to answer congressional questions about both how programs are operating and what their effects are. For example, the evaluation synthesis can provide an estimate of how many people are actually receiving program services. The report entitled Disparities Still Exist in Who Gets Special Education on the Education for all Handicapped Children Act used 14 existing studies and two data bases to describe the handicapped children receiving special education services (GAO, September 1981). This report was able to use different sources not only to provide an estimate of how many children were receiving services but also to describe their racial and ethnic background and the severity of their handicaps. No study provided estimates on each description, nor did multiple estimates necessarily agree.

Similarly, we have used the evaluation synthesis to determine how many people need a program service. The special education report again serves as an example. The studies allowed for an examination of this issue, including estimates of particular handicapping conditions underrepresented and grade and age levels with particular underrepresentation.

As for program effects, GAO's 1982 report on the Comprehensive Employment and Training Act (CETA), for example, examined the effects of CETA programs on disadvantaged adult enrollees (GAO, June 1982). Entitled CETA Programs for Disadvantaged Adults—What Do We Know About Their Enrollees, Services, and Effectiveness? the report was able to provide estimates of CETA participants' experiences before and after program participation with respect to wages earned and time employed, public benefits received, and private sector employment. Additionally, estimates were provided for participants'

experiences by type of CETA service received. Follow-up reports from the Continuous Longitudinal Manpower Survey provided the data base. Another report used the evaluation synthesis method to study the effectiveness of expanded home health care services to the elderly (GAO, December 1982). Estimates of effect were provided for client outcomes and cost. Twelve major studies were used in determining the estimates.

We have also used the evaluation synthesis to compare the performance of two or more programs. For example, *Lessons Learned From Past Block Grants: Implications for Congressional Oversight* examined the question of whether the poor and other disadvantaged groups have been served equally under block grants and categorical programs (GAO, September 1982). Eight basic evaluation studies, some comprising a series of reports, were used.

As these examples show, the evaluation synthesis brings together existing studies, assesses them, and uses them as a data base for answering specific congressional questions. It enables evaluators to determine what is actually known about a particular topic, estimate the confidence (based on study methodology and execution) that can be placed in the various studies used in the data base and their findings, and identify gaps that remain in evaluative research with regard to the congressional questions.

Designed to be performed in a short time period, the evaluation synthesis has the important advantage of low cost. One or two persons with sufficient expertise typically can provide an evaluative summary of the state of knowledge in a particular area. The precise amount of time necessary depends on the narrowness of the topic area, the size of the data base available, and the familiarity of the evaluators with the topic and the data base.

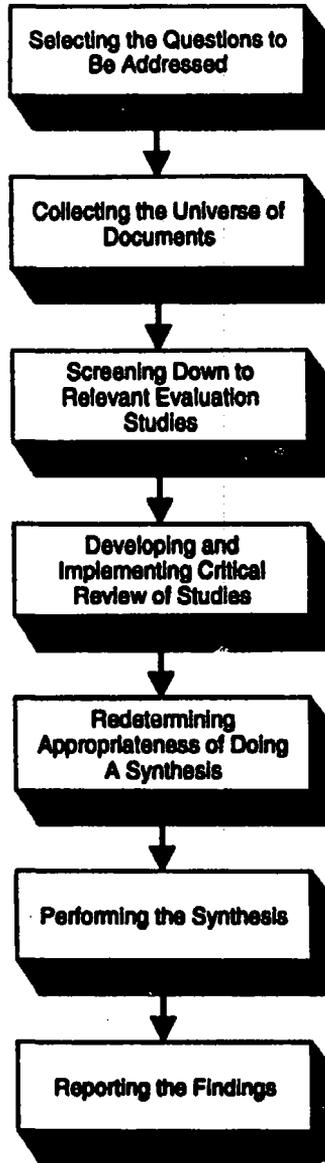
An additional advantage of the evaluation synthesis method is that by integrating evaluation findings, it establishes an easily accessible base of knowledge and identifies knowledge gaps or needs with respect to a specific topic upon which future evaluations can build. It can integrate administrative data and findings from studies with either qualitative or quantitative emphasis. It improves the use made of evaluative information since, in and of itself, it helps ensure the systematic legislative use of evaluations that have already been completed.

What differentiates the evaluation synthesis from the many other efforts involving the review and analysis of evaluative literature is that, as part of an overall strategy, it is designed backward from the end-use. That is, the evaluation synthesis is driven not by the quest to increase knowledge but by a specific need—requested or anticipated—for certain information. This means that the work must always begin with a framework of questions that impart logical cohesion to the effort. Some of the questions may be answerable by the available information but others may not be. Those left unanswered serve to identify gaps in the desired array of information.

Steps in an Evaluation Synthesis

Throughout this document, we will give detailed suggestions on how to organize and carry out an evaluation synthesis. We also give several illustrations that clarify how to implement each suggestion. But it is helpful to begin by summarizing the seven steps that all evaluation syntheses require. They are shown in figure 1.1.

Figure 1.1: Sequence of Steps in Evaluation Synthesis



Chapter 1
Defining Evaluation Synthesis

The seven steps fall under the two broad categories in chapters 3 and 4 of developing the synthesis and performing the synthesis. In the seven steps, the evaluator should

- **specify the questions: how questions are stated can determine how an evaluation synthesis is organized;**
- **gather the documentation: collect journal articles, bibliographies from computerized data bases, and unpublished evaluations and research reports and ask evaluators in the field to identify key studies;**
- **develop criteria for choosing studies: justifying the initial decision of which studies to include is crucial;**
- **organize and implement a reviewing strategy: assess studies against basic standards for research design, conduct, analysis, and reporting;**
- **redetermine the appropriateness of the synthesis method: this takes place after a preliminary review of the available evidence;**
- **implement the evaluation synthesis and check for problems: the synthesis can be done using quantitative and qualitative evidence, and it is particularly helpful to anticipate problems that may occur and to take steps at the outset to minimize them;**
- **present the findings: be sure to state the objectives, to describe the scope and methodology, and of course to respond clearly and concisely to the questions that were asked.**

Why Do Evaluation Synthesis?

Evaluation synthesis is a formal technical procedure for combining the results from several empirical studies. We use the word "formal" to indicate that the execution of the synthesis is not specific to a particular evaluator or a particular set of studies. In fact, its systematic nature is its primary strength. Two evaluators using the same synthesizing procedure should arrive at the same statistical output, although their interpretations of the output may differ.

Why Evaluation Synthesis Is Important

Faced with tens or even hundreds of studies on a single topic, an evaluator unarmed with systematic procedures is forced to use subjective criteria for deciding how to synthesize. The evaluator may choose several favorite studies, relatively well done from a classical experimental design standpoint. Or evaluators may favor studies carried out by investigators they respect. In either case, their impressionistic conclusions will often differ from those of other well-intentioned evaluators. A good example of two evaluators' differing dramatically in their interpretations of the same set of studies is provided by the debate between Munsinger (1974, 1978) and Kamin (1978) concerning studies of adopted children's IQ's.

Several researchers in the 1970's (Glass, 1977; Kulik, Kulik, and Cohen, 1979; Light and Smith, 1971; Rosenthal, 1978) commented on the unsystematic ways that social science research findings were being synthesized. They argued that the typical literature review was highly subjective and fell far short of rigorous scientific standards for the accumulation of evidence. In response, they tried to develop procedures for combining the results of independent studies. The goal was to draw, in a systematic manner, as much information as possible from existing evidence. (Hedges, 1988; Wachter and Straf, 1990)

So one clear importance of synthesis relates to research. Narrative overviews of prior findings, while offering a certain contextual richness that a single technical index cannot, generally do not provide the systematic information a researcher needs to design more powerful future investigations. (Rubin, 1990)

A second importance involves rendering scientific research useful to public policy. Policy decisions need to be made. If research findings are to inform policy, they must be put into an understandable form and provide answers or partial answers. The answers are occasionally clear-cut, but more often they are likely to be more complex, reflecting the real-world relationships between policy variables and outcomes. For example, the question "What are the effects of title I legislation?" does not have a simple answer. Certain programs under that legislation may work while others may seem to fail. The effects of a particular program depend upon a variety of factors such as who participates, the size of the community, and how the money is distributed.

Even when a policy question is complex, there is a strong need for summary information. A narrative description of 100 studies is frequently not enough. If there is not a single "main effects" answer, and if a program's success depends largely on setting-by-treatment interactions, synthesis may succeed in identifying and summarizing these interactions concisely. This can lead to guidelines about where and how to implement a particular program, improving the chances for its success. (Cordray, 1990)

The Strengths and Limitations of Synthesis

Strengths

The major advantage of the evaluation synthesis is its ability to provide relatively inexpensive, comprehensive, and timely information. It is designed to be conducted by one or two persons, with methodological expertise, and can be performed usually in 6 to 9 months. By integrating findings from already completed studies, the evaluation synthesis can potentially serve a client's needs for relatively short-term evaluative information. The focus of the evaluation synthesis is tailored to specific concerns.

Another strength is that the evaluation synthesis can increase the power of the individual study finding. Confidence in a number of well-done studies with the same finding is greater than in the finding of a single well-done study.

By drawing together information about a specific question from a disparate number of completed evaluation studies, the evaluation synthesis also creates a common knowledge base about a particular topic. It clearly sets out what is known—and with what level of confidence—and what is not known about the topic, thus enabling program managers and evaluation units to determine where they might best commit future evaluation resources. Thus, a particularly valuable feature of the synthesis is the identification of remaining unanswered questions.

Finally, the evaluation synthesis can serve, to a limited extent, as a check on the quality of the evaluations being performed concerning a particular program. The technical review of each study identifies methodological strengths and weaknesses that

influence a sponsor's posture with regard to future studies. GAO's synthesis of special education studies, for example, found that many study reports did not adequately describe the methodology they employed. The U.S. Department of Education indicated in its comments on the report that it had reviewed the studies the report used and agreed that the criticism was valid. Since most of the studies were conducted under contract, the department indicated that with approval from its Office of Procurement and Management, a requirement to include a methods description in final reports could be written into future requests for proposals. (Bornstein, 1989; Bowers and Clum, 1988; Chalmers et al., 1981; Cordray, 1990; Dush, Hirt, and Schroeder, 1989)

Limitations

The main limitations of the evaluation synthesis methodology stem from its reliance on extant data. The methodology is best applied to areas in which there is a base of evaluation information. Policy concerns for which there is little or no existing study information cannot be satisfactorily investigated. Thus, the methodology will not be appropriate for new programs where evaluation studies have not been completed (or perhaps even initiated) and no existing information base has applicability.

Even when a substantial information base is available, the evaluation synthesis is limited in that it can answer questions only to the extent that the existing studies have addressed them. Thus, for example, findings in response to a particular question may or may not be generalizable to the nation, depending on the nature of the relevant studies conducted on this topic.

Poor reporting also limits the evaluation synthesis. Procedures may have been described in so brief a manner that judgments cannot be made about a study's technical adequacy. Additionally, in experimental or quasi-experimental studies,

treatments may have been so minimally described that judgments cannot be made about the similarities and differences across studies or variables of interest may not have been reported consistently across studies. Some studies may report demographic data such as sex, age, and education, for example, while other studies focusing on the same questions do not. The evaluation synthesis is limited by the form and quality of the reports it uses.

Finally, the evaluation synthesis is only as current as the studies it analyzes. If studies are several years old, they may have identified findings that program managers have already taken steps to address and that are no longer characteristic of the program. The methodology is no substitute for primary data collection, but it is useful when questions can be answered using information from existing studies and when time is short. (Feingold, 1988; Hazelrigg et al., 1987; Johnson and Eagly, 1989; Parker et al., 1988; Yeaton and Wortman, 1989)

Evaluation Synthesis Can Guide Future Research

We have emphasized looking carefully at existing data to see where things stand now. But some syntheses are undertaken primarily to help guide future research. Their goal is to suggest to the designer of the eleventh study what can be learned from the first 10. The evaluation synthesis can provide such guidance in at least two ways.

First, the synthesis can help by identifying the most promising experimental manipulations and comparisons. With finite resources, it is not possible to build all variables formally into each effort. A review can examine a large number of possible variables that might be important and eliminate many of them as serious candidates for new research. If hospital size is not related to surgical success in 10 well-done studies, it is unlikely to emerge as crucial in the eleventh. Using a review to reduce the number of

experimental variables should improve the statistical power and guide the allocation of resources in a new study.

Second, the synthesis can help researchers choose between organizing one big new effort at a single site and organizing a series of small efforts at many sites. Suppose funds are available to evaluate a new treatment for breast cancer involving 1,000 women. Is it better to conduct one study with all 1,000 women at one hospital or to commission five smaller studies in five different hospitals with 200 patients each? Existing research can guide this decision. On the one hand, suppose past evaluations show little variation in the success of cancer treatment across different hospitals. Then the wisest decision probably is to focus the entire new effort at one site. The large sample size will help identify subtle ways in which the new treatment differs from current practice. On the other hand, suppose a review of earlier findings shows the value of cancer treatment to vary widely across sites. Then it could be a mistake to focus on one particular setting or hospital. Setting-by-treatment interactions should be expected. This expectation can only be assessed by trying the new cancer treatment in several places.

The particular guidance a research review provides will differ from one substantive area to another. These examples illustrate the benefits of designing into new research the messages of the old. The implication for evaluators is that simply concluding with the usual "more research is needed" is not enough. Evaluators must make a conscious effort to identify what specific directions new initiatives should take. This linking of past and present is crucial if research is to achieve its full potential for enhancing both science and policy.

Developing the Synthesis

The process of developing the synthesis is iterative. Through a series of steps shown in figure 1.1, the synthesis topic and information base are defined and reexamined. The first five steps of the evaluation synthesis covered in this chapter are specifying the question, gathering the studies, developing criteria for choosing studies, organizing a reviewing strategy, and redetermining the appropriateness of the synthesis method.

Specifying the Questions

The three most common questions that may be answered with an evaluation synthesis are

1. For any program or treatment, what is its effect on the average?
2. Where and with whom is the program or treatment particularly effective or ineffective?
3. Will it work here? In other words, what are practical guidelines for implementing the program or treatment in a particular place?

Different reviewers can approach the same group of studies with quite different goals. Policymakers often face decisions requiring an estimate of average program performance. For example, Blue Cross and Blue Shield must decide whether to offer third party payments for psychological counseling to persons who have just had cancer surgery. The goal is to enhance recovery rate and reduce morbidity. Here, an administrative regulator may simply want an answer to the question, "On the average, does psychological counseling after cancer surgery help people?" Researchers may think this is far too broad a question. But a policymaker's main concern is not with arranging perfect matches between psychologist and client. It is the need for a decision about whether postcancer therapy services should or should not be reimbursed.

Researchers might come at the problem differently. The researcher might decide that averaging across several mental health protocols, with different types of clients, misses the main objective. The more important questions here might be, "What kind of counseling is usually best and what sort of client benefits most from that type of counseling?" Researchers will not be surprised if a particular treatment does not work for everyone. Indeed, it is sometimes exhilarating to discover that a new treatment works for anyone. So a researcher will usually organize a review to go beyond an "on the average" question to examine what works, how well, and for whom.

Local program managers may have yet another purpose in mind. While interested in the question of what treatment is best for whom, their main focus is feasibility. Can an innovative treatment for breast cancer be implemented successfully in specific, real-world locales? It is one thing to learn that streptokinase administered at a certain time can help. It is another thing to build this finding into practice with good results. A local hospital director or physician will want to know what treatment works best in general, but any concrete evidence about what it takes to implement a treatment successfully in a specific environment (such as a small, rural hospital rather than an urban teaching center) will be particularly valuable. A review for this purpose will emphasize any available reports about implementation efforts at similar institutions.

To summarize, an evaluation synthesis designed to answer the "on the average" question emphasizes a search for main effects. A synthesis that asks "who benefits most from what" will focus on a search for interaction effects. The synthesis that asks "how it will work here" should emphasize qualitative details of the setting, the locale, and the context for a treatment. (Glass et al., 1981; Green and Hall, 1984; Hedges, 1986; Hedges and Olkin, 1985; Rosenthal, 1984)

Evaluation synthesis can be used to answer a wide variety of questions, including descriptive, normative, and impact (cause and effect) questions. For example, a report on the handicapped focused on "who was receiving" questions and whether these groups were over- or underrepresented with respect to the receipt of special education services. The specific questions were: Who does this program serve? (a descriptive question) To what extent are the intended beneficiaries being served? (a normative question). An evaluation that attempted to assess the effect of race on death penalty sentencing answered the following impact question: Does the race of either the victim or the defendant influence the likelihood that defendants will be sentenced to death?

The kinds of questions for which the evaluation synthesis may be appropriate, at least for service delivery types of programs, are, however, likely to fall into two distinct categories. These are program operations and program effects, both themselves components of the broad question of whether the program is working. While the specific wording of the questions will vary, examples are as follows. The first three are program operations questions.

Who does the program serve and to what extent are the intended beneficiaries being served? The report on the handicapped, for example, asked not only who was receiving services but also what groups were over- and underrepresented with respect to the receipt of special education services.

What are the program's services, what services are delivered to whom, what is the service delivery process, and are these consistent with program objectives? In a report on CETA, for example, GAO examined shifts in the mix of services over time in CETA programs. Services included classroom training, on-the-job training, work experience, and public service employment. We also investigated

differences in the characteristics of persons receiving these services—in other words, how were the services targeted?

What administrative processes and procedures are implemented? How is the program administered? In a study of lessons learned from past block grants, GAO investigated studies of the costs of administering block grants and the effects of fixed percentage caps on administration.

Here are four typical program effects questions.

What are the general outcomes for program recipients? A study on home health care, for example, investigated studies of the effects of expanded home health care on client longevity, satisfaction, physical functioning, and mental health.

Do program outcomes vary by type of recipient or types of service? The CETA study examined whether differences across service types (classroom training, on-the-job training, work experience, and public service employment), in the characteristics of participants, and in their occupational areas of employment and training were reflected in data on their experiences before and after CETA.

What is the program effect on other than program recipients? A major question in a study of expanded home health care services was the effect of expanded home health care on nursing home and hospital use.

How effective is the program in terms of costs, alternative programs, or different versions of the program? The CETA study, for example, investigated the effectiveness of CETA in terms of postprogram earnings that could be attributed directly to CETA participation in adult-oriented services. It also examined gains by service type to determine whether one type of service was more effective than another

type (for example, on-the-job training versus classroom training).

Any one or more of these general questions may serve as the basis for a limited yet comprehensive subset of questions that can be used to respond to the congressional need for program information. These questions provide a framework not only for conducting the evaluation synthesis but also for reporting the findings.

The process of selecting the precise topic and identifying the actual study questions drives the evaluation synthesis method. This is particularly important because the evaluation synthesis can answer only questions for which there already exists study information. Even then, it can answer questions only to the depth or extent that the evaluation studies have addressed them, and it can be only as current as the studies themselves.

It is important during question specification to conduct a preliminary review of the kinds of data available. Before settling on the study topic and questions, the evaluator must have some familiarity with the nature and extent of the evaluative information available on the proposed topic. The actual questions for investigation must be carefully formulated so that they are neither so broad that addressing all the pertinent evaluation information is not possible in a short time nor so narrow that little evaluation information is available for responding to them.

There is little limitation on the type of topical area suitable for evaluation synthesis. The method is as appropriate to defense topics, for example, as to social service delivery topics. Given the need, however, for a base of completed evaluation studies, the method is generally less applicable to new policies or programs. Conversely, for a program with a long life, it may be desirable to set a cut-off point for the time

period of program operations to be covered in the synthesis.

An important consideration in the early formulation of study questions is the degree of precision needed in the answers to be found. For instance, the client may wish to know how many people need a service or how many are receiving a service. An exact answer will be impossible. The answer will either be a formal confidence interval or, if the analysis is based on case studies or less rigorous methods, have the flavor of a confidence interval. (For this and other statistical concepts, see Ullman, 1978.) We mean by this that any synthesis should specify a range of possible values with some confidence that the true value is included in that range. How narrow that range of possible values must be to make the synthesis practically useful and how high the confidence level must be that the specified range includes the true value will vitally influence each of the next steps of evaluation synthesis.

The need to define questions, to determine the degree of precision needed in the answers, to assess the appropriateness of evaluation synthesis versus other possible methods, perhaps to renegotiate the original questions after having looked at the available evidence—these steps suggest an iterative, collaborative approach between the information-users and the evaluator.

Gathering the Studies

Once the specific questions have been developed—remembering that the questions can be developed soundly only if they are guided by at least some prior knowledge of the topical area and the existing evaluation literature—relevant evaluative information should be compiled. While the federal agency administering a policy or program is a natural place to begin, the evaluation synthesis method requires that the investigation go beyond this information base and include nonagency-sponsored

literature. Without including nonagency-sponsored literature, only a part of the universe of relevant studies is likely to be obtained, and it will not be clear how large a part of the universe has been obtained or how biased or representative it is.

In going to the agency, the objective is a thorough and comprehensive search for information. Background information such as legislative and funding histories and regulations should be obtained as well as relevant administrative or management information system data and evaluation studies. Summaries of data tapes (or the actual computer tapes) may additionally be collected as part of the data base. Secondary data analysis, while not a necessary part of the approach, may be appropriate in cases where existing data sets have not been fully exploited.

While the short time period and the focus on secondary data collection implied by the method dictate that interviews of agency officials and others be kept to a minimum, interviews may be needed to complete an understanding of the program and its evaluation and to identify ongoing and planned evaluation studies for which reports are not yet available. Again, visits to project sites are not routinely indicated, but they may also be informative.

Nonagency-sponsored literature covers all evaluation studies other than those initiated by the federal agency administering the policy or program. This includes studies conducted by other federal agencies in the executive branch; studies conducted by legislative agencies such as the General Accounting Office, Congressional Research Service, and Congressional Budget Office; studies undertaken independently by state and local agencies, national associations, and members of the academic community; or studies focusing on the same topic done in other countries. (An evaluation synthesis of the "guestworker" program experience, for example,

might need to consider the European literature and experience.) While it may be time-consuming and otherwise problematic to attempt to explore all these information sources, such efforts underlie and enhance the credibility and worth of the evaluation synthesis and, at a minimum, must be considered.

One pitfall in collecting the literature for the synthesis, as documented by White (1982), is that focusing only on published reports can lead to erroneous conclusions. White found that published research reports tended to have more significant positive findings than unpublished reports. Studies with less significant findings were less "newsworthy" and, therefore, usually not published. Thus, just examining published reports might lead to an inflated view of a program's effect (Abrami et al., 1988; Rosnow and Rosenthal, 1989; Shadish et al., 1989). Being sure that no major published or unpublished study has been omitted is usually a considerable challenge in an evaluation synthesis. One approach useful in preventing an omission is to ask the assistance of outside experts to help identify the literature and to review the literature collected in this way.

There are at least three specific steps that we recommend for organizing a systematic search for published articles for an evaluation synthesis. The first step is to use a computerized data base, accessing the data base by choosing key words. For example, a recent evaluation synthesis by Lipsey (1990) examined the following data bases for a synthesis of criminal justice research: AIM/ARM, Arts and Humanities Citation Index, Books in Print, British Books in Print, British Education Index, Child Abuse and Neglect, Criminal Justice Periodical Index, CRISP: National Institute of Mental Health, Dissertation Abstracts Online, ERIC, Family Resources, Federal Research in Progress, GPO Publications, Library of Congress Books, Medline, Mental Health Abstracts, National Criminal Justice Reference Service, National Technical Information

Service, PAIS International, Psychological Abstracts, Social Science Citation Index, Sociological Abstracts, SSIE Current Research, and U.S. Political Science Documents.

While a search of computerized data bases is crucial, and will generally identify most key articles and research reports, a good second step is to examine the lists of references at the end of key research reports. Such cross checking will often turn up additional cites, often cites that you did not initially locate simply because you used a key word that the authors of the original article did not use as their key word. If you identify several articles in this way, it is constructive to see if the large computerized data base actually has these additional articles in its list, and if yes, what key word these articles use. Finding a new key word may lead you to additional relevant articles.

Finally, a third step is to ask knowledgeable colleagues and fellow scholars around the country. If ever there were a good use for expert advice, this is the time.

The purpose that drives these approaches to identifying and gathering original studies is that of being all-inclusive. There is some risk, in the real world, that the computerized data base searches will turn up so many articles that the synthesis can become unwieldy. So an evaluator must be prepared to deal with the potentially enormous size of a full-fledged search and must be willing to tolerate, for some topic areas, an enormous set of potential studies to include in an evaluation synthesis. (Cooper, 1988, 1989)

Developing Criteria for Choosing Studies

Once the relevant literature has been identified and collected, the question becomes: What types of studies should the synthesis include? A goal of evaluation synthesis is the identification and control of potential sources of bias in the technical sense. If the studies used in the evaluation synthesis share common, usually unknown, sources of bias, the synthesis as a whole will take on that bias.

When determining which studies to include in an evaluation synthesis, the evaluator must also apply GAO's standards of evidence. Was the study sufficient—that is, did it provide enough factual and convincing evidence to support its findings and conclusions? This would include an assessment of whether statistical methods were appropriate. Was the evidence used in the study relevant? Evidence is relevant if it has a logical relationship to the assignment issues. The evaluator must also determine if the evidence used in the study is competent—valid and reliable. Using this criterion, the evaluator should independently assess the studies to be included. If such an assessment is not made, it must be stated in the body of the report.

This identification and control of bias requires, in part, an understanding of how variations in study methodology may influence results. For instance, Wortman and Yeaton (1983) were careful in their synthesis of studies on coronary bypass surgery to include both randomized and quasi-experimental studies. The two sets of studies produced markedly different estimates of the effect of the surgery. The investigation set out to account for the gap in the findings of the two sets of studies. They concluded that although the randomized experiments led to a different estimate than the quasi-experiments, a small part of the gap between the two estimates was attributable to biases in the randomized studies. Some patients were randomly assigned to have medical rather than surgical treatment, and the evaluators were able to account for a source of bias.

As the example above shows, whenever possible the evaluator should seek studies that use a variety of methods. Variations in study types may control bias and prove helpful in accounting for discrepancies in study findings, leading to more accurate answers to the client's questions. To illustrate again, suppose a congressional committee wanted to find out, first, how many people have been victimized by violent crime in each of the past 5 years and, second, how many of these victims have received services from programs providing aid to victims of violent crime. To answer the first question, studies might have used a variety of methods. For instance, some studies might be based on police reports, which tend to underestimate the number of crimes because many crimes go unreported. Other studies might have used surveys of a sample of people selected at random from a defined population. But, among a number of problems such studies may have, the populations might have been defined locally (so that all the people in a given city were equally likely to be surveyed) and since local crime rates vary, variations in estimates may reflect variations in local crime rates. This example underlines the importance of enlisting a representative sample of studies and study types so that the evaluation synthesis as a whole does not take on the bias of a single study type (Cordray, 1990).

If the congressional committee were interested in finding out how many people are receiving aid to victims of violent crime, there are again fundamentally different ways individual studies may be designed to provide an answer. One method, for example, is to identify all government programs providing aid to victims of violent crime, to retrieve evaluative information on these programs, and to derive from these records a count of people receiving aid. A second method is to consult victim surveys concerning violent crime as to whether people received government aid. Again, the key point is that in conducting a synthesis, one should include both kinds of studies, if

Chapter 3

Developing the Synthesis

they are available. The two methods may have built-in biases, and unless both are included, the synthesis takes on the bias of the individual studies providing data for it (Cooper, 1986, 1988).

A common criticism of narrative research summaries is that they are not objective or that they are too impressionistic. More-quantitative efforts, by contrast, should more "objectively" synthesize the available evidence. But if many studies of a particular treatment, such as a certain method for treating breast cancer, are available, a reviewer must decide which to include. Several options are available, and a decision must be made early in the review process.

The simplest option is to include every available study: published and unpublished reports, doctoral theses, academic studies, and contract research studies. When a reviewer has no prior hypothesis and wants to explore broadly what is known about a treatment, including such diversity may help. Scientific precision is less important than identifying interesting trends, patterns, or outliers.

But an evaluator faces difficult trade-offs in any plan to track down and include everything. For example, if it is clear that a certain study is fundamentally flawed, say with obvious numerical errors, it is hard to argue for its inclusion. Wrong information is not better than no information. Another example is that the details about a treatment may have changed over time. Including very old studies, even if they were well done, when the question driving a review is how well the treatment currently works, is foolish. A concrete illustration comes from Wortman and Yeaton's (1983) pooling of data from randomized trials of coronary artery heart bypass surgery. The survival rate has risen dramatically as surgical technique has advanced. If this improvement is quite obvious, do we want to include very old studies? Probably not, but this

decision will depend upon the specific policy problem.

A second option is to use a panel of experts to generate a list of studies for inclusion. Hauser-Cram and Shonkoff (1986) used this approach to choose studies for their review of the effectiveness of early intervention programs for young children with cerebral palsy, developmental delay, and Down's syndrome. A search of published literature yielded hundreds of studies that had some potential for inclusion in their summary. They used experts to sharply narrow the list of candidates. A quick caution here is that sometimes experts pay more attention to large studies of modest quality than to well-designed, smaller studies. Such bias should be controlled for.

Organizing and Implementing a Reviewing Strategy

Given the substantial number of evaluation studies that concern a topic of interest, some will probably have focused exclusively on the topic, while for others, addressing the topic may have been only a secondary study purpose. Some studies, as discussed in a previous section, are likely to have similar types of designs while others will have differed on design type and therefore also on the types and sources of data. As a group, it is likely that the studies will have varied in the soundness or rigor of procedures and execution and perhaps even the appropriateness of the design.

While it is important to include different types of studies in the evaluation synthesis, what does the evaluator do with studies that vary in quality? This is a question that has provoked heated debate. A critical issue in this debate is what constitutes a "good study." It seems reasonable that all studies included in a synthesis should be assessed against basic standards for research design, conduct, analysis, and reporting.

Thus, the evaluation synthesis requires an assessment of the overall soundness of each individual study.

Major weaknesses of study design, conduct, analysis, or reporting that affect the reliability or validity of each study's findings must be identified and considered in using the study and placing confidence in the study findings. Whether experiment, case study, survey, or content analysis, each study should be questioned as to its reliability and validity. Questions such as the following will determine the overall usefulness of the individual study to the evaluation synthesis:

- **Are the study's objectives stated? Were the objectives appropriate with respect to the developmental stage of the program?**
- **Is the study design clear? Was the design appropriate given the study objectives? Was the indicated design in fact executed?**
- **Did the variables measured relate to and adequately translate to the study objectives and are they appropriate for answering the client's questions?**
- **Are sampling procedures and the study sample sufficiently described? Were they adequate?**
- **Are sampling procedures such that policymakers can generalize to other persons, settings, and times of interest to them?**
- **Is an analysis plan presented and is it appropriate?**
- **Were data-collector selection and training adequate?**
- **Were there procedures to ensure reliability across data collectors?**
- **Were there any inadequacies in data collection procedures?**
- **Were problems encountered during data collection that affect data quality?**
- **Are the statistical procedures well specified and appropriate to the task?**
- **Are the conclusions supported by the data and the analysis?**
- **Are study limitations identified? What possibly confounds the interpretation of the study findings?**

This list shows some of the issues that should be raised in reviewing the studies. The information derived by answering these questions should lead to an overall judgment of the usefulness of each study. It does not mean, however, that studies with design or other weaknesses are automatically excluded from the synthesis. Instead, if such studies are included, a judgment should be made about the confidence that can be placed in the study findings in relation to other study findings.

Of particular concern, however, is the consistency or reliability of judgments of study quality. In a synthesis, for example, Stock et al. (1982) had coders judge a random sample of 30 primary research documents. Among the items requiring a coding decision was one global item called quality of the study. Correlation coefficients among the coders were not acceptable with a mean level of .52. The study suggests strategies for improving reliability, including summing ratings across methodological variables (as superior to a single global item rating), coder training and retraining, and group rather than individual judgments of quality.

At a minimum, the issue of coder reliability should be raised in the evaluation synthesis. It seems reasonable to describe steps taken to address the reliability issue, or as several GAO evaluation syntheses have done, to describe the strengths and weaknesses of the study that led to a summary judgment of quality and utility. A report synthesizing studies on special education, for example, included the actual review of each study as a technical appendix, making the basis for the judgment available for each reader to assess.

**Redetermining
the
Appropriateness
of the Synthesis
Method**

Is the available research sufficient to answer the client's questions? In developing the evaluation synthesis, it is useful to classify each study or data base that is to be included in the synthesis according to both the questions in the study framework that it addresses and the study design. This procedure ensures that all studies to be included in the synthesis are relevant, and it quickly shows commonalities as well as information gaps.

Sometimes, although preliminary evidence appeared sufficient, it may simply not be possible to answer a client's question using evaluation synthesis. For example, a GAO report collected a number of studies attempting to estimate the size of the illegal alien population in the United States. However, the range in estimates was enormous. It was possible to identify biasing factors in some cases. One household survey conducted in Mexico, for instance, quite clearly underestimated the number of Mexican citizens who had illegally emigrated to the United States. While this study put a lower bound on the true value, the quality of the remaining studies was so questionable, their results so discrepant, and potential explanatory factors so numerous in relation to the number of studies available that the evaluators concluded that a major new research effort rather than evaluation synthesis was required to answer the question. In this instance, the main use of synthesis was to help identify whether and what research was needed to uncover important features requisite for the design of such research.

There is a danger that a methodology that solves certain thorny problems of applied research will promise more than it can deliver. Evaluation synthesis is no exception. The purposes of redetermining the appropriateness of the synthesis method are the following:

1. To clarify information-user expectations before the evaluator becomes involved in the details of the synthesis itself.

Chapter 3
Developing the Synthesis

- 2. To enlist the collaboration of the client in addressing likely difficulties in the work.**
- 3. To prevent months of labor being wasted when synthesis is unlikely to meet the client's information needs.**
- 4. When synthesis is found inappropriate, to formalize and systematize the process whereby new research is recommended on the basis of gaps in past knowledge.**
- 5. If synthesis is found appropriate, to sharpen understanding of research questions just prior to immersion in the details of the work.**

An Example

To illustrate the steps in the evaluation synthesis process discussed in this chapter, let us consider what was done in a GAO study of the Special Supplemental Food Program for Women, Infants, and Children, or WIC (GAO, January 1984). The U.S. Senate Committee on Agriculture, Nutrition, and Forestry asked GAO to synthesize all available evidence about WIC. This program, funded at over \$1 billion a year, provides nutrition supplements to approximately 3 million people each year. These people are pregnant women from low-income families and children from birth to age 5 in low-income families who are considered at high nutritional risk. The Senate committee's request was motivated by the sharply conflicting testimony that it received about WIC's effectiveness. Some witnesses argued that it was a highly effective program and that it had clear positive effects in increasing children's birthweight, reducing fetal and neonatal mortality, improving nutrition in mothers and children, and reducing mental retardation in children. Other witnesses argued that there was no concrete evidence for these positive assertions. They testified that while it seemed hard-hearted to oppose the distribution of food vouchers to low-income mothers, the facts did not support assertions that

Chapter 3
Developing the Synthesis

women or their children benefited in any concrete way.

The questions that were finally agreed to as being the most relevant to the committee were as follows:

1. Does participating in WIC affect birthweights?
2. Does participating in WIC prevent miscarriages, stillbirths, and the mortality of the newborn?
3. Does participating in WIC affect the health and nutrition of pregnant women?
4. Does participating in WIC affect the incidence of anemia in infants and children?
5. Does participating in WIC affect the incidence of mental retardation in infants and children?

In proceeding to identify and collect the relevant universe of documents that possibly provided insights to the answers of all or any of these questions, the evaluators cast as broad a net as possible, including agency bibliographies; journals; discussions with many professionals in the field, among them nutritionists, health professionals, and researchers; and an iterative mailout of a list of documents to experts requesting additions as appropriate. Over 100 documents were identified, some containing more than one evaluation study report. From their first reading of this set of documents, the evaluators found 54 to be relevant because they contained information pertaining to one or more of the evaluation questions posed above. The evaluators then identified, within these 54 documents, 61 studies to be included in the synthesis to be performed.

The evaluators then developed a reviewing strategy that included establishing a nine-point scale to be used by expert reviewers to rate the credibility of each

Chapter 3
Developing the Synthesis

study for each question. The reviewers used a set of criteria regarding the soundness and appropriateness of the methodology underlying each study's findings and then assigned a numerical rating. In this way, each study was judged to be somewhere on a scale between high and low credibility. Each study was read by more than one evaluator. This review resulted in the matching of studies to questions and led to results like the following.

Question 1: effect on birthweights, 39 relevant studies, 6 of high or medium credibility and 33 of low credibility;

Question 2: effects on mortality, 12 relevant studies, of high or medium credibility;

Question 3: effects on maternal nutrition, 24 relevant studies, 6 of high or medium credibility and 18 of low credibility.

These steps then led to the point where the synthesis of information available for each question could proceed, if at least some studies of high or medium credibility had been identified. These procedures will be discussed in the next chapter.

Performing the Synthesis

Given a set of studies that have been individually assessed and deemed usable for the synthesis, the next steps are to implement the evaluation synthesis and check for problems and to present the findings. These are discussed in this chapter. The question is: How are the different studies compared? There is no standard approach, but two major factors will influence how the studies are compared. First, different evaluative questions are likely to require different approaches for synthesizing the information and, second, the nature of the study designs will limit the possible analyses.

As mentioned previously, the question that motivates the synthesis in large part drives the specific procedure used to synthesize. For example, in examining how well a program is working, the targeted question might be, Who does the program serve under ideal circumstances? Alternatively, Who does the program serve on the average? In the first instance, the evaluator might want to investigate a number of case studies and provide a narrative description of the findings. In the second instance, the evaluator might take the arithmetic average of the answers given by the individual studies available or might express the answer as the range between the highest and lowest estimates. A problem here is that, since the evaluation synthesis is employed to answer questions given existing information, the evaluator will not often find the ideal quantitative analysis possible.

As with the discussion on what studies to include in the synthesis, this is an area where considerable literature exists. The literature assumes for the most part, however, that the study designs are experimental or at least quasi-experimental in nature, which may, of course, not be the case. (Cooper, 1989; Hedges, 1986, 1988; Hedges and Olkin, 1982, 1985, and 1986)

This chapter discusses both quantitative and nonquantitative approaches to evaluation synthesis. Quantitative approaches are ideal for certain questions but nonquantitative approaches are what most evaluators will have to wrestle with when responding to questions driven by policy.

Quantitative Approaches for Evaluation Synthesis

The literature describes two basic quantitative approaches for synthesizing the findings of experimental or quasi-experimental studies. These approaches, detailed in the following sections, are (1) computing an average effect size and (2) conducting a combined significance test. It may be relatively uncommon to use these specific techniques in GAO work because of the character of the questions posed as well as the disparate, fragmented nature of existing evaluations. Quantitative approaches are, however, powerful tools when the basic assumptions can be met. (Bryk and Raudenbush, 1988; Green and Hall, 1984)

Computing an Average Effect Size

The key descriptive statistic that Glass (1981) employed in his pioneering synthesis is the effect size. When one compares a treatment to a control, a common definition of effect size is simply the difference between the two group averages, expressed in terms of the control group's standard deviation.

To illustrate, suppose we were studying two groups of teenagers, one group receiving a certain type of job training and the other receiving none. After a year on the job market, each person in both groups is asked about his or her income. If the average annual income for the group that received training is \$10,500, and the average for the group receiving no training is \$10,000, with a standard deviation of \$1,000, then the effect size for this program is simply 0.5, or half a standard deviation. There are several elaborations on

this basic idea, some of which incorporate the treatment group's standard deviation and others that are based on the idea of change over time. Our example provides a working definition that is congruent with Glass's extensive work. (Colditz et al., 1988)

Assuming that an effect size is reported (or can be computed) for each of several studies, the average effect size for the entire set is easily calculated. An important aspect of computing an average effect size is that it provides a single summary value for an entire area of study: "Most of our work is aimed at simple and sweeping generalizations that stick in the reader's memory. If what an integrative analysis shows cannot be stated in one uncomplicated sentence, then its message will be lost on all but a few specialists" (Glass, 1978, p. 3). For example, Glass and Smith (1976) computed the average effect size for psychotherapy across 400 separate studies to be .68. They concluded that, on the average, psychotherapy is beneficial, since "the average person receiving some form of psychotherapy was about two-thirds standard deviation more improved on an outcome measure than the average control group member" (Glass, 1977, p. 363).

Effect size averaging requires that we know the group means and the control group standard deviation. Estimating an average effect size is most clearly useful when a group of study outcomes seem neatly, perhaps normally, distributed around their mean. In this case, an average gives a useful single summary of results. But when study outcomes appear to conflict, or have an unusual distribution, a single average is less useful. (Feldman, 1971; Guzzo et al., 1987; Hedges, 1982, 1984; Hyde and Linn, 1988)

**Conducting a
Combined
Significance Test**

The relationship between sample size and the power of a statistical test is well known: the larger the sample size, the more likely that a certain effect will be detected as statistically significant. For example, an observed difference of 10 IQ points between Head Start and non-Head Start children may not be statistically significant with 10 children per group; however, this same 10-point difference can be highly significant with group sizes of 100.

When there exist a number of studies on the same topic, the various smaller data sets often can be pooled into a single overall analysis. This increases effective sample size and will dramatically improve the power of statistical tests. This approach is especially appealing when sample sizes of individual studies are small. Suppose we have several studies investigating the effectiveness of highly structured versus less-structured curriculums. All the studies may turn up concordant results, without any of the individual findings reaching statistical significance. Yet an overall test on the pooled data may show highly significant results.

When multiple independent studies all compare two treatments that are similar across studies and the group differences are tested statistically in each instance, one strategy for drawing a single "grand" conclusion from these results involves combining the separate significance tests into an overall test of a common null hypothesis. This is generally that both treatment groups have the same population mean.

A number of procedures using this idea have been suggested. Rosenthal (1978) summarized many of them and provided guidelines as to when they are likely to be most useful. To illustrate one technique, we take the method of adding Z scores (standard normal deviates). If two groups are compared in each study, there is a Z score associated with each reported p value. The Z's are added across studies, and their

Chapter 4 Performing the Synthesis

sum is divided by the square root of the number of studies that are combined. The probability value associated with the resulting overall Z score provides the level of significance for the combined statistical test. (See Rosenthal, 1978, for a detailed explanation and computational examples of other, conceptually similar techniques.)

A strength of the combined significance tests when conditions for their use can be met is that they generally accomplish the goal of increasing power. (Rosenthal added the caveat that the studies should have tested the same directional hypothesis.) We can illustrate this approach by assuming that curriculum A is more effective than curriculum B but that the true difference for large populations is small. If A and B are repeatedly compared using small samples, one would expect to find, on the average, small differences favoring A. But many of the differences would not be statistically significant. An informal review of this research might conclude that the effect is not statistically reliable or that the plurality of studies find no difference at all. However, if the studies are combined (for example, by adding Z scores) the overall statistical test is much more likely to be significant.

In general, techniques for conducting a combined significance test seem most useful when the separate studies can be considered independent and essentially random samples, estimating a single "true" difference between populations, so that variation among study outcomes is attributable to chance. In this case, when the treatments are in fact differentially effective, an overall comparison will often detect this difference because it increases the effective sample size used in the test. When the variation among outcomes of different studies cannot be attributed simply to random variation, however, the combined significance test is less useful. The overall test will still provide an "answer" as to whether or not the common null

hypothesis should be rejected, but a single answer may not be a useful representation of reality.

A key point is that since many separate studies are combined into one "big test," its use should be preceded by efforts to determine if the variation in outcomes can be viewed as random. This is a crucial step. In cases where conflicts exist, an analyst may choose to use other techniques that are more sensitive to variation among study outcomes. (Bryant and Wortman, 1985; Bullock and Bvyantek, 1985; Kulik and Kulik, 1986, 1988; McGaw, 1988; White et al., 1986)

Special Problems of Quantitative Synthesis

This discussion has focused primarily on statistical procedures: computing effect sizes and conducting significance tests. The following three issues also come up in most quantitative evaluation syntheses.

Different Outcome Measures Across Studies

Combining studies is easiest when they all use the same outcome measure. But given the diverse priorities and resources of different researchers, such uniformity is extremely rare. Take day care as an example. Investigators have used various cognitive, physical, health, social, and emotional indexes to assess its effect on participating children (Belsky and Steinberg, 1978).

When outcome measures differ, the reviewer faces a dilemma. Is it reasonable to combine across seemingly different measures? The problem is not primarily a technical one. Whenever means and standard deviations are available, effect sizes can be computed and averaged. Whether or not to do so is a substantive question. The answer is ultimately dictated by good sense rather than any rote formula. The key issue is conceptual clarity. Suppose a review of day-care findings includes cognitive measures for 3-year-olds in some studies and emotional measures for 6-year-olds

in others. Then the reviewer must decide whether an overall quantitative summary will be useful and substantively sound. Just throwing together disparate measures because the title of each study contains the term "day care" can be foolish, no matter how statistically elegant or precise the review. (Anderson et al., 1983; Bayarri and Degroot, 1987; Bredderman, 1984; Fienberg et al., 1985; Hall et al., 1986; Himel et al., 1986; Steinkamp and Maehr, 1984; Willson, 1983)

Multiple Measures Within Studies

A second issue is how to treat studies that report more than one outcome. Take day care again. Suppose some studies compare day-care and home-reared children on both cognitive and social development with several measures of each, while other studies rely on only a single index. How should we balance their respective contributions in a review?

One way is to compute a separate effect size for each measure within each study. A study comparing children in day care to home-reared children on five different outcomes then contributes five effect sizes to the review. This approach disaggregates the unit of analysis to each comparison rather than to each study. It uses all available information. But perhaps an unintended consequence is that studies with multiple measures will be weighted more heavily than those with only one or two. Also, several comparisons within any study are not independent. They were done by one investigator on one group of participants. This could lead to repeated bias.

One solution is to categorize outcomes by what they measure—such as emotional, social, or cognitive abilities—and then conduct separate analyses for each subgroup. However, since many studies use more than one cognitive measure, or emotional measure, this might not always be sufficient.

A second solution treats each study as the unit of analysis and gives each study equal weight. It involves computing a "grand" effect for each study by averaging across the several measures (for example, Kulik and Kulik, 1982). This way, each study rather than each comparison gets one "vote" in the review. The trade-off here is loss of information within studies.

We recommend following and reporting both procedures. This will expand a final report. But since averaging within studies requires computing effects for individual comparisons anyway, presenting both analyses raises costs minimally. Doing both allows a reader to explore any differences between analyses. For instance, suppose a large average effect size emerges from a summary using each comparison as a unit of analysis. Then we can ask whether such findings depend unduly on one or two studies with multiple measures. (Bangert-Drowns, 1986; Becker and Hedges, 1984; Eysenck, 1984; Pillemer, 1984)

Missing Numbers

A quantitative review is impossible unless studies report the necessary statistical information. Data requirements for computing effect sizes are minimal. All we need are means and standard deviations or exact test statistics such as *t* and sample sizes. Yet it is surprising how often this information is unavailable. For example, one analyst recently looked at 24 studies of day care's effect on children's intellectual development. Over half did not report sufficient information for computing simple effect sizes.

What are a reviewer's options when confronted with missing or insufficient data? One is to try to obtain missing information directly from authors if time and resources permit. Since the statistics needed are quite basic—means and standard deviations—one would expect such efforts to be successful. The chance of success probably depends quite idiosyncratically upon the field, the investigators, and other factors

such as how dated the studies are. (Becker, 1986; Hedges, 1986; Tamir, 1985; Wolf, 1986)

A second strategy is to fill in conservative estimates of effect sizes when studies have missing data. Usually this means assigning effect sizes of zero. We do not know the treatment effect when statistics are missing. So if we plug in a zero, we are assuming minimum treatment effect. If, despite this policy, the review shows the treatment to be effective, we can be confident that this overall conclusion would not change, even if missing statistics were available.

This seemingly conservative strategy, however, is not always conservative. It depends upon your point of view. In some cases, such as research on the effect of day care or reduced cost reimbursement for hospitalization, finding no effect of the new program can be a happy outcome. We may not expect day care to raise IQ's or to make children happier; we are satisfied if it simply does no harm. We rarely expect reducing costs to improve health; the goal is to not do significant harm. In such cases, plugging in conservative statistics may bolster such an optimistic conclusion unjustifiably.

When effect sizes cannot be extracted from several studies, and when efforts to get this information directly from authors fails, it makes sense to focus quantitative analyses on the subgroup of studies with good information. Basing analyses on data that seem firm increases confidence in the review as a whole.

Table 4.1 illustrates the concepts so far. From the WIC report, it shows how the results of the WIC studies, finally used in the synthesis, can be combined to increase confidence in the findings.

Chapter 4
Performing the Synthesis

**Table 4.1: Mean
Birthweight Quantitative
Summary**

Study	Year and location
Kotelchuck	1978, Mass.
Metcoff	1980-82, Oklahoma City
Stockbauer	1979-81, Mo.
Silverman	1971-77, Allegheny County, Pa.
Bailey	1980, 2 Fla. counties
Kennedy	1973-78, Mass.
Summary	
Average	
Weighted average ^a	
Range	
Lowest	
Highest	

**Chapter 4
Performing the Synthesis**

Reported birthweight in grams ^a		Quantitative Indicators		
WIC	Non-Wic	Raw difference	% difference ^b	Statistically significant
3,281 (4,126)	3,260 (4,126)	21.0	0.6	Marginally
3,254 (238)	3,263 (172)	91.0 ^c	2.9	Yes
3,254 (6,657)	3,238 (6,657)	16.0	0.5	Yes
3,189 (1,047)	3,095 (1,361)	94.0	3.0	Yes
3,229 (37)	3,276 (42)	-47.0	-1.4	No
3,261.4 (897)	3,138.9 (400)	122.5	3.9	Yes
3,244.7	3,195.1	49.6	1.55 ^d	
3,257.8	3,225.9	31.3	0.97 ^d	
3,189.0	3,095.1	47.0	-1.4	
3,281.0	3,276.0	122.5	3.9	

^aThe numbers in parentheses are sample sizes.

^bRaw difference divided by non-WIC birthweight.

^cAdjusted.

^dAverage raw difference divided by average non-WIC birthweight.

^eEach mean is weighted by the number of participants or controls in its group and an overall average is obtained by dividing by the total number of participants or controls in the six studies. The raw difference is based on the total of participants or controls.

Source: U.S. General Accounting Office, WIC Evaluations Provide Some Favorable but No Conclusive Evidence on the Effects Expected for the Special Supplemental Program for Women, Infants, and Children, GAO/PEMD-84-4 (Washington, D.C.: January 30, 1984), p. 16.

Nonquantitative Approaches in Evaluation Synthesis

Many evaluation studies do not meet the assumptions or contain sufficient information to allow the use of the statistical approaches described above. Case studies and other kinds of information have often been available for synthesis. There are at least five types of information valuable to evaluation synthesis for which the statistical approaches described above are not applicable. The discussion that follows details these five types of information, describes general situations in which this information should be synthesized, and outlines some guidelines for incorporating such information.

Five Types of Information

The five types of information potentially valuable for the evaluation synthesis that are not suitable for statistical analysis are (1) single case designs, (2) nonquantitative aggregate studies, (3) nonquantitative information in quantitative studies, (4) expert judgments, and (5) narrative reviews of collections of research studies. We will review each type of information in turn.

Single Case Design

Detailed studies of single cases are common, and techniques for analyzing such information have been developed (Herson and Barlow, 1976; Kratochwill, 1977, 1978). Observations of single individuals have contributed heavily to the theories of Freud, Piaget, and Skinner—among the most influential psychologists of modern times. Dukes (1965) and Herson and Barlow (1976) presented many examples of “N = 1” research in psychology. Case studies are also frequently used in public policy analysis to examine the effects of nonexperimental events such as political decisions by cities and towns (Yin and Heald, 1975).

The term “case study” can refer to the study of a single event or desegregated studies of multiple events (Kennedy, 1979). Even if a case study uses a quantitative outcome, it is not possible to compute an

effect size in the traditional manner. If each individual is viewed as a separate study, there is no direct measure of within-group variation and no control group. Many of the studies used in the GAO synthesis on special education were case studies of local school districts. (Curtis and Shaver, 1987; Salzberg et al., 1987; Sampson et al., 1987; Scruggs et al., 1987; Slavin, 1986; Strube et al., 1985)

**Nonquantitative
Aggregate Studies**

Some research areas have important outcomes that are difficult to measure objectively or numerically. A clinical psychologist may report that obese people usually show general life improvements after weight loss or that hypnosis is effective in helping cancer patients adjust to chemotherapy. While an implicit baseline must exist, the benefits may not have been assessed with objective tests. In fact, an investigator may believe that the psychological effects of weight loss or hypnosis cannot be accurately assessed with a simple numerical measurement. A reviewer of such studies may still want to include these nonquantitative insights.

As Zimiles pointed out, this problem is particularly common in evaluations of complex programs:

"Most programs for children, especially educational programs, are aimed at producing a multiplicity of outcomes. As already noted, many of the psychological characteristics they are concerned with fostering—whether it be ego strength, or resourcefulness, or problem solving ability—are difficult or impossible to measure, especially within the time and cost constraints of an evaluation study. The usual response to this dilemma is to sift through the roster of multiple outcomes and single out for assessment, not the most important ones, but those that are capable of being measured" (Zimiles, 1979, p. 7).

Here an evaluator is faced with a trade-off between precision and meaning. Organizing a synthesis forces evaluators to confront a similar dilemma. Which outcomes appearing in the studies should be included

in a synthesis? If they decide not to rely exclusively on quantitative measures, they must figure out how to incorporate nonquantitative evidence.

A related situation occurs when quantitative studies do not contain sufficient information for statistical synthesis. For example, weak experimental designs may include a quantitative assessment. The reading performance of a group of children may be assessed with a standardized test following a special tutoring session. But without a comparison group, an effect size cannot be computed. Other studies compare a treatment group to a control but do not report sufficient information for producing a statistical summary.

Many of the studies included in various GAO syntheses fall into this category. For example, in the block grant report, administrative costs were calculated before and after program consolidation, but the calculation of comprehensive and reliable estimates of effect was hindered by differing definitions of administrative activities and other accounting procedures, inadequacy in data collection procedures, and weakness in sampling. These characteristics of the studies led to a choice of either omitting them or treating them in some nonquantitative manner. (Becker and Hedges, 1984; Carlberg and Walberg, 1984; Carlberg et al., 1984; Center et al., 1986; Slavin, 1987)

**Nonquantitative
Information in
Quantitative Studies**

In preparing a study report, researchers and evaluators do not simply list numerical results. The treatment and participants are carefully described; caveats or limitations are painstakingly laid out. Often the effort put into these nonquantitative descriptions far surpasses that involving the numerical information. It is not always either appropriate or desirable to reduce a study to one or several numerical indexes. Numbers may not accurately be interpreted without taking into account factors such as subject attrition, changes in

study procedure, and a variety of unexpected or otherwise notable happenings that become major study limitations. Most evaluators will need to include information in the evaluation synthesis that goes beyond numerical outcomes. (Chipman, 1988; Chow, 1988; Mullen and Rosenthal, 1985; Stanley, 1987)

Expert Judgment

An evaluator may choose to include expert opinion at early stages of the synthesis, such as in evaluating individual studies. Instead, the evaluator may want to systematically compare studies relying on expert judgments about program effectiveness. Syntheses should be able to incorporate these inputs.

Narrative Reviews of Collections of Research Studies

As Cook and Leviton (1980) have pointed out, a careful narrative review, explicit about its analytic procedures, can be extremely valuable. Narrative reviews of collections of research studies may frequently, for example, identify methodological weaknesses of certain broad types or groups of studies in a particular topic area. The evaluator will need to consider these points in deciding whether or not to include these studies in the synthesis and, if they are included, in interpreting findings from these studies. (Noblit and Hare, 1988; Wachter, 1988; White, 1987)

Indications of the Need for Nonquantitative Approaches

Under special circumstances, nonquantitative approaches to the evaluation synthesis are particularly appropriate. Four are when (1) treatments may be individual or more concerned with process than outcomes, (2) program effects are assessed across multiple levels of effect, (3) uncontrolled treatment groups are compared with the treated control group, and (4) the "wrong" treatment is studied.

1. Treatments may be individualized and focused on process objectives. Some educational and social

programs are tailored idiosyncratically to the person or community receiving services (Yin and Heald, 1975). Such treatment variations do not result from haphazard implementation. Rather, there is an intentional effort to individualize.

An example is the Education for All Handicapped Children Act (Public Law 94-142), passed by the Congress in the mid-1970's. The act requires that every handicapped child receive an appropriate, or individualized, program of special education and related services. It covers many handicaps, including physical, cognitive, and emotional handicaps, and so the services provided are extremely diverse and specialized. The desired outcomes vary as much as the treatments, both within and across handicapping conditions. That is, the desired outcomes and treatments might vary as much for two partially deaf children as they would for a partially deaf child and an emotionally disturbed child. Additionally, treatment lengths are individually determined.

Nonquantitative information is important in that the act stresses the process aspects of each treatment rather than the outcomes. The handicapped child's parents, for example, are to receive notice of a proposed change in their child's educational program, they are to be provided the opportunity to participate in the program, and the child's treatment and treatment outcomes are to be reviewed at least once a year.

Thus, aggregated and later synthesized child outcome data would be of little use to a policymaker who wants to know if Public Law 94-142 is working well on the whole and how it should be changed. A variety of descriptive data from various sources would be more useful. For example, descriptions of the quality of parent and school interaction might be helpful. (Guzzo and Katzell, 1987; Jackson, 1980; Levin, 1987; Walberg, 1986; Ward et al., 1987)

2. Assessing program effects across multiple levels of effect. Quantitative approaches can be employed when all the studies have assessed program effects at the same "level" or unit of effect. This level is often the individual participant. For example, most day-care studies examine the behavior of participating children. But programs can have an effect at other levels as well (Yin and Heald, 1975). With day care, for example, its availability can influence families and the labor market as well as children (Belsky and Steinberg, 1978).

If a program's influence is felt at several levels, an overall decision about it may force the aggregation of results across the different levels as well as across outcomes measured at the same level. While synthesis at any particular level can profit from quantitative methods (when the assumptions for using such methods are met and it is feasible to use them), the aggregation across levels usually demands many qualitative decisions about trade-offs.

3. Uncontrolled treatment groups and treated control groups. Salter (1980) has pointed out that when several studies compare people who receive a treatment to others who do not, subtle differences between similarly labeled treatments are common. Nonquantitative information can offer valuable guidance in helping a reviewer decide how similar the treatments are.

An example of this comes from a study by Fosburg et al. (1981). They reviewed a series of studies of a children's nutrition program sponsored by the U.S. Department of Agriculture. The simplest quantitative analysis would have involved computing an effect size for each study comparing the health of children who received food supplements with those who did not and then averaging findings across the studies. But nonquantitative information included in many of the individual studies convinced them this would be fruitless. While for administrative purposes the

treatment was the same in each study, information about "plate waste" (food not eaten) of the supplementary food suggested important differences among sites. In some cases, the plate waste was high; other studies reported almost none. In every case, these data were informal and descriptive. But the reviewers decided they were crucial. Combining treatments that had the same administrative name, in this setting, would have amounted in fact to combining groups receiving vastly different treatments. They were "uncontrolled."

The same dilemma arose for the control groups. They were not all "pure" control groups, in textbook fashion. Many studies reported that children at sites not receiving assistance from the U.S. Department of Agriculture, rather than receiving nothing at all, were getting some food assistance under title XX of the Social Security Act. This title provides various forms of aid to low-income families. So control groups in some of the studies in the review were actually quite heavily "treated," while others were in fact "pure" control groups, receiving no food assistance at all.

In this case, the qualitative descriptions of what actually happened to children in treatment and control groups in each study led the analysts to reorganize their synthesis into subgroups. These subgroups recognized differences between treated versus untreated controls. A simple effect size averaging over all available studies would have missed this step. (Bangert-Drowns, 1986; Becker, 1986; Begg, 1985; Cooper, 1982, 1988)

4. Studying the "wrong" treatment. Occasionally, when synthesizing outcomes, in cases in which quantitative approaches have proved feasible, one finds that a relationship between a program and an outcome is not as strong as was the originally planned treatment that might explain the differential success. Here,

descriptive or nonquantitative data can play an important role.

A quantitative analysis can systematically examine, across many research studies, the relationship between planned program and outcome variables. But descriptive information in one or several studies can give a clue to an evaluator that there exists a different feature of the treatment, one not formally built into a study's experimental design, that may be more important than the original planned treatment.

How Nonquantitative Information Can Influence Policy

A major impetus for developing quantitative synthesis methods was a wish to make research findings more useful for policy. When presented with a simple numerical summary of the average effect of psychotherapy (Smith and Glass, 1977) or personalized instruction (Kulik, Kulik, and Cohen, 1979) or class size (Glass and Smith, 1979), a policymaker can evaluate program effects without wading through volumes of research reports or vague rhetoric.

The "best" format for presenting research findings remains an open and complicated question. But there are cases in which qualitative findings have had a clear effect on policy. One example of how qualitative information led to actual administrative changes comes from studies of the comparative effectiveness of professional versus paraprofessional "helpers." Durlak conducted a systematic review of 42 comparative studies. He reported consistent findings across different patient populations that for certain clinical services "paraprofessionals achieve clinical outcomes equal to or significantly better than those obtained by professionals" (1979, p. 80).

This is not the sort of finding that many physicians expect when they review the literature on the effectiveness of nurse practitioners, yet it led to a practical outcome. Lewis et al. (1974) and

Merenstein, Wolfe, and Barker (1974), looked earlier for qualitative information about why the nurse practitioners seemed to be so effective. A key observation was that nurses allocated their time among patients differently from physicians. The two groups also gave different weighting to the importance of various symptoms and incidents. The result of these qualitative findings was that physicians made adjustments in their time allocations. It is interesting to trace the sequence of events here. Because of the quantitative information underlying the original comparative studies, the physicians viewed them as surprising but took them seriously as scientific evidence. This willingness to accept surprising findings led to a qualitative search for an explanation and, ultimately, to adjustments in the way some physicians allocate their time and resources. (Cooper, 1989; Hazelrigg et al., 1987; Light and Pillemer, 1982, 1984; Slavin, 1984; Smith, 1980; Strube and Hartman, 1983)

Presenting the Findings

The information generated through the evaluation synthesis process is brought together in a report that must be carefully formatted to respond to the questions that were formulated in conjunction with the study's requester. The introductory chapter should briefly describe the history of both the study and the particular program under discussion and should present the study objectives, scope, and methodology.

The latter section might include a framework showing the evaluation studies and data bases, a table showing the relationship between the evaluation questions and the available studies, and a description of the analytic steps undertaken. At a minimum, however, this section should describe the search to identify the evaluation studies, including any limits that were put on the search (such as a requirement that all studies have experimental designs). The section should answer the following types of question: How was the information obtained? From what sources? What limits, if any,

were put on the effort? How confident are the investigators that all relevant information, or a representative sample of that information, was obtained?

If possible, other report chapters should correspond to the client's questions. The body of the report of course includes discussion of the adequacy of the data available for response to a particular question. A technical appendix might systematically describe each study across such dimensions as title, report reference, study purpose, data collection period, sample selection, data collection, and data analysis. Data bases should also be described, although not all the same dimensions will be appropriate.

For several reasons, caution should be exercised in drawing conclusions from the synthesized data and in formulating recommendations. The evaluation synthesis cannot substitute for a carefully designed study with primary data collection for investigating the question of interest. Sources for the evaluation synthesis may be dated; additionally, all aspects of particular issues may not have been thoroughly explored. Confirmation from the agency administering the particular program under review may be needed to determine that the conclusions drawn from past studies are still applicable.

One of the most common concluding sentences in research reports is, "More research is needed." When is this statement based on a systematic assessment of available evidence, and when is it a casual remark that simply concludes an evaluation study? A synthesis can help answer this question. For example, the GAO synthesis of findings about the Special Supplemental Food Program for Women, Infants, and Children (discussed at the end of chapter 3) concluded that while the evidence that the program resulted in fewer low-birthweight babies was strong, there was no comparably convincing evidence as to its effect on children's mental retardation. Conclusions such as

Chapter 4
Performing the Synthesis

this can help policymakers understand exactly what is known, and exactly what is not known, about the problems they pose. Such conclusions also point to where good, new information would be particularly valuable.

Evaluation Synthesis Can Answer Questions a Single Study Cannot

What can a synthesis of evaluation studies do that a single study cannot? In this chapter, we discuss six issues that synthesis helps resolve. The most frequently cited virtue of synthesis is that the increased sample size can increase statistical power. This virtue has been discussed widely. However, the six properties of synthesis emphasized here have little to do with sample size. What they have in common is that they help us say when a social, medical, educational, or some other type of program works, not just whether or not it works on the average.

One way we can identify when a program works is by focusing on interaction. Statisticians often use this word to indicate nonlinearity. That is how we interpret the word here. In a program evaluation context, we can ask two questions. First, does the program work well for certain kinds of people and less well for others? Second, does the program work well in certain settings and less well in others? Both these questions are about interactions. A single study can find certain kinds of interactions, but synthesis of several studies can turn up much richer, more useful information. (Wachter and Straf, 1990; Yeaton, 1989)

Why Interaction Effects Are Important

Usually, social, educational, and health programs are evaluated to see how well they work. Good evaluations also examine how changes in program format could incrementally help improve them. One way of asking whether a program works is to ask whether it works on the average. Another way is to ask whether it works for a subgroup of people or in special settings.

For policy purposes, the interaction question can be as important as the main-effects question. For example, when a physician considers what anesthesia to give a patient prior to surgery and has a choice between two drugs, it is useful to learn which of the two is better on the average. However, it is even more

valuable to learn which of the two is preferable for the precise surgery the patient will have. Or which of the two has a better track record for the particular kind of patient. It would not be surprising to find, for example, that the anesthesia best suited for a 20-year-old in excellent general health is different from the anesthesia best suited for a 70-year-old in poor health.

Finding such interactions is important not only when making decisions for individuals but also when assessing the effectiveness of large-scale programs. Suppose that Head Start works generally well for children under 4 but far less well for children 5 years old or older. That would be worth knowing. If resources for the program were limited, such knowledge could tell us where to concentrate them. Or, if substantial resources were available, this finding of interaction would suggest that the Head Start curriculum should be modified for older children. So, whether the main purpose of an evaluation is to target resources or to change a program incrementally, finding an interaction can guide decisions. (Raudenbush and Bryk, 1985; Rosenthal and Rubin, 1986; Shapiro, 1985)

Why Synthesis Is Useful in Identifying Interactions

Let us recall how a single research study can identify an interaction effect. Basically, there are two ways. One way is to build a search for the interaction directly into the study design. For example, let us hypothesize that job training program A works better for high school dropouts than it does for high school graduates and that the reverse is true for training program B. Then, if we have control over treatment assignments, we can test this hypothesis by making sure that all four combinations of people and program type are represented. Ideally, randomization will be used to develop the four groups—dropouts given A, dropouts given B, graduates given A, and graduates given B. Then, comparisons of the four effect sizes will give a clear indication of what program type

works best, on the average, for what type of person. These findings will either refute or strengthen the initial hypothesis.

The other way of identifying interaction effects in a single study involves the use of post hoc procedures. Suppose that a search for interaction has not been formally designed into a study. In that case, such procedures as regression analysis and other applications of the general linear model can be applied retrospectively. The dilemmas and caveats involved in this process are well known. If people were free to choose their own treatment, there might be self-selection. There may be a confounding of background variables. For example, most of the high school dropouts may come from middle-income families in rural areas. Suppose that a single study of this type did not assign people to training programs at random. Then, because the study was not designed to examine interaction, its findings could well be confounded by graduation status, setting, and family type.

Against this background, we can now address the central theme of this chapter—that research synthesis can be far more effective in identifying interactions than any single study. Any one study is conducted in a particular context, under a particular set of constraints. Unless the study is extraordinarily large in scope, it has a limited group of participants who are assigned to treatments in a certain way. Each of these facts is good for a single study. It is important to know exactly what population is in and what population is out. It is important to know how people chose, or were assigned to, a treatment.

The advantage of looking at a group of evaluation studies is that the individual studies often take place in different contexts. And we can learn much about interactions from noticing how findings relate to context. To illustrate this idea concretely, we can now

address the six evaluation issues that synthesis can resolve better than any single study.

**Issue 1: Matching
Treatment Type With
Recipient Type**

The Head Start program was created in the early 1960's in response to a growing belief that something had to be done to help poor children start school on a stronger footing. In 1964, Sergeant Shriver, director of the Office of Economic Opportunity (OEO), formed a committee chaired by the pediatrician Robert Cooke. Its charge was to develop a program for reducing the effects of poverty on children. These efforts led to the creation of Head Start, which had seven concrete goals, including improving the child's mental processes and skills, with particular attention to conceptual and verbal skills.

The program was formally authorized to begin in summer 1965. Between 50,000 and 100,000 children were expected to participate in the first summer program. In fact, 560,000 did. By 1967, Head Start funding had grown to \$349 million. OEO decided to evaluate its performance and contracted in 1968 with Westinghouse Learning Corporation and Ohio State University to conduct a formal evaluation. The findings were released in 1969, and they stunned the education community.

The key sentence in the Westinghouse final report says: "Although this study indicates that full-year Head Start appears to be a more effective compensatory program than summer Head Start, its benefits cannot be described as satisfactory" (Cicirelli, 1969, p. 43). According to Datta,

"children who participated in Head Start summer programs did not score higher at the beginning of first, second, and third grades in such programs on all measures of academic achievement, linguistic development, and personal/social development than children who had not participated. Children who had attended the full-year programs and were tested in the first grade achieved higher scores on the Metropolitan Reading Test and some subtests of the Illinois

Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot

Test of Psycholinguistic Abilities. Scores of children who had attended full-year programs and were tested in the second and third grade were not different from the scores of comparison children" (1976, p. 134).

The disappointing findings of this evaluation generated great controversy. Smith and Bissell (1970), Cambell and Erlebacher (1970), and others criticized the methodology severely. Supporters of preschool education found many problems with the study's design and implementation. Yet, despite the criticism, the study had a great effect on policy. Supporters of Head Start were placed on the defensive. For example, both Alice Rivlin and Christopher Jencks, who supported such remedial programs as Head Start in the late 1960's, became more cautious after the Westinghouse-Ohio study. Rivlin noted that "Jencks and his associates dismiss the whole preschool child development movement in a few skeptical paragraphs, citing the Westinghouse-Ohio study's findings that, on the average, Head Start children showed no long-term cognitive gains over non-Head Start children" (1971, p. 32).

How should we interpret the findings of this single, large study, which had such a great effect? A synthesis of early education programs conducted by Bissell (1970) throws much light on Head Start and related preschool programs. Her review emphasized a search for interactions. Bissell reanalyzed data collected by three researchers: Karnes in Urbana, Illinois; DiLorenzo in New York state; and Weikart in Ypsilanti, Michigan. She chose these three data sets because each author compared two or more specific curriculums, each project had well-formulated goals, and each project was conducted and documented carefully.

Taken together, these three data sets compare five types of curriculum, each of which has supporters in the preschool community: the Karnes Ameliorative

curriculum, a highly structured cognitive curriculum; the Bereiter-Engelmann curriculum, a highly structured informational program; a traditional enrichment program emphasizing language development, with a relatively permissive low-structure environment; a traditional enrichment program emphasizing psychosocial development, with a relatively permissive low-structure environment; and a Montessori program with a structured environment.

Bissell found small main effects. For example, programs with strong quality control, well-trained staff, a high degree of staff supervision, and a low pupil-to-teacher ratio produce bigger cognitive gains than other programs. Her big finding involved interaction. To quote her: "Directive, highly structured preschool programs tend to be more effective with the more disadvantaged of poor children. . . . In contrast, nondirective, less-structured programs tend to be more effective with the less disadvantaged of poor children" (Bissell, 1970, p. 62).

Bissell's data make her point sharply. The reanalyses of scores on three standardized tests—the Binet, the Peabody Picture Vocabulary Test, and the Illinois Test of Psycholinguistic Abilities—show that when a child is well matched with the optimal program (for example, exceptionally down-and-out children and highly structured programs), the average difference between experimental and control groups is between two thirds and three quarters of a standard deviation. If the match is poor (as when down-and-out children from poor backgrounds are exposed to a relatively open curriculum), the comparative gains are minimal. A few of the comparisons even find a marginally negative program effect.

A synthesis such as Bissell's has at least three virtues. First, since the individual evaluations examined projects organized to serve different children in different places with different programs, we get a

Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot

broad panorama of findings. Second, since the data collected by several independent investigators display similar interaction patterns—that highly structured programs are best for the poorest children—the credibility of this overall finding is enhanced. Third, the synthesis of several evaluations puts the results of the single, big Westinghouse study in a new light. Most of the early Head Start sites, such as those examined by Westinghouse, had clearly open and permissive styles. They offered relatively little formal cognitive work. To quote Bissell again, “directors favor supportive, unstructured, socialization programs rather than structured informational programs for poor children” (1970, p. 81). Knowing this about the early Head Start centers that Westinghouse and Ohio State University examined and combining this fact with Bissell’s review findings, we can see why the study found little success. There is also reason for optimism that student performance should improve as more structure is introduced at local Head Start sites. (Katz et al., 1985; Levin et al., 1984; National Institute of Education, 1984; Proleau et al., 1983)

Issue 2: Explaining
Important Treatment

In 1968, Rosenthal and Jacobson wrote:

“As teaching training institutions begin to teach the possibility that teachers’ expectations of their pupils’ performance may serve as self-fulfilling prophecies, there may be new expectancy created. The new expectancy may be that children can learn more than had been believed possible, an expectation held by many educational theorists, though for quite different reasons” (p. 141).

Three years later, Baker and Crist asserted the opposite:

“Teacher expectancy probably does not affect pupil IQ. This conclusion is supported by a background of decades of research suggesting the stability of human intelligence and its resistance to alterations by environmental manipulation, by the reanalysis of the Rosenthal and Jacobson (1968) study . . . , and by the failure of all replication studies to demonstrate effects on IQ” (1971, p. 56).

Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot

So, here we find arguments from distinguished scholars that disagree sharply. The expectancy hypothesis is central to classroom conduct in education, because it has both substantive and ideological components. Suppose that teachers' expectations for a particular student's performance actually play a role in determining the student's performance. Some people see schools as exacerbating or even perpetuating inequality among children's life achievement. For these people, the expectancy argument offers a strong explanation for why poor children do less well in school than other children. Educators have vigorously debated the importance of teachers' expectations. Ryan (1971) and Kohl (1971) both argued that teachers expect less from poor children and therefore receive less. Elashoff and Snow (1971) argued the reverse—that methodological flaws in the study by Rosenthal and Jacobson (1968) undercut their findings.

To assess the importance of teacher expectancy on student IQ's, Raudenbush (1983) synthesized 18 such experimental studies. Seventeen of these studies had a strong research design, in which children were assigned at random to treatments. While the 18 studies included children of different ages and income groups, they all used IQ as an outcome measure. Raudenbush used several different methods for combining studies in a quantitative synthesis (Edgington, 1972; Fisher, 1973; Mosteller and Bush, 1954; Winer, 1971). He emphasized the effort to explain variation among outcomes (Pillemer and Light, 1980). His conclusion was not at all obvious for someone simply looking at the findings of 18 studies:

"The effect sizes of the studies, in standard deviation units, range from .55 down to -.13. Five of the eighteen achieved statistical significance, three at the .05 level and two at the .01 level. For the thirteen other studies, in five the experimental children scored higher than the controls, while in the other eight the control children scored higher" (Raudenbush, 1983).

Raudenbush's findings are clear and important. He found a small average effect size across the 18 studies of .11. But, as he reported, this main effect summary "certainly conceals more than it clarifies." That is because the studies can be divided into two broad groups. In one group, teachers were given information about each student (the "treatment") after a few weeks of initial contact. In the other group, teachers got information before they met students.

This difference between the two subgroups proved to be the key finding. Teachers who obtain information before they meet students show a strong expectancy effect. Teachers who obtain information after knowing students for several weeks show essentially no expectancy effect. To quantify this difference, the correlation between timing of the treatment induction and outcomes is $r = .68$. Raudenbush summarized: "When no teacher-student contact preceded the experiment, the average probability level was .06. After the second week of teacher-student contact, only one reported a probability of less than .05" (1984).

So, this synthesis sheds light on a controversy that has raged for many years. All these years, the debate focused on main effects: Does expectancy have a big effect or not? There seems to be a main effect, but it is very small. The synthesis tells us that the important treatment component lies in when the induction is given. It would be impossible to learn this from any one of the 18 studies alone. In fact, one major finding of this synthesis is the consistency of treatment effect in studies where there was no prior teacher-student contact. Similarly, outcomes of studies where expectancy induction took place 2 weeks or more into the school year show very little variance. Raudenbush found that the big news in these studies is that when the treatment is implemented matters a lot. Knowing this enables us to understand how teacher expectancy works.

Issue 3: Explaining
Conflicting Results

Most researchers and policymakers have at some point reviewed a group of studies to come up with "overall findings." A natural inclination in doing such a review is to hope that all or nearly all the outcomes will agree and that we can feel reasonably comfortable with these results. But it would be a shame if our natural hope to find agreement among study outcomes led us to view contradictory outcomes with frustration. Indeed, a major strength of data synthesis is that it helps us view conflicting outcomes in a constructive way. The conflicts may be offering valuable information. (Yeaton and Wortman, 1984)

In the late 1970's, discussions of job training emphasized the importance of "integrated services." Evaluations of the Comprehensive Employment and Training Act, the broad umbrella jobs program budgeted at several billion dollars a year, were finding marginal success at best. Some of these evaluations (National Academy of Sciences, 1978, 1979) suggested that job training alone, when narrowly defined, could not break a family's cycle of poverty and unemployment. These assessments found that integrated services, which included matching a family's needs for education, health services, and job training with a well-coordinated group of "helpers," offered far more promise than a stand-alone jobs program.

To assess this idea, the U.S. Department of Labor initiated several studies of integrated services programs. The key idea was to coordinate a series of services for poor families in which job training was an important component but not the only component. Several demonstration programs took place at several different sites. But the results were conflicting. While these conflicting findings about the value of integrated services were discouraging, the investigators ultimately capitalized on the varying outcomes to learn a great deal about the contexts in which integrated services worked well and worked badly and

about how to organize a good matching plan between services and recipients.

The broad question, then, is, How can a synthesis harness different findings from several studies to enhance our understanding about a program's effectiveness? It is not rare in an effort to pull together information about a program's effects across studies to find that the studies provide severely conflicting information. These conflicts can be frustrating, but it is precisely such conflicts that may give evaluators some insights into the matching problem. The job training example shows this. Let us look at some data, rounded off for illustrative purposes. Weeks of employment is the outcome measure.

Take the case of two studies conducted in different states by different investigators. Each compared an integrated service program with a single-service job training program. The study in one state looked at 80 men. Forty receive one program and 40 received the other. This study found that the integrated services group had an average of 80 weeks of employment, while the single-service group had 70 weeks. Integrated services seem to be more effective. The study in the other state found precisely the opposite. It also examined 80 men. Again, 40 received one program, and 40 received the other. This second study found that the integrated services men worked only 60 weeks, while the single-service men in the comparison group worked an average of 70 weeks.

What can be done with conflicting results? An effort can be made to see if we can discover from the conflict something about matching people's needs to the services that are offered in an integrated plan. Here is how synthesis could explain the conflict: categorize the 80 men in each study by their "problem" constellation. Such categorization can be difficult when the people served have multiple problems, but let us simplify here and assume that there are two

Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot

broad types of problem sets, problem set A and problem set B. First, look at the allocation of people with each type of problem to each program in the two studies: then separate that allocation from the average effect found for the people in each program.

A simple table might show that although the grand means of the two studies caused the results to conflict, the two studies had identical effects for the groups of men receiving each treatment. Both studies found that men with problem set A who received integrated services were employed an average of 90 weeks, while men with problem set A who received the single service worked an average of only 50 weeks. Both studies found that the reverse was true for men with problem set B. Why, then, was there a conflict? The conflict was caused by different allocations of problem type across the two service types in the two studies. In the first study, more A men than B men received integrated services. In the second study, the reverse was true. This difference in allocations, combined with the consistent finding of both studies that integrated services were more effective for men with problem set A and less effective for men with problem set B, created the conflict.

What policy implications can be drawn from this synthesis? If these data are in fact a good description of reality, we learn how integrated services should be targeted to a subgroup of people whose needs best match those services. The inkling that this might be the case emerged from an observation that two studies comparing integrated with single services reached opposite findings. It would not have emerged from either of the studies alone. We learn here that by examining program effects and allocations of people across studies, we can improve the matching process and target integrated services to those who will benefit the most. (Chelmsky, 1983; Cook, 1984; Eagly and Crowley, 1986)

**Issue 4: Determining
Whether Relative or
Absolute Performance
Is the Critical
Outcome**

Most programs can be looked at in at least two different ways. One way is to see whether an intervention has taken hold as intended. For example, does the child really know how to count better? Does the drug for hypertension actually lower the patient's blood pressure? Is the prisoner who is about to be released actually a competent carpenter? The other way of assessing a program is to see what happens in the end. Does the child who now counts better get higher marks in school? Do the patients who now have lower blood pressure also have a lower incidence of heart attacks? Does the newly released carpenter earn a reasonable income with the new skills?

A synthesis of evaluation findings that is well done usually looks at the studies that it examines in both ways. But it is worth noticing that synthesis has a special comparative advantage over any single study in answering the second question. This is because, while some programs can confer benefits or inculcate skills that indeed take hold, it is not always the case that these skills or benefits can be translated into concrete positive outcomes in the end. In particular, some benefits are valuable only in a comparative sense, because of the limited number of opportunities in which certain skills are useful. If too many others have the same skills, they become less valuable to any particular person who has them.

Job training can illustrate both points—the point about the comparative benefits and the point about the special value of synthesis. Suppose that a job skills program undertaken in one city is evaluated. The research design is excellent. One hundred applicants are divided at random into two groups. One group receives training to be carpenters, while the other group does not. If this evaluation finds that 2 years after the job program the trainees clearly earn more than the control group, what can we conclude? We would probably conclude that the job training works—and well we should, since the one available

Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot

study has positive findings, and they come from a randomized experimental design allowing causal inferences.

Let us assume that this positive finding is noticed and that the same program is offered at 10 other sites around the country. Learning from the excellent example set in the initial evaluation, each of the 10 new sites organizes its own randomized trial. The results become available 2 years later, and they are difficult to interpret. At three sites, the training is a clear success; the trainees have good jobs. At two sites, it is at best marginal; only some trainees have jobs. At the five other sites, it didn't work at all.

Efforts to organize these 10 findings into an evaluation synthesis can move forward in two quite different ways. One way is to emphasize the skills question: Did the trainees at all 10 sites become reasonably good carpenters? The other way is to emphasize the outcome question: Why did the findings differ so much across the 10 sites? By tackling both questions, synthesis can generate valuable insights. For example, a finding that sites varied enormously in their trainees' knowledge of carpentry provides management information. Clearly, the substantive training component needs to be improved across sites, and it needs to be strengthened in certain weak places. However, a finding that trainees learned carpentry quite well at all 10 sites would be even more informative, because it would force us to ask why trainees differed so substantially across the 10 sites in their ability to get jobs.

One possible explanation is that the benefit of the carpentry training for any one recipient depends on the number of other people who receive the training. Synthesis could support this explanation by examining the correlation across sites between the fraction of trainees who got jobs and the opportunity for success as measured by, say, the total population at each site. If the correlation is clearly positive, we

Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot

learn three things. First, we learn that training works in a predictable way. Trainees in bigger cities have better prospects than trainees in smaller towns. Second, we learn two important things about the training itself, that it indeed confers skills on participants and that when relative performance is not a constraint on any one trainee—that is, when the trainee lives in a big city—the program succeeds. Third, we learn how to organize and manage such training programs better in the future: they are best targeted to settings in which there are opportunities for trainees to put their training to use.

To summarize, then, synthesis can identify programs whose value depends not only on their substantive features but also on the number of people who participate in them. That is, synthesis can point out when programs are constrained by limited opportunities for success. A single study cannot answer questions of this nature.

Issue 5: Assessing the
Stability of Treatment
Effectiveness

Usually, any single study is organized by a single investigator or a small group of investigators, and it takes place in one or a very few sites. A single study at a single site allows us to assess whether a treatment worked overall. We can even examine the variance among outcomes at several sites. But we cannot tell how robust the treatment is when it is provided by several different investigators or organizations. Only a synthesis of results across several studies allows us to answer this question. When each of several organizations implements the same program in different places, the variation in outcomes offers a good signal of the program's robustness. If it works extraordinarily well in a few places and poorly in others, we must try to explain why. But, whatever the explanation, we will have discovered that the program is not robust. We learn that it is sometimes effective but that its strength is easy to undercut. At some sites, the poor performance may be explained by weak

implementation or by a poor match between recipients and program. The only way to assess the stability of a program in different settings is to see how it functions in different settings.

Issue 6: Assessing the
Importance of
Research Design

Some scholars spend a large fraction of their time arguing that research design matters. Gilbert et al. (1975), Chalmers (1981), and Hoaglin et al. (1985) worked hard to convince the evaluation community that randomization is a crucial ingredient for evaluation, since it underlies our ability to make causal inferences. The efforts of these investigators made some headway. However, randomization is sometimes difficult or impossible, so we must turn to alternatives and do the best that we can despite their imperfections. Alternatives include case studies, quasi-experiments, observational designs, studies of management records, and computer simulations when appropriate (Hoaglin et al., 1985).

None of this is news. But when a researcher or policymaker faces a concrete problem, such as whether a certain nutrition program is effective or whether a new surgical procedure is worth using, any single study is almost certain to have a single research design. There are a few exceptions, but they are rare. The reader of this one study must then ask two questions. First, does the study stand well on its own merit: Is it well done? Second, does the research design introduce any constraints, limitations, or biases? It is difficult to answer this second question with evidence from only one study, even one that has been well designed and executed. But a synthesis helps us a lot. It allows us to compare findings—and the research designs that lead to those findings—across a group of studies. If there are correlations, we learn two things. First, we see what specific design led to what specific outcomes. Second, we can organize future research knowing more about the consequences of specific designs.

**Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot**

Several concrete illustrations show how research design can matter. One example comes from surgery. Chalmers (1982) reviewed the findings from 95 studies of portacaval shunt surgery. These 95 reports were published over a period of many years by different investigators who worked at different hospitals. Chalmers asked two questions about each study. First, did its research design have adequate controls, poor controls, or no controls at all? Second, what did the investigator say about the surgery: Was there marked enthusiasm, moderate enthusiasm, or no enthusiasm? The conclusion is that poorly controlled studies of this surgery are far more likely than well-controlled studies to lead to positive results, perhaps illustrating Hugo Muench's law of clinical studies: results can always be improved by omitting controls (Bearman, Loewenson, and Gullen, 1974). (Something to be guarded against.)

A second example comes from the dilemma of how best to control spiraling health care costs. In 1982, the Committee on Labor and Human Resources of the U.S. Senate asked GAO to examine all available evidence about the effects on medical costs of increasing the amount of health care provided at home for elderly citizens. It was proposed in Senate debate that if more health care were provided at home, total service costs would drop, because the chronically ill would make less use of hospitals. GAO's findings were striking. The case studies, mostly small-sample narrative reports, almost unanimously suggested that costs would decline. But the quantitative studies found the opposite: total costs would not decline and, indeed, they might even increase slightly.

The quantitative studies turned up a clear reason for this surprising result. Rather than leading some elderly recipients of service to change the site from hospital to home, the new opportunity for home care considerably expanded the total number of people requesting care. People not receiving services began

Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot

to request them. So, while the offering of reimbursable home care as an alternative to hospitalization can reduce the cost per recipient for those who accept the alternative, it seems also to create a substantial new group of service recipients, and total service costs do not drop.

This example is not here to argue for or against home care. Some people argue that the home care option is a good idea even if costs are higher. Others disagree. But whatever one's values about the trade-offs between hospital and home care, the point is that a study's design is closely related to its outcome. An evaluator could examine every published case study and conclude that the evidence for lower costs with home care was overwhelming. Meanwhile, another evaluator who examined only studies with comparison or control groups would find overwhelming evidence in the other direction. Knowing that different types of studies generally lead to different sorts of findings offers guidance for the future. Whoever designs the next study to examine the costs of a home care program can try to incorporate the strengths of both types of design.

Summary

Research synthesis is not a remedy for all ills. Each effort faces dilemmas. Perhaps because certain value judgments must be made, such as the weight that must be placed on findings from different research designs, some investigators may be tempted to fall back on traditional narrative reviews. This would be a mistake. Just because a synthesis turns up conflict or requires a judgment call is not good reason to shoot the messenger. The messenger gives us information that is vital in two ways. First, synthesis points to the features of a treatment or program that seem to matter. Is there a crucial background variable? Does research design matter much? How stable are the findings across a group of studies? Second, synthesis helps us design the next study. Examining the first 10

Special Topics in Evaluation Synthesis

Once the evaluator has grasped the tools and techniques of evaluation synthesis, he or she is prepared for some of the finer points of the methodology. In this chapter, we discuss some of these, including (1) comparing and contrasting the studies and their findings, (2) merging the quantitative and nonquantitative approaches, (3) exploiting differences in study findings, and (4) anticipating problems.

Comparing and Contrasting Studies and Their Findings

The general GAO synthesis approach has been to compare and contrast the studies and their findings. In comparing the studies, we look for the nature and extent of similar findings or trends across them and try to rule out alternative explanations for their findings. The key questions asked are: What rules out placing support in similar findings across studies? What factors, if any, might increase our confidence in findings across the studies? To what extent can we place confidence in the findings?

In contrasting the studies, we focus on the exceptions and conflicts. We try to identify the study characteristics that might result in outcome variations. These may suggest tentative hypotheses for further investigation.

Begin with a review of the individual study, or study type, to identify the strengths and weaknesses that will affect confidence in the findings. If there is major weakness, low confidence in the individual study findings will, of course, be indicated. For example, the synthesis on home health care referred to earlier found that project evaluations using comparison groups experienced problems such as the presence of special populations, noncomparability of sites, and selection bias but that more confidence could be placed in studies with random assignment to groups. In evaluating the effectiveness of CETA, studies that considered only the postprogram experiences of CETA trainees without regard to participants'

Chapter 5
Evaluation Synthesis Can Answer
Questions a Single Study Cannot

studies and learning which program features are important and which are not help us develop an effective research plan for the eleventh study. Findings from a synthesis help make a study as powerful as possible in answering a specific policy question or resolving a policy dilemma. In a world of scarce resources, such targeting is valuable. While any one study is important, a great virtue of synthesis is that it makes systematic use of existing data and helps answer policy questions that single studies cannot answer.

preprogram experiences or without comparison groups were omitted from the synthesis.

Weak studies are not always omitted, however. For example, the synthesis on block grants examined administrative costs. All studies identified had many methodological problems. Rather than either place weight on any single estimate or take the position that no data were available, the evaluators examined the studies to see if any general patterns were discernible across the entire set of estimates. Given the weaknesses of the data, patterns were considered suggestive rather than definitive.

Even when studies are sound, issues such as generalizability may limit confidence in the applicability of the findings. Information available to address a particular question might come, for example, from several sound but small case studies. While the information is readily synthesized, confidence in generalizing from the findings would remain a problem.

Differences in findings across studies can sometimes be explained through the nonquantitative approach. For example, the special education synthesis showed large differences in two data sets in counts of handicapped children. Narrative analysis of the specific discrepancies in the efforts—including data collection methods, timing, and reporting content procedures—were shown as reasonable explanations for the differences in estimates.

While the findings across studies may be contradictory, they can also be complementary. In fact, findings from a study with a comparatively weak design may be reconsidered if they are consistent with those of other studies. For example, confidence in findings from a small case study may increase when they are similar to those of a more powerful study. Likewise, a series of independently conducted case studies consistent in their findings may yield a stronger vote

of confidence than would any study taken individually. Process evaluations are always helpful in interpreting the results of evaluations of effect.

In brief, the nonquantitative approach has generally required that we describe the characteristics, strengths, and weaknesses of the available sources of information. This requires analysis of individual studies and of studies taken as groups. It then dictates further analysis of similarities and differences in the findings of the studies.

Merging the Quantitative and Nonquantitative Approaches

Ideally, the nonquantitative approaches to evaluation synthesis should complement the quantitative approaches. Several of the types of information on quantitative studies illustrate how nonquantitative information can supplement the quantitative, when it is in fact feasible to implement quantitative approaches. In some situations, such as the uncontrolled treatment groups and treated control groups, the quantitative analysis would be, at best, misleading without the insights provided by the nonquantitative information.

Nonquantitative approaches to evaluation synthesis are especially helpful in dealing with conflicting findings among studies that have surfaced in a quantitative approach such as the blocking technique or cluster approach. Investigating conflict can sometimes reveal important information about programs that would not otherwise be available. The conflicts act as warning flags, suggesting that it may be useful to look at studies that show how a similar program was implemented at different sites or to examine variation across studies in relation to design characteristics and analysis strategies. From this perspective, variation among study findings uncovered through one approach to synthesis and investigated through another can be a useful, constructive, information-laden occurrence.

Three Strategies for Combining Quantitative and Nonquantitative Evidence

There are three broad strategies for using different kinds of information in the same synthesis: (1) putting nonquantitative information into a quantitative format, (2) discussing quantitative indexes narratively, and (3) using the two types of information in combination while maintaining the integrity of each one.

Quantifying Nonquantitative Information

One way to try to integrate qualitative and quantitative information is to translate the former into a numerical format. Here are three suggestions.

1. Case studies and nonquantitative aggregate studies. A first strategy here is to somehow summarize each case or aggregate with a numerical index and combine across studies. For example, outcomes of individual cases could be assigned values of +1 (successful), 0 (neutral), or -1 (unsuccessful), depending on a reviewer's overall evaluation of treatment success. This quantification can be done at a more detailed level by assigning numbers to several individual components of a case study and summing the ratings or by developing weights for different indicators of success (Laxarafeld and Robinson, 1940). This produces a single numerical index for each study, which can then be averaged or shown in a distribution.

The "case survey method" developed by Yin and Heald (1975) offers a more sophisticated way to quantify case studies. Each study is rated on several dimensions, such as research quality, program characteristics, and outcomes. These multiple ratings are cumulated across studies, providing an overall numerical summary. Scorers also indicate their level of confidence for each judgment, allowing reliability comparisons for "sure" and "unsure" ratings. A weakness of this "numbering" is that much rich descriptive detail is lost.

2. Qualitative information in quantitative studies.

Quantitative research studies usually report much information beyond statistical summaries. Most journals require authors to carefully describe the treatment, give information about participants and research settings, and discuss limitations and special features. A key insight often appears in the "discussion" rather than the "results" section of a research report.

Glass et al. (1981) suggest that all this "other" information be coded when possible and brought into the formal quantitative analysis. Walberg and Haertel (1980) present many specific reviews where background features are coded and statistically related to program effectiveness.

An advantage of this approach is that it helps us identify qualitative features of studies that are formally related to the quantitative outcomes we are testing. A reviewer often faces too many studies to conduct an efficient search for important qualitative information without statistical tools. This is especially true in the evaluation of educational innovations, where such relationships are often modest.

However, there is a familiar drawback. By quantifying study characteristics to facilitate statistical comparisons, we lose information and obscure important real-world differences. Similarly, it is hard to formally quantify idiosyncratic features that characterize a particular study, such as a report that the testing took place on a particularly hot day or that the children in a certain class had more opportunities for informal practice than children in another class.

3. Expert judgment. A third way to quantify narrative information is perhaps the most controversial, yet interesting. It involves an effort to incorporate into a review the wisdom of researchers and practitioners. Some people will have invested years of study and

thought and have had intimate experience with a program or curriculum. Scientists consult frequently with colleagues, and both researchers and practitioners are called upon routinely to provide expert testimony in policy matters. While no individual opinion can encompass all the detailed evidence from published literature, sometimes wisdom and fresh insight may transcend the "sterile" data of research reports.

We can suggest two ways of translating expert judgment into quantitative formats for use in synthesis. First, a reviewer can incorporate expert evaluation of studies prior to statistical integration. One way of doing this is to weight each study according to an expert's judgment of its overall value. Techniques already exist for weighting individual study outcomes by their sample size that can be adapted to experts' ratings (Rosenthal, 1978). This may serve to "formalize" what experts do when subjectively "weighing" the results of different studies to reach an overall conclusion. If an expert believes that a study provides especially strong evidence, the results from that study will receive extra weight.

Incorporating experts' judgments could enrich a review. For example, one can compare syntheses using weightings of different experts and also compare the various results incorporating weightings to a simple unweighted analysis. This would make explicit where experts disagree. If certain studies are rated positively by some experts and negatively by others, the discrepancies should be explored. Lack of agreement may pinpoint methodological, substantive, or ideological issues that lie at the core of controversy about an issue. When expert evaluations are consistent, we can be more confident about the innovation in question.

The use of weights ties expert evaluations to specific research studies. Our second suggestion involves

obtaining an expert's overall judgment about a specific issue, based on a global integration of his or her knowledge. Experts are often asked questions like "How big a risk does day care pose to an infant's emotional development?" or "Is reading program A really better than program B?" While it is possible to give a precise numerical answer to such questions, experts may prefer to supply judgments or assessments verbally (for example, it is "unlikely" that emotional development will be impeded, or it is "very possible" the new curriculum is better).

**Presenting Quantitative
Studies Narratively**

A second broad strategy works to do the reverse: take quantitative evidence and present it narratively. Rather than summarizing a series of results with numerical indexes, evaluators discuss studies individually. Strengths and weaknesses are identified and weighed, and overall conclusions are offered without precise quantitative documentation.

Critics of narrative reviews have described their characteristics as drawbacks. If studies are rigorous, precision is lost when a reviewer gives an approximate or impressionistic summary. However, certain purposes may be served by the discursive format. For example, narrative reviews may be more accessible to practitioners and policymakers who are unfamiliar with formal techniques and unwilling to rely solely on numerical indexes. When writing for a broad audience, an evaluator may choose to supplement effect sizes and significance tests with discussion of specific studies.

Narrative presentation may be especially useful when the purpose of the synthesis is not to summarize but, rather, to stimulate research or program improvements. Reviews often explore questions such as: How are studies designed? What are their major strengths and weaknesses? How easy or difficult was it to implement the treatment? Are there important but

“overlooked” program characteristics? Answering such questions gives newcomers to a field and nonspecialists a broad picture of “what the issues are.” It gives policymakers some ideas about strengths and weaknesses of “overall” findings and how confident one can be in adopting some of the suggestions. It may offer researchers important insights not only about how to interpret findings of existing studies but also about how to improve future efforts.

Allying Statistical and Descriptive Evidence

The two strategies above treat the synthesis process as one of translation. Words and numbers are different “languages.” So, for consistency, words are transformed into numbers or vice versa. However, both strategies have a crucial weakness. When one perspective is transformed into the other, its unique benefits are weakened. Statistical summaries lose their precision and the advantage of data reduction when transformed into narratives. Similarly, summarizing case descriptions with a simple numerical index loses much richness.

We think it is worth while for evaluators to work hard toward building an “alliance of evidence”: including both quantification and description within the same synthesis, while maintaining the integrity of each. Each type of information offers unique benefits. Similarly, rather than choosing between numbers and narrative when combining results across several studies, we need instead to determine where each is most useful and use them in synchrony.

A review can be not only primarily quantitative or descriptive but also strong or weak on both dimensions. Cook and Leviton (1980) put it well: the best synthesis makes the most out of both types of information.

**Three Benefits of
Combining
Information**

There are three ways numerical and descriptive information interact. They illustrate the benefits of working toward an alliance of evidence. We believe they show how the benefits of combining quantitative and descriptive studies outweigh the simplicity offered by an exclusive choice between them.

1. Using statistics to identify relationships not apparent from visual inspection. One view of formal quantitative methods is adversarial. Statistical significance is a dreaded hurdle that must be overcome before a study is considered "legitimate" and worthy of discussion. But some comparisons of statistical versus visual criteria for assessing change suggest that statistics are more often ally than adversary: by relying solely on visual inspection and subjective judgment, we are often likely to overlook small but reliable effects.

This view can be generalized to methods of combining studies. Cooper and Rosenthal (1980) had university faculty and graduate students summarize the results of seven investigations of sex difference in task persistence. Half of the reviewers were asked to "employ whatever criteria you would use if this exercise were being undertaken for a class term paper or a manuscript for publication" (p. 445), while the other half were instructed in how to use statistical combinatorial procedures. While several of the individual studies did not show significant sex differences, the statistical procedure demonstrated an overall significant effect favoring females ($p = .016$).

Descriptive reviewers were significantly more likely than statistical reviewers to find little or no support for the hypothesis of a sex difference in persistence. "Traditional reviewers either neglect probabilities or combine them intuitively in an overly conservative fashion" (Cooper and Rosenthal, 1980, p. 448). However, the statistical reviewers did not unquestioningly accept the hypothesis as "proven."

No one in either group of reviewers concluded that there "definitely" was support of the hypothesis. Furthermore, the type of reviewing procedure was not strongly related to recommendations for future research or to judgments about the methodological adequacy of studies. Statistical reviewers cautiously interpreted rather than blindly accepted numerical indexes.

These findings suggest that statistical procedures can help an evaluator identify relationships that may not be large enough to detect informally. Their worth should increase as the number of studies grows large or when a program effect is small. One might wonder why an evaluator should be excited about turning up positive but small effects. We can suggest two reasons. First, the limits on the degree of control that can be exerted over program participants in educational or medical innovations are likely to lead to small or incremental gains rather than "slam-bang" effects (Gilbert, Light, and Mosteller, 1975; Gottman and Glass, 1978). Second, when a small effect is detected, it sometimes can be enhanced by program refinement. This requires a judgment about whether a modest finding is worth pursuing. In such instances, process analysis and expert judgment become particularly important. This brings us to suggesting another way in which descriptive evidence can be allied with quantitative findings.

2. Using nonquantitative evidence after detecting a program effect. Statistical procedures can help both to identify small effects and to formalize the search for unusually successful or unsuccessful program outcomes or outliers. But such findings, standing alone, are not very informative. Suppose a reviewer looking at a dozen Head Start evaluations finds that, on the average, curriculum A slightly outperforms B, or that a review of 10 studies of urban high schools shows 1 to be unusually effective. What is one to make of these results? Formal procedures can detect subtle

differences but they cannot explain them. They offer a starting point, not a final answer.

After an effect is identified statistically, the reviewer must try to explain why this finding exists. Is it replicable? What program characteristics are responsible? Can it be enlarged or improved? Answering these questions requires further efforts that often rely heavily on case studies and descriptive evidence. Qualitative information may be necessary to explain the quantitative findings.

A more general point is that qualitative case descriptions are particularly valuable in helping program managers interpret statistical findings. Most managers are conscientious and want to strengthen their programs as much as possible. For them, it is especially useful to have descriptive data such as: What are the characteristics of successful implementations? How were the teachers trained? How were parents involved? What were details of the educational program? This information helps managers incrementally improve programs, using comparative findings from a review that gives insights about why certain versions of a program work better than others. Case study and other discursive information can help a manager at a "micro" program level, and at the same time it can inform "macro" divisions about program effectiveness, sometimes across hundreds of local sites.

3. Using the alliance to capitalize on conflicting outcomes. We have emphasized the value of using quantitative and descriptive studies as allies rather than adversaries for data synthesis. For example, in a review some years ago, the two different sorts of studies led to sharply contrasting findings but nonetheless illustrate our argument. In the 1940's, a group of educators and psychologists working with mentally retarded individuals came to believe that glutamic acid would improve a person's capacity to learn and

that this would be reflected by higher IQ scores. In the late 1940's and early 1950's, a series of uncontrolled studies and case reports appeared in the medical and psychological literatures, most of them finding a modest improvement in IQ's of retarded people receiving this drug (Kane, 1953; Levine, 1949; Zimmerman and Burgmeister, 1950).

These findings did not go unchallenged. Skeptics pointed out many threats to the validity of the studies and questioned how this drug could work physiologically to improve IQ. A series of controlled clinical trials were carried out to examine the effects of glutamic acid more systematically (McCulloch, 1950; Quinn and Durling, 1950; Zabrenko and Chambers, 1953). For example, McCulloch used matched experimental and control groups, with the controls receiving a placebo. Caretakers and examiners were not informed of subjects' group membership. Several of these experiments showed quickly and convincingly that glutamic acid did not outperform the placebo, although both groups showed an improvement over people receiving no treatment at all (the usual custodial care common in the 1940's). In 1960, Astin and Ross summarized the discrepant findings between case reports and experimental studies and concluded that the experimental evidence was far more convincing: glutamic acid was ineffective.

It is tempting to conclude from this example that the controlled, experimental, quantitative studies were "right" while the uncontrolled studies were "wrong" and that the latter served no useful scientific purpose. We come to a different conclusion: the conflicting results carry valuable information about how to improve the lives of retarded people. The controlled experiments are indeed convincing that glutamic acid does not raise IQ. But something was working in the patients' behalf, since most of the earlier case reports documented IQ gains. Scientists were pressed to account for the improvement.

Contrasting the controlled and uncontrolled studies prompts us to examine the context in which the drug was administered. Including the uncontrolled studies in a synthesis reveals an example of "studying the wrong treatment." People receiving glutamic acid also got environmental stimulation far beyond what was "usual." Increased attention and expectations also seemed to improve the performance of the "placebo group" in the experimental trials. One study (Zabrenko and Chambers, 1953) focused on the "environmental stimulation" hypothesis directly and confirmed its positive effect on IQ.

This example illustrates how different forms of evidence, taken together, can lead to insights with important policy implications. The seemingly inconsistent findings end up displaying information about both glutamic acid and supportive environments. Conflicts in outcomes have not hindered us. They have enriched educational practice.

The glutamic acid controversy occurred over 40 years ago, but the lesson still applies today. The theme is that different types of evidence may be complementary and that singlemindedness about either quantitative or qualitative approaches to synthesis imposes unnecessary limits on what we can learn from the work of others. The pursuit of good science should transcend personal preferences for numbers or narrative.

Exploiting Differences in Study Findings

To benefit from discrepancies among studies, whether uncovered through a quantitative or nonquantitative approach, we must repeatedly ask the question, What may explain the different findings? Trying to answer this forces a systematic inquiry that may or may not be quantitative. There are at least five specific ways to seek out and confirm explanations for conflicting findings.

Four were described in chapter 2: determine if similarly labeled treatments and programs differ in important ways, look for setting-by-treatment interactions, investigate different research designs used across studies, and examine different analysis strategies used in different studies. A fifth way is to relate background variables to findings.

One strategy for doing the latter involves coding information about participants' background characteristics and the design characteristics of the research (for example, the method of assigning subjects to groups) and relating this information to study findings. The work of Hall (1979) illustrates this synthesis strategy. She related several features of each study to the size of the effect of sex differences in decoding nonverbal cues. These features include both background characteristics of the participants and research design descriptions. For example, she found no relationship between participants' age and effect size, while the year in which the study was conducted turned out to be important (more recent studies tended to show the largest effects).

A second strategy follows Klitgaard's (1975) suggestion "to use the unusual as a guide to the usual," since "the unusually successful (or unsuccessful) may provide a clearer picture of processes operating to a lesser extent elsewhere" (p. 531). Comparing extremely successful programs to particularly unsuccessful ones may produce a list of other clear differences between them. For example, comparing a successful title I program to one that failed miserably may point out differences in staffing, expenditures, or curriculums.

With a few key explanatory factors identified, a policy analyst can form specific hypotheses about how they may influence findings. For instance, one might expect staff-to-child ratio to influence Head Start effectiveness, but there may also be complex

interactions between this variable and others, such as the amount of money spent per child or total number of children in the program. The hypotheses can be examined using data from less extreme studies. For example, if staff-to-child ratio in Head Start is universally important, there should be some evidence of this across the entire range of study outcomes. In fact, since public policies or regulations will often influence the "usual" more than the "unusual," this step can be critical.

A third strategy looks at what is "typical." Focusing on atypical programs should not deter an analyst from examining the major bulk of the studies for background features related to outcome differences. First, just because a study outcome falls in the middle of a distribution, this does not indicate that the program is typical. It is possible that a highly successful program or curriculum is paired with unusually needy participants, or poor resources, resulting in a mediocre final performance level. In these instances, an analyst would ideally want to adjust for some background factors before searching for effective or ineffective programs (see Klitgaard, 1975, for further discussion). A "typical" program may appear quite "atypical" after adjustments are made for background characteristics related to outcomes.

The examination of studies that have roughly "average" outcomes can be valuable in another way. Focusing on extremes puts our emphasis on identifying program or participant differences in order to explain divergent findings. But in large syntheses, involving many potential background variables, the other side of the coin is important as well. Examining studies with similar outcomes may be useful in identifying inoperative variables. For example, suppose that 10 Head Start programs produce relatively consistent results. Suppose also that while the program curriculums and participants are quite similar, the formal educational level of the teachers varies

dramatically across centers. This fact by itself would not prove that teacher education is unimportant, since it may interact with other measured or unmeasured variables. But it would strongly suggest that teacher education should not be our number one candidate for a variable that will explain outcome differences. Since there are usually enormous numbers of variables that we think might be important, this process of looking at "typical" outcomes can help limit the field for first-cut analysis and study designs.

Anticipating Problems That Might Emerge

Publication Bias

Publication bias results when not all studies of any drug or therapy are equally likely to appear in a refereed journal. A finding may be more likely to be published in a journal if it turns up as statistically significant. So if a synthesis includes only published studies, one might suspect a bias toward large, or statistically significant, effects.

Reviewers in education and psychology have found empirical support for this view. For example, Smith (1980) gathered groups of studies assessing innovations in education. She found empirically that average effect sizes were noticeably smaller for unpublished studies than for published studies. Greenwald (1975), also suspecting publication bias in psychological research, collected data from a group of referees for the *Journal of Personality and Social Psychology*. He surveyed authors who had recently published in that journal. He asked two key questions. First, "After an initial full-scale test of the focal hypothesis that allows rejection of the null hypothesis, what is the probability that you will submit the results for publication

immediately, before more data collection?" The mean response for reviewers was more than 40 percent, and for authors it was 58 percent. The second question was, "After an initial full-scale test of the focal hypothesis that does not allow rejection of the null hypothesis, what is the probability that you will submit the results before further data collection?" The mean response for reviewers was 5 percent, and for authors it was 6 percent. If these results are even roughly in the ballpark of reality, we see that a statistically significant finding is nearly 10 times more likely than a non-significant finding to be submitted for publication in a refereed journal.

This idea led Rosenthal (1979) to coin the term the "file drawer problem." His thought is that for every published study there may be several sitting in a researcher's file drawer, unsubmitted or unpublished because the researcher did not turn up statistically significant results. Ignoring this problem and looking only at published studies can lead an evaluator to overestimate a treatment effect, perhaps dramatically. (Lane and Dunlap, 1978; Orwin, 1983; Rosenthal, 1979; Shadish et al., 1987; Simes, 1987; Sommer, 1987)

Combining Results Across Different Treatments

Are the treatments given in different studies similar enough so that results can be combined in a sensible fashion? Answering this question is probably harder in social science research than in medicine, but it should be asked in drug trials nonetheless. In the field of job hunting, the National Academy of Sciences issues a report aggregating findings from many analyses of the broad job training program called YEDPA that trains unemployed youth. Their biggest finding is that the specific protocol for this training program varies enormously from site to site, despite the common "template" over the training workshop's door at each site. Boruch (1980) raises, in this academy report, the fundamental question of whether the results from

these several sites are combinable at all. (Chalmers et al., 1987; Cordray, 1990; Hoaglin et al., 1985; Louis et al., 1985; Sacks et al., 1985)

Examining Control Groups in Different Studies

Have control groups in different studies been examined for similarities and differences? This question applies specifically to comparative studies. When some studies show a treatment group outperforming the controls while others show no difference, the reviewer asks why. One possible explanation is that control groups in various studies are fundamentally different.

Control groups might differ simply by how different researchers define them. Some studies compare a new treatment to a "control" that is no treatment at all. Others compare a new treatment to a "control" that is an old or standard or existing treatment. Still others compare a new treatment to "controls" that are really alternative new treatments. In each of these circumstances, there is a clear comparison group but the group's fundamental purpose varies.

An example of these different definitions from the day-care literature comes from the work of Ruopp (1979). Ruopp examined many studies of a program called "developmental day care" for young children, as part of a project for the U.S. Department of Health and Human Services. By examining control groups in depth, this researcher found at least four different kinds: children cared for full time by a parent at home, children in nursery school, children in less costly care called "custodial day care," and children cared for in a private home by adults other than their parents. Simply aggregating findings across these four kinds of comparative studies did not make sense. The results turn out to depend heavily upon which kind of comparison group is used.

The fundamental point here, of course, is that aggregating across all available studies regardless of the form of the control group can dilute rather than strengthen the inferences from a research review. (Bailar et al., 1986; Begg and Berlin, 1988; Light and Pillemer, 1984; Wortman, 1984, 1985)

**The Evaluator's
Attitude Toward
Conflicting and
Discrepant Outcomes**

We have left this question for last because we believe it is the most important and yet the hardest to deal with concretely. It is astonishing how often evaluators are surprised that different studies of the same drug or treatment produce discrepant results. What do they expect? It would be remarkable if each of 30 independent studies evaluating a new drug for high blood pressure found that it brought pressure down by exactly 10 systolic units. Indeed, it would be more than remarkable—it would be suspect. Some chance variation among findings is expected.

Usually reviewers have the opposite problem. Many summaries flounder because individual studies give highly discrepant results. So a productive initial step in quantitative analysis is searching for orderly patterns of results. Probably the easiest way to do this is with a simple graph. Plotting study outcomes on the X axis and their frequency on the Y axis can offer surprisingly rich insights. Light and Pillemer (1984) describe a number of simple graphic procedures for examining variation among findings. Here we will mention only the briefest summary of inferences from a simple graph of study outcomes.

First, if treatments in several studies are really similar, the graph should be well-behaved. It should look approximately like a normal distribution, suggesting that differences among findings are basically the result of sampling error. If outcomes look grossly irregular, a reviewer must question whether all studies come from the same population. For example, a bimodal distribution would be a first indication that a

group of studies should not be combined in too facile a way—there might be two underlying populations. The challenge for a reviewer is then to identify the factors that divide studies into two groups.

Second, a graph should make outliers more noticeable. These extreme observations may or may not bother a reviewer, depending upon the purpose of the review. If its purpose is to identify a typical or central value, a few scattered outliers carry no special information. But if the reviewer's purpose is to spot the rare failure of a new drug, or an exceptionally successful circumstance for that drug, identifying outliers can be the most important part of the entire process.

To tie this back to the earlier discussion of what question drives the evaluation synthesis, outliers are especially important when a researcher is looking for subject-by-treatment interactions—say, certain types of illness in which specific drugs work especially well or especially poorly.

After finding outliers that seem important, the reviewer must look for explanations. Why must this have happened? Is it just a chance finding? Suppose a group of heart bypass surgery studies have a small cluster of particularly successful reports. Then the challenging question is whether they share any special feature. Perhaps the exceptionally successful studies all involve younger patients. Perhaps they were all done at large urban hospitals with exceptional facilities. There are usually many possible explanations: similarly labeled treatments or programs may differ in important ways, there may be setting-by-treatment interactions (that is, a program or treatment may be more or less effective depending on who participates in it, where it is administered, or some other situational factor), different studies may have been designed differently, and analysis strategies used in different studies may vary.

Chapter 6
Special Topics in Evaluation
Synthesis

Determining a convincing reason or reasons is a real challenge to the evaluator. This brings home the enormous value of successfully combining substantive and technical knowledge in syntheses. It is easy enough to graph outcomes and spot outliers. It is much harder to identify what features consistently distinguish the exceptional studies from the others. (Berlin et al., 1989; Light and Pillemer, 1984; Olkin, 1990; Toth and Horwitz, 1983; Yeaton and Wortman, 1984, 1985).

BLANK PAGE

Bibliography

Abrami, P. C., P. A. Cohen, and S. D'Apollonia. "Implementation Problems in Meta-analysis." Review of Educational Research, 58 (1988), 151-79.

Anderson, R. D., et al. "Science Education: A Meta-analysis of Major Questions." Journal of Research in Science Teaching, 20 (1983), 379-85.

Bailar, J. C., III, and F. Mosteller (eds.). Medical Uses of Statistics. Waltham, Mass.: New England Journal of Medicine Books, 1986.

Baker, P. J., and U. L. Crist. "Teacher Expectancies: A Review of the Literature." In R. E. Snow and J. D. Elashoff (eds.), Pygmalion Reconsidered. Worthington, Ohio: Jones Publishing Company, 1971.

Bangert-Drowns, R. L. "Review of Developments in Meta-analytic Method." Psychological Bulletin, 99 (1986), 388-99.

Bayarri, M. J., and M. DeGroot. "Bayesian Analysis of Selection Models." The Statistician, 7 (1987), 137-46.

Bearman, J. E., R. B. Loewenson, and W. H. Gullen. "Muench's Postulates, Laws, and Corollaries, or Biometricians' Views on Clinical Studies." Biometrics Notes, number 4. Bethesda, Md.; Office of Biometry and Epidemiology, National Eye Institute, National Institutes of Health, 1974.

Becker, B. J. "Influence Again: An Examination of Reviews and Studies of Gender Differences in Social Influence." The Psychology of Gender: Advances Through Meta-Analysis, 178-209. Baltimore, Md.: Johns Hopkins University Press, 1986.

Becker, B. J., and L. V. Hedges. "Meta-analysis of Cognitive Gender Differences: A Comment on an Analysis by Rosenthal and Rubin." Journal of Educational Psychology, 76 (1984), 583-87.

Bibliography

Begg, C. "A Measure to Aid in the Interpretation of Published Clinical Trials." Statistics in Medicine, 4 (1985), 1-9.

Begg, C., and J. A. Berlin. "Publication Bias: A Problem in Interpreting Medical Data (with Discussion)." Journal of the Royal Statistical Society, series A, 151 (1988), 419-63.

Belsky, J., and L. D. Steinberg. "The Effects of Day Care: A Critical Review." Child Development, 49 (1978), 929-49.

Berlin, J. A., C. B. Begg, and T. A. Louis. "An Assessment of Publication Bias Using a Sample of Published Clinical Trials." Journal of the American Statistical Association, 84 (1989), 381-92.

Bissell, J. W. "The Effects of Preschool Programs for Disadvantaged Children." Ph.D. dissertation, Harvard Graduate School of Education, Cambridge, Mass., June 1970.

Bornstein, R. F. "Exposure and Affect: Overview and Meta-analysis of Research 1968-1987." Psychological Bulletin, 106 (1989), 265-89.

Boruch, R. F. An Appraisal of Education Program Evaluations: Federal, State, and Local Agencies. Washington, D.C.: U.S. Department of Education, June 30, 1980.

Bowers, T. G., and G. A. Clum. "Relative Contribution of Specific and Nonspecific Treatment Effects: Meta-analysis of Placebo-controlled Behavior Therapy Research." Psychological Bulletin, 103 (1988), 315-23.

Bredderman, T. "The Influence of Activity Based Elementary Science Programs on Classroom Practices: A

Quantitative Synthesis." Journal of Research in Science Teaching, 21 (1984), 289-303.

Brown, R. "The Issue of Independence of Effect Sizes in Meta-Analyses." Presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 16-20, 1986.

Bryant, F. B., and P. M. Wortman. "Issues in Data Synthesis." New Directions for Program Evaluation, No. 24. San Francisco, Calif.: Jossey-Bass, 1984.

Bryant, F. B., and P. M. Wortmann. "Methodological Issues in the Meta-analysis of Quasi-experiments." Evaluation Studies Review Annual, 10 (1985), 629-48.

Bryk, A. S., and S. W. Raudenbush. "Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretations." Psychological Bulletin, 104 (1988), 396-404.

Bullock, R. J., and D. J. Svyantek. "Analyzing Meta-analysis: Potential Problems, an Unsuccessful Replication, and Evaluation Criteria." Journal of Applied Psychology, 70:1 (1985), 108-15.

Campbell, D. T., and R. Boruch. "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensating Education Tend to Underestimate Effects." In Evaluation and Experiment, C. A. Bennett and A. A. Lumsdaine (eds.). New York: Academic Press, 1975.

Carlberg, C. G., et al. "Meta-analysis in Education: A Reply to Slavin." Educational Researcher, 13:8 (1984), 16-27.

Carlberg, C. G., and H. J. Walberg. "Techniques of Research Synthesis." The Journal of Special Education, 18 (1984), 12-49.

Center, B. A., R. J. Skiba, and A. Casey. "A Methodology for the Quantitative Synthesis of Intra-subject Design Research." The Journal of Special Education, 19 (1985-86), 388-400.

Chalmers, T. C., et al. "A Methodology for Assessing the Quality of Randomized Control Trials." Controlled Clinical Trials, 2 (1981), 31-49.

Chalmers, T. C., et al. "Meta-analysis of Clinical Trials as a Scientific Discipline, II: Replicate Variability and Comparison of Studies That Agree and Disagree." Statistics in Medicine, 6 (1987), 733-44.

Chelimsky, E. "The Definition and Measurement of Evaluation Quality as a Management Tool." New Directions for Program Evaluation, No. 18. San Francisco, Calif.: Jossey-Bass, 1983.

Chipman, S. F. "Far Too Sexy a Topic." Review of The Psychology of Gender: Advances Through Meta-Analysis. Educational Researcher, 17:3 (1988), 46-49.

Chow, S. L. "Significance Test or Effect Size?" Psychological Bulletin, 103 (1988), 105-10.

Cicirelli, V. "The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development." Clearinghouse for Federal Scientific and Technical Information, Washington, D.C., 1969.

Colditz, G., J. Miller, and F. Mosteller. "The Effect of Study Design on Gain in Evaluation of New Treatments in Medicine and Surgery." Drug Information Journal, 22 (1988), 343-52.

Cook, T. D. "What Have Black Children Gained Academically from School Integration? Examination of the Meta-analytic Evidence." In "School

Desegregation and Black Achievement," T. Cook et al. (eds.). Unpublished report, National Institute of Education, Washington, D.C., 1984.

Cook, T. D., and L. C. Leviton. "Reviewing the Literature: A Comparison of Traditional Methods with Meta-Analysis." Journal of Personality, 48 (1980), 449-72.

Cooper, H. M. "Scientific Guidelines for Conducting Integrative Research Reviews." Review of Educational Research, 52:2 (1982), 291-302.

Cooper, H. M. "Literature Searching Strategies of Integrative Research Reviewers: A First Survey." Knowledge: Creation, Diffusion, Utilization, 8:2 (1986a), 372-83.

Cooper, H. M. "On the Social Psychology of Using Research Reviews: The Case of Desegregation and Black Achievement." In The Social Psychology of Education, pp. 341-63, R. Feldman (ed.). Cambridge, England: Cambridge University Press, 1986b.

Cooper, H. M. "Organizing Knowledge Synthesis: A Taxonomy of Literature Reviews." Knowledge in Society, 1 (1988), 104-26.

Cooper, H. M. Integrating Research: A Guide for Literature Reviews, 2nd ed. Newbury Park, Calif.: Sage, 1989.

Cooper, H. M., and R. Rosenthal. "Statistical Versus Traditional Procedures for Summarizing Research Findings." Psychological Bulletin, 87:3 (1980), 442-49.

Cordray, D. S. (ed.). "Utilizing Prior Research in Evaluation Planning." New Directions for Program Evaluation, No. 27. San Francisco, Calif.: Jossey-Bass, 1985.

Cordray, D. S. "Quasi-experimental Analysis: A Mixture of Methods and Judgment." New Directions for Program Evaluation, 31 (1986), 9-27.

Cordray, D. S. "Meta-analysis: An Assessment from the Policy Perspective." In The Future of Meta-Analysis, K. Wachter and M. Straf (eds.). New York: Russell Sage Foundation, 1990.

Curtis, C. K., and J. P. Shaver. "Modifying Attitudes Toward Persons with Disabilities: A Review of Reviews." International Journal of Special Education, 2:2 (1987), 103-29.

Datta, L. "The Impact of the Westinghouse/Ohio Evaluation on the Development of Project Head Start." In C. C. Abt (ed.), The Evaluation of Social Programs. Beverly Hills, Calif.: Sage Publications, 1976.

Dickersin, K., et al. "Publication Bias and Clinical Trials." Controlled Clinical Trials, 8 (1987), 343-53.

Dukes, W. F. "N = 1." Psychological Bulletin, 64 (1965), 74-79.

Durlak, J. A. "Comparative Effectiveness of Paraprofessional and Professional Helpers." Psychological Bulletin, 86 (1979), 80-92.

Dush, D. M., M. L. Hirt, and H. E. Schroeder. "Self-statement Modification in the Treatment of Child Behavior Disorders: A Meta-analysis." Psychological Bulletin, 106 (1989), 97-106.

Eagly, A. H., and M. Crowley. "Gender and Helping Behavior: A Meta-Analytic Review of the Social Psychological Literature." Psychological Bulletin, 100:3 (1986), 283-308.

Bibliography

Edgington, E. S. "An Additive Model for Combining Probability Values from Independent Studies." Journal of Psychology, 80 (1972), 351-63.

Eysenck, H. J. "An Exercise in Mega-Silliness." American Psychologist, 33 (1978), 517.

Eysenck, H. J. "Meta-analysis: An Abuse of Research Integration." The Journal of Special Education, 18 (1984), 97-106.

Feingold, A. "Matching for Attractiveness in Romantic Partners and Same-Sex Friends: A Meta-analysis and Theoretical Critique." Psychological Bulletin, 104 (1988), 226-35.

Feldman, K. A. "Using the Work of Others: Some Observations on Reviewing and Integrating." Sociology of Education, 44 (Winter 1971), 86-102.

Fineberg, S. E., M. E. Martin, and M. L. Straf (eds.). Sharing Research Data. Washington, D.C.: National Academy Press, 1985.

Fisher, R. A. Statistical Methods for Research Workers. London: Oliver and Boyd Publishing Co., 1973.

Fosburg, S., and F. Glantz. "Analysis Plan for the Child Care Food Program." Submitted to the Food Nutrition Service, U.S. Department of Agriculture, by Abt Associates, Inc., Cambridge, Mass., April 1981.

Gallo, P. S., Jr. "Meta-Analysis—A Mixed Meta-phor." American Psychologist, 33 (1978), 515-17.

Gilbert, J. P., R. J. Light, and F. Mosteller. "Assessing Social Innovations: An Empirical Base for Policy." In Evaluation and Experiment, C. A. Bennett and A. A. Lunsdaine (eds.). New York: Academic Press, 1975.

Glass, G. V. "Integrating Findings: The Meta-Analysis of Research." Review of Research in Education, 5 (1977), 351-79.

Glass, G. V. "Bibliography of Writings on the Integration of Research Findings." Manuscript, Laboratory of Educational Research, University of Colorado, Denver, Colo., 1978.

Glass, G. V., B. McGaw, and M. L. Smith. Meta-analysis in Research. Beverly Hills, Calif.: Sage Publications, 1981.

Glass, G. V., and M. L. Smith. "Primary, Secondary, and Meta-Analysis of Research." Educational Researcher, 5 (November 1976), 3-8.

Glass, G. V., and M. L. Smith. "Reply to Eysenck." American Psychologist, 33 (1978), 517.

Gottman, J. J., and G. V. Glass. "Analysis of Interrupted Time Series Experiments." In T. R. Kratochwill (ed.), Single Subject Research. New York: Academic Press, 1978.

Green, B. F., and J. A. Hall. "Quantitative Methods for Literature Reviews." Annual Review of Psychology, 35 (1984), 37-53.

Greenwald, A. G. "Consequences of Prejudice Against the Null Hypothesis." Psychological Bulletin, 85 (1975), 845-57.

Guzzo, R., S. Jackson, and R. Katzell. "Meta-analysis Analysis." In Research in Organizational Behavior, vol. 9, B. Staw and L. Cummings (eds.). Greenwich, Conn.: JAI, 1987.

Hall, B. W., A. W. Ward, and C. B. Comer. "Published Educational Research: An Empirical Study of Its Quality." Presented at the annual meeting of the

American Educational Research Association, San Francisco, Calif., April 16-20, 1986.

Hall, J. A. "Gender Effects in Decoding Nonverbal Cues." Psychological Bulletin, 85 (1978), 845-57.

Hall, J. A. Statistical Methods for Meta-Analysis. New York: Academic Press, 1985.

Hall, J. A., and I. Olkin. "Vote Counting Methods in Research Synthesis." Psychological Bulletin, 88:2 (1980), 359-69.

Haney, W. "Units of Analysis Issues in the Evaluation of Project Follow Through." Prepared for U.S. Office of Education, Contract No. OEC-0-74-03-94. The Huron Institute, Cambridge, Mass., 1974.

Hauser-Cram, P., and J. Shonkoff. "Report of a Meta-Analysis from the Early Intervention Childhood Project." Prepared for the National Institutes of Health, Bethesda, Md., 1986.

Hazelrigg, M. D., H. M. Cooper, and C. M. Borduin. "Evaluating the Effectiveness of Family Therapies: An Integrative Review and Analysis." Psychological Bulletin, 101 (1987), 428-42.

Hedges, L. V. "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators." Journal of Educational Statistics, 6 (1981), 107-28.

Hedges, L. V. "Estimating Effect Size from a Series of Independent Experiments." Psychological Bulletin, 92 (1982a), 490-99.

Hedges, L. V. "Fitting Categorical Models to Effect Sizes from a Series of Experiments." Journal of Education Statistics, 7 (1982b), 119-37.

Hedges, L. V. "Fitting Continuous Models to Effect Size Data." Journal of Educational Statistics, 7 (1982c), 245-70.

Hedges, L. V. "A Random Effects Model for Effect Sizes." Psychological Bulletin, 93 (1983), 388-95.

Hedges, L. V. "Advances in Statistical Analysis." New Directions for Program Evaluation, 24 (1984a), 25-42.

Hedges, L. V. "Advances in Statistical Methods for Meta-Analysis." In Issues in Data Synthesis, pp. 25-44, W. H. Yeaton and P. M. Wortman (eds.). New Directions for Program Evaluation, No. 24. San Francisco, Calif.: Jossey-Bass, 1984b.

Hedges, L. V. "Estimation of Effect Size Under Non-random Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences." Journal of Educational Statistics, 9 (1984c), 61-85.

Hedges, L. V. "Estimating Effect Size from Vote Counts or Box-score Data." Presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 1986a.

Hedges, L. V. "Issues in Meta-analysis." Review of Research in Education, 13 (1986b), 353-98.

Hedges, L. V. "Meta-Analysis." Review of Research in Education, 17 (1986c), 1-55.

Hedges, L. V. "Improving Statistical Procedures for Validity Generalization." In Test Validity for the 1990's and Beyond, pp. 191-212, H. Braun and H. Warner (eds.). Hillsdale, N.J.: Lawrence Erlbaum, 1988.

Hedges, L. V. "The Meta-analysis of Test Validity Studies: Some New Approaches." In Test Validity, H.

Bibliography

Wainer and H. Braun (eds.). Hillsdale, N.J.: Lawrence Erlbaum, 1988b.

Hedges, L. V., and I. Olkin. "Analyses, Reanalyses, and Meta-analysis." Review of G. V. Glass, B. McGaw, and Mary L. Smith, Meta-analysis in Social Research. Contemporary Education Review, 1 (1982), 157-65.

Hedges, L. V., and I. Olkin. Statistical Methods for Meta-analysis. New York: Academic Press, 1985.

Hedges, L. V., and I. Olkin. "Meta-analysis: A Review and a View." Educational Researcher, 15:8 (1986), 14-21.

Herson, M., and D. H. Barlow. Single-Case Experimental Designs: Strategies for Studying Behavior Change. New York: Pergamon Press, 1976.

Himel, H. N., et al. "Adjuvant Chemotherapy for Breast Cancer: A Pooled Estimate Based on Published Randomized Control Trials." Journal of the American Medical Association, 256 (1986), 1148-59.

Hine, L. K., N. Laird, and T. C. Charlmers. "Meta-Analysis of Randomized Control Trials of Routine Antiarrhythmic Therapy of Post Acute Myocardial Infarction Patients Indicates Urgent Need for More Trials." Manuscript, Harvard School of Public Health, Boston, Mass., n.d.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey (eds.). Exploring Data Tables, Trends, and Shapes. New York: Wiley, 1985.

Howard, K. I., et al. "The Dose-response Relationship in Psychotherapy." American Psychologist, 41:2 (1980), 159-64.

Hunter, J. E., F. L. Schmidt, and G. B. Jackson. Meta-Analysis: Cumulating Research Findings Across Studies. Beverly Hills, Calif.: Sage, 1982.

Hyde, J. S., and M. C. Linn. "Gender Differences in Verbal Ability: A Meta-analysis." Psychological Bulletin, 104 (1988), 53-69.

Iyengar, S., and J. B. Greenhouse. "Selection Models and the File Drawer Problem." Statistical Science, 3 (1988), 109-35.

Jackson, G. B. "Methods for Integrative Reviews." Reviews of Educational Research, 50 (1980), 438-60.

Johnson, B. T., and A. H. Eagley. "Effects of Involvement on Persuasion: A Meta-analysis." Psychological Bulletin, 106 (1989), 290-314.

Kamin, L. Comment on Munsinger's review of adoption studies. Psychological Bulletin, 85 (1978), 194-201.

Kane, E. D. "Differential Indications for the Use of Glutamic Acid." American Journal of Psychiatry, 109 (1953), 699-700.

Katz, B. M., L. A. Marascuilo, and M. McSweeney. "Nonparametric Alternatives for Testing Main Effects Hypotheses: A Model for Combining Data Across Independent Studies." Psychological Bulletin, 98 (1985), 200-8.

Kennedy, M. M. "Developing an Evaluation Plan for Public Law 94-142." New Directions for Program Evaluation, 2 (1978), 19-38.

Kennedy, M. M. "Generalizing From Single Case Studies." Evaluation Quarterly, 3 (1979), 661-78.

Klitgaard, R. "Going Beyond the Mean in Educational Evaluation." Public Policy, 23 (1975), 59-79.

Kohl, H. "Great Expectations." In R. E. Snow and J. D. Elashoff (eds.), Pygmalion Reconsidered. Worthington, Ohio: Jones Publishing Company, 1971.

Kratochwill, T. R. "N = 1: An Alternative Research Strategy for School Psychologists." Journal of School Psychology, 15 (1977), 239-49.

Kratochwill, T. R. Single Subject Research. New York: Academic Press, 1978.

Kulik, J. A., and C. C. Kulik, "Operative and Interpretable Effect Sizes in Meta-analysis." Presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 16-20, 1986.

Kulik, J. A., and C. C. Kulik. "Meta-analysis in Education." International Journal of Educational Research, 13:3 (1989), 221-340.

Kulik, J. A., and C. C. Kulik. "Meta-analysis: Historical Origins and Contemporary Practice." Presented at the annual meeting of the American Educational Research Association, New Orleans, La., April 1988.

Kulik, J. A., and C-L. C. Kulik. "Operative and Interpretable Effect Sizes in Meta-Analysis." Presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 16-20, 1986.

Kulik, J. A., C. C. Kulik, and P. A. Cohen. "A Meta-Analysis of Outcome Studies of Keller's Personalized System of Instruction." American Psychologist, 34 (1979), 307-18.

Lane, D. M., and W. P. Dunlap. "Estimating Effect Size: Bias Resulting from the Significance Criterion in

Editorial Decision." British Journal of Mathematical and Statistical Psychology, 31 (1978), 107-12.

Leifer, A., N. Gordon, and S. Graves. "Children's Television More Than Mere Entertainment." Harvard Educational Review, 44 (1974), 153-71.

Levin, H. "Cost-Benefit and Cost-Effectiveness Analysis." New Directions for Program Evaluation, 34 (1987), 83-99.

Levin, H. M., G. V. Glass, and G. R. Meister. Cost-Effectiveness of Four Educational Interventions. Stanford, Calif.: Institute for Research on Educational Finance and Governance, School of Education, Stanford University, 1984.

Levine, E. S. "Can We Speed Up the Slow Child?" Volta Review, 51 (1949), 169-70.

Lewis, C. E., et al. "An Evaluation of the Impact of School Nurse Practitioners." Journal of School Health, 44 (1974), 331-35.

Light, R. J. "Capitalizing on Variation: How Conflicting Research Findings Can Be Helpful for Policy." Educational Researcher, 8:9 (1979), 7-11.

Light, R. J., and D. B. Pillemer. "Numbers and Narrative: Combining Their Strengths in Research Reviews." Harvard Educational Review, 52 (1982), 1-26.

Light, R. J., and D. B. Pillemer. Summing Up: The Science of Reviewing Research. Cambridge, Mass.: Harvard University Press, 1984.

Light, R. J., and P. V. Smith. "Accumulating Evidence: Procedures for Resolving Contradictions Among Different Studies." Harvard Educational Review, 41 (November 1971), 429-71.

Linn, M. C., and A. C. Peterson. "Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-analysis." Child Development, 56 (1986), 1479-98.

Lipsey, M. W. "Juvenile Delinquency Treatment: A Meta-analytic Inquiry into the Variability of Effects." Paper for the Research Synthesis Committee of the Russell Sage Foundation, February 1990.

Louis, T. A., H. V. Fineberg, and F. Mosteller. "Findings for Public Health from Meta-analysis." Annual Review of Public Health, 6 (1985), 1-20.

McCall, R. "Challenges to a Science of Developmental Psychology." Child Development, 48 (1977), 333-34.

McCulloch, T. L. "The Effect of Glutamic Acid Feeding on Cognitive Abilities of Institutionalized Mental Defectives." American Journal of Mental Deficiency, 55 (1950), 117-22.

McGaw, B. "Meta-analysis." In Educational Research Methodology and Measurement: An International Handbook, pp. 678-85, J. P. Keeves (ed.). New York: Pergamon Press, 1988.

Matt, G. E. "Decision Rules for Selecting Effect Sizes in Meta-analysis: A Review and Reanalysis of Psychotherapy Outcome Studies." Psychological Bulletin, 105 (1989), 106-15.

Merenstein, J. H., H. Wolfe, and K. M. Barker. "The Use of Nurse Practitioners in a General Practice." Medical Care, 12 (1974), 445-52.

Mosteller, F., and R. Bush. "Selecting Quantitative Techniques." In G. Lindzey (ed.), Handbook on Social Psychology. Vol. 1. Theory and Method. Cambridge, Mass.: Addison Wesley Publishing Company, 1954.

Mullen, B., and R. Rosenthal. Basic Meta-analysis: Procedures and Programs. Hillsdale, N.J.: Lawrence Erlbaum, 1985.

Munsinger, H. "The Adopted Child's IQ: A Critical Review." Psychological Bulletin, 82 (1974), 623-59.

Munsinger, H. Reply to Kamin. Psychological Bulletin, 85 (1978), 202-6.

National Institute of Education. "School Desegregation and Black Achievement." Unpublished report, U.S. Department of Education, Washington, D.C., 1984.

Noblit, G. W., and R. D. Hare. Meta-ethnography: Synthesizing Qualitative Studies. Qualitative Research Methods, vol. 11. Beverly Hills, Calif.: Sage, 1988.

Olkin, I. "History and Goals." In The Future of Meta-analysis, K. W. Wachter and M. L. Straf (eds.). New York: Russell Sage Foundation, 1990.

Orwin, R. "A Fail-Safe N for Effect Size in Meta-Analysis." Journal of Education Statistics, 8 (1983), 157-59.

Orwin, R., and D. S. Cordray. "Effects of Deficient Reporting on Meta-Analysis: A Conceptual Framework and Reanalysis." Psychological Bulletin, 97:1 (1985), 134-47.

Parker, K. C. H., R. K. Hanson, and J. Hunsley. "MMPI, Rorschach, and WAIS: A Meta-analytic Comparison of Reliability, Stability, and Validity." Psychological Bulletin, 103 (1988), 367-73.

Pillemer, D. B. "Conceptual Issues in Research Synthesis." The Journal of Special Education, 18 (1984), 27-40.

Pillemer, D. B., and R. J. Light. "Using the Results of Controlled Experiments to Construct Social Programs: Three Caveats." Evaluation Studies Review Annual. Beverly Hills, Calif.: Sage Publishing Company, 1979.

Pillemer, D. B., and R. J. Light. "Synthesizing Outcomes: How to Use Research Evidence from Many Studies." Harvard Educational Review, 50 (1980), 176-95.

Prioleau, L., M. Murdock, and N. Brody. "An Analysis of Psychotherapy vs. Placebo Studies." The Behavioral and Brain Sciences, 6:2 (1983), 275-85.

Quinn, K. V., and D. Durling. "New Experiment in Glutamic Acid Therapy: 24 Cases Classified as Mental Deficiency, Undifferentiated, Treated with Glutamic Acid for Six Months." American Journal of Mental Deficiency, 55 (1950), 227-34.

Raudenbush, S. W. "Magnitude of Teacher Expectance Effects on Pupil IQ as a Function of the Credibility of Expectance Induction: A Synthesis of Findings from 18 Experiments." Journal of Educational Psychology, 76 (1984), 85-97.

Raudenbush, S. W., and A. S. Bryk. "Empirical Bayes Meta-Analysis." Journal of Educational Statistics, 10 (1985), 75-98.

Rivlin, A. Systemtatic Thinking for Social Action. Washington, D.C.: The Brookings Institution, 1971.

Rosenthal, R. "Combining Results of Independent Studies." Psychological Bulletin, 85 (1978), 185-93.

Rosenthal, R. "The 'File Drawer Problem' and Tolerance for Null Results." Psychological Bulletin, 86 (1979), 638-41.

Rosenthal, R. "Comparing Effect Sizes of Independent Studies." Psychological Bulletin, 92 (1982), 500-4.

Rosenthal, R. Meta-Analytic Procedures for Social Research. Beverly Hills, Calif.: Sage, 1984.

Rosenthal, R. "Experimenter Expectancy, Covert Communication, and Meta-Analytic Methods." Presented at the annual meeting of the American Psychological Association, New Orleans, La., 1985.

Rosenthal, R., and L. Jacobson. Pygmalion in the Classroom. New York: Holt, Rinehart and Winston, 1968.

Rosenthal, R., and D. B. Rubin. "Interpersonal Expectancy Effects: The First 345 Studies." The Behavioral and Brain Sciences, 3 (1978), 377-415.

Rosenthal, R., and D. B. Rubin. "Comparing Significance Levels of Independent Studies." Psychological Bulletin, 86 (1979a), 1165-68.

Rosenthal, R., and D. B. Rubin. "A Note on Percent Variance Explained as a Measure of the Importance of Effects." Journal of Applied Social Psychology, 9 (1979b), 395-96.

Rosenthal, R., and D. B. Rubin. "A Simple, General Purpose Display of Magnitude of Experimental Effect." Journal of Educational Psychology, 74 (1982), 166-69.

Rosenthal, R., and D. B. Rubin. "Meta-Analytic Procedures for Combining Studies with Multiple Effect Sizes." Psychological Bulletin, 99:3 (1986), 400-6.

Rosnow, R. L., and R. Rosenthal. "Statistical Procedures and the Justification of Knowledge in Psychological Science." American Psychologist, 44 (1989), 1276-84.

Rubin, D. B. "Using Empirical Bayes Techniques in the Law School Validity Studies." Journal of the American Statistical Association, 75 (1980), 801-16.

Rubin, D. B. "Estimation in Parallel Randomized Experiments." Journal of Educational Statistics, 6:4 (1981), 377-401.

Rubin, D. B. "A New Perspective." In The Future of Meta-Analysis, K. Wachter and M. Straf (eds.). New York: Russell Sage Foundation, 1990.

Ruopp, R. R. Children at the Center: Report of the National Day Care Study. Cambridge, Mass.: Abt Books, 1979.

Ryan, W. Blaming the Victim. New York: Pantheon Books, 1971.

Sacks, H. S., et al. "Should Mild Hypertension Be Treated? An Attempted Meta-analysis of the Clinical Trials." Mount Sinai Journal of Medicine, 52 (1985), 265-70.

Sacks, H. S., et al. "Meta-analysis of Randomized Controlled Trials." New England Journal of Medicine, 316 (1987), 450-55.

Salter, W. J. "Conducting Social Program Evaluations." Prepared for Bolt, Beranek, and Newman, Inc., Cambridge, Mass., August 1980.

Salzberg, C. L., P. S. Strain, and D. M. Baer. "Meta-analysis for Single-Subject Research: When Does It Clarify, When Does It Obscure?" Remedial and Special Education, 8 (1987), 43-48.

Sampson, G. E., et al. "The Effects of Teacher Questioning Levels on Student Achievement: A Quantitative Synthesis." The Journal of Educational Research, 80 (1987), 290-95.

Scruggs, T. E., M. A. Mastropieri, and G. Casto. "The Quantitative Synthesis of Single-Subject Research: Methodology and Validation." Remedial and Special Education, 8 (1987), 24-33.

Shadish, W. R., L. M. Montgomery, and M. Doherty. "How Many Studies Are in the File Drawer? An Empirical Estimate." Presented at the meeting of the American Evaluation Association, Boston, Mass., October 1987.

Shadish, W. R., et al. "Marital/Family Therapy Effectiveness: Meta-Analysis of 163 Randomized Trials." Presented at the American Psychological Association, New Orleans, La., August 1989.

Shapiro, D. A. "Recent Applications of Meta-analysis in Clinical Research." Clinical Psychology Review, 5:1 (1985), 13-34.

Simes, R. J. "Confronting Publication Bias: A Cohort Design for Meta-analysis." Statistics in Medicine, 6 (1987), 66-29.

Slavin, R. E. "Meta-analysis in Education: How Has It Been Used?" Educational Researcher, 13 (1984), 6-15.

Slavin, R. E. "Best-evidence Synthesis: Why Less Is More." Educational Researcher, 16:4 (1987), 15-16.

Smith, M. L. "Publication Bias in Meta-Analysis." Evaluation in Education: An International Review Series, 4 (1980), 22-24.

Smith, M. L., and G. V. Glass. "Meta-analysis of Psychotherapy Outcome Studies." American Psychologist, 32 (1977), 752-60.

Sommer, B. "The File Drawer Effect and Publication Rates in Menstrual Cycle Research." Psychology of Women Quarterly, 11 (1987), 233-42.

Stanley, J. C. "Note About Possible Bias Resulting When Under-Statisticized Studies Are Excluded from Meta-analyses." Journal of Educational Measurement, 24:1 (1987), 72-76.

Steinkamp, M. W., and M. L. Maehr. "Gender Differences in Motivational Orientations Toward Achievement in School Science: A Quantitative Synthesis." Journal of the American Educational Research Association, 21 (1984), 39-59.

Stock, W., et al. "Rigor in Data Synthesis: A Case Study of Reliability in Meta-analysis." Educational Researcher, 11 (1982), 10-14.

Strube, M. J., W. Gardner, and D. P. Hartmann. "Limitations, Liabilities, and Obstacles in Reviews of the Literature: The Current Status of Meta-analysis." Clinical Psychology Review, 5 (1984), 63-78.

Strube, M. J., and D. P. Hartman. "Meta-analysis: Techniques, Applications, and Functions." Journal of Consulting and Clinical Psychology, 51 (1983), 14-27.

Tamir, P. "Meta-analysis of Cognitive Preferences and Learning." Journal of Research in Science Teaching, 22 (1985), 1-17.

Toth, P. J., and R. I. Horwitz. "Conflicting Clinical Trials and the Uncertainty of Treating Mild Hypertension." The American Journal of Medicine, 75 (1983), 482-88.

Tracz, S. M., and P. B. Elmore. "The Effect of the Violation of the Assumption of Independence When Combining Correlation Coefficients in a

Meta-analysis." Multiple Linear Regression Viewpoints, 14 (1985), 61-80.

Ullman, Neil R. Elementary Statistics: An Applied Approach. New York: John Wiley and Sons, 1978.

Urkowitz, A. G., and R. E. Laessig. "Assessing the Believability of Research Results Reported in the Environmental Health Matrix." In Program Evaluation: Patterns and Directions, E. Chelimsky (ed.). Washington, D.C.: American Society for Public Administration, 1986.

U.S. General Accounting Office. Disparities Still Exist in Who Gets Special Education, IPE-81-1. Washington, D.C.: September 30, 1981.

U.S. General Accounting Office. CETA Programs for Disadvantaged Adults: What Do We Know About Their Enrollees, Services, and Effectiveness? IPE-82-2. Washington, D.C.: June 14, 1982.

U.S. General Accounting Office. Lessons Learned From Past Block Grants: Implications for Congressional Oversight, IPE-82-8. Washington, D.C.: September 23, 1982.

U.S. General Accounting Office. Problems and Options in Estimating the Size of the Illegal Alien Population, IPE-82-9. Washington, D.C.: September 24, 1982.

U.S. General Accounting Office. The Elderly Should Benefit From Expanded Home Health Care but Increasing These Services Will Not Insure Cost Reduction, IPE-83-1. Washington, D.C.: December 7, 1982.

U.S. General Accounting Office. WIC Evaluations Provide Some Favorable but No Conclusive Evidence on the Effects Expected for the Special Supplemental

Program for Women, Infants, and Children,
GAO/PEMD-84-4. Washington, D.C.: January 30,
1984.

Wachter, K. W. "Disturbed by Meta-Analysis."
Science, November 1988, pp. 1407-08.

Wachter, K. W., and M. L. Straf. "Introduction." In
The Future of Meta-analysis. New York: Russell Sage
Foundation, 1990.

Walberg, H. "Synthesis of Research on Teaching." In
Handbook of Research on Teaching, 3rd ed., M. Wit-
trock (ed.). New York: Macmillan, 1986.

Ward, M. M, G. E. Swan, and M. A. Chesney.
"Arousal-Reduction Treatments for Mild Hyperten-
sion: A Meta-analysis of Recent Studies." In
Handbook of Hypertension, Vol. 9., Behavioral Fac-
tors in Hypertension, pp. 1-19, S. Julius and D. R.
Bassett (eds.). New York: Elsevier Science Publishers,
1987.

White, K. R. "The Relation Between Social Economic
Status and Academic Achievement." Psychological
Bulletin, 91 (1982), 461-81.

White, K. R., D. W. Bush, and G. C. Casto. "Learning
from Reviews of Early Intervention." The Journal of
Special Education, 19 (1985-86), 401-16.

White, O. R. "Some Comments Concerning "The
Quantitative Synthesis of Single-Subject Research.""
Remedial and Special Education, 8 (1987), 34-39.

Willson, V. L. "A Meta-analysis of the Relationship
Between Science Achievement and Science Attitude."
Journal of Research in Science Teaching, 21 (1983),
289-303.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw Hill Publishing Company, 1971.

Wolf, F. M. Meta-analysis: Quantitative Methods for Research Synthesis. Sage University Paper series on Quantitative Applications in the Social Sciences. Beverly Hills, Calif.: Sage, 1986.

Wortman, P. M. "School Desegregation and Black Achievement: An Integrative View." In "School Desegregation and Black Achievement," T. Cook et al. (eds.). Unpublished report. National Institute of Education, Washington, D.C., 1984.

Wortman, P. M. "School Desegregation and Black Achievement." Sociological Methods and Research, 13 (1985), 289-324.

Wortman, P. M., and W. H. Yeaton. "Synthesis of Results in Controlled Trials of Coronary Artery Bypass Graft Surgery." Evaluation Studies Review Annual, 8. Beverly Hills, Calif.: Sage Publications, Inc., 1983.

Yeaton, W. H., and P. M. Wortman. "Evaluation Issues in Medical Research Synthesis." In Issues in Data Synthesis, pp. 43-56, W. H. Yeaton and P. M. Wortman (eds.). New Directions for Program Evaluation Series. San Francisco, Calif.: Jossey-Bass, 1984a.

Yeaton, W. H., and P. M. Wortman (eds.). "Issues in Data Synthesis." New Directions for Program Evaluation, no. 24. San Francisco: Jossey-Bass, 1984b.

Yeaton, W. H., and P. M. Wortman. "Medical Technology Assessment: The Evaluation of Coronary Artery Bypass Graft Surgery Using Data Synthesis Techniques." International Journal of Technology Assessment in Health Care, 1 (1985), 125-46.

Bibliography

Yeaton, W. H., and P. M. Wortman. "Reconceptualizing Reliability in Meta-analytic Reviews." Manuscript, State University of New York, Stonybrook, New York, 1989.

Yin, R. K., and K. A. Heald. "Using the Case Survey Method to Analyze Policy Studies." Administrative Science Quarterly, 20 (1975), 371-81.

Zabrenko, R. N., and G. S. Chambers. "An Evaluation of Glutamic Acid in Mental Deficiency." American Journal of Psychiatry, 108 (1953), 881-87.

Zimiles, H. M. "Generalizing From Single Case Studies." Evaluation Quarterly, 3 (1979), 661-78.

Zimmerman, F. T., and B. B. Burgmeister. "The Effect of Glutamic Acid on Borderline and High-Grade Defective Intelligence." New York State Journal of Medicine, 50 (1950), 693-97.

Glossary

Bias	The extent to which a measurement, sampling, or analytic method systematically underestimates or overestimates the true value of an attribute.
Case Study	A method of learning about a complex instance, based on a comprehensive understanding of that instance, obtained by extensive description and analysis of the instance, taken as a whole and in its context.
Construct Validity	The extent to which a measurement method accurately represents a construct and produces an observation distinct from that produced by a measure of another construct.
External Validity	The extent to which a finding applies (or can be generalized) to persons, objects, settings, or times other than those that were the subject of study.
Generalizability	Used interchangeably with "external validity."
Internal Validity	The extent to which the causes of an effect are established by an inquiry.
Null Hypothesis	In hypothesis testing, we should state the assumed or hypothesized value of the population figure before we begin sampling. The assumption that we want to test is called the <u>null hypothesis</u> . The term had its origin in earlier agricultural and medical applications of statistics.

Glossary

Outliers	Instances that are aberrant or do not fit with other instances; instances that, compared to other members of a population, are at the extremes on relevant dimensions.
Program Evaluation	The application of scientific research methods to assess program concepts, implementation, and effectiveness.
Qualitative Data	Information based on judgments (such as the estimated speed of a UFO) that may be expressed in numerical or nonnumerical ways and data that may not be based on judgments (such as state of birth) but are not meaningfully expressed numerically. The data sources are often textual and observational and expressed in words.
Quantitative Data	Information based on measures that do not rely on judgments and that are meaningfully measured. These are usually expressed numerically and often use continuous rather than discrete or categorical levels of measurement and scales with interval or ratio properties.
Reliability	The extent to which a measurement process produces similar results on repeated observations of the same condition or event.
Representative Sample	A sample that has approximately the same distribution of characteristics as the population from which it was drawn.

Glossary

**Simple Random
Sample**

A method for drawing a sample from a population such that all samples of a given size have equal probability of being drawn.

Contributors

**Linda G. Morra
Richard J. Light
Richard T. Barnes
Christine A. Fossett
Penny Pickett**

Papers in This Series

This is a flexible series continually being added to and updated. The interested reader should inquire about the possibility of additional papers in the series.

The Evaluation Synthesis. Transfer paper 10.1.2, formerly methods paper 1.

Content Analysis: A Methodology for Structuring and Analyzing Written Material. Transfer paper 10.1.3, formerly methodology transfer paper 3.

Designing Evaluations. Transfer paper 10.1.4, formerly methodology transfer paper 4.

Using Structured Interviewing Techniques. Transfer paper 10.1.5, formerly methodology transfer paper 5.

Using Statistical Sampling. Transfer paper 10.1.6, formerly methodology transfer paper 6.

Developing and Using Questionnaires. Transfer paper 10.1.7, formerly methodology transfer paper 7.

Case Study Evaluations. Transfer paper 10.1.9, formerly methodology transfer paper 9.

Prospective Evaluation Methods: The Prospective Evaluation Synthesis. Transfer paper 10.1.10, formerly methodology transfer paper 10.

Ordering Information

The first copy of each GAO report is free. Additional copies are \$2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20877

Orders may also be placed by calling (202) 275-6241

United States
General Accounting Office
Washington D.C. 20548

**Official Business
Penalty for Private Use \$300**

First Class Mail
Postage & Fees Paid
GAO
Permit No. G100

END