

The Evaluation Synthesis

THE EVALUATION SYNTHESIS

FOREWORD

This document outlines the methodology for an evaluative approach, currently being applied by the General Accounting Office (GAO), which we call the Evaluation Synthesis. The evaluation synthesis represents a cluster of techniques by which questions about a Federal program are developed collaboratively with congressional committee staff, existing studies addressing those questions are identified and collected, 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 congressional committees 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 congressional 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 GAO evaluation or audit work.

The evaluation synthesis has two major benefits. First, the ability to draw on a large 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 to the Congress: 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.

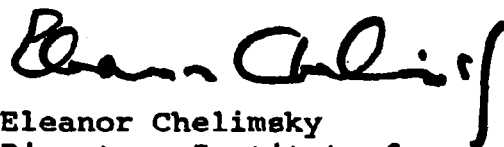
The methodology outlined here is not definitive. It reflects the work performed on four completed evaluation syntheses. There are parts of the evaluation synthesis methodology we have not described with the specificity that we

think is desirable. It is still too early in our experience to be able to do this. We will, however, be reviewing and revising the methodology both as we receive additional reactions and suggestions from people in the evaluative and legislative communities and as we conduct additional syntheses. Reactions from readers in the evaluation community therefore continue to be both needed and sought.

The paper will serve as a guideline for ourselves both to aid those staff members who have not yet conducted an evaluation synthesis, and to allow us to see what changes we need to make in developing the methodology further as we apply it to different topical areas. Thus, the document serves as a preliminary standard and as a point of departure. For example, we are developing one variation which we term the Information Synthesis. The information synthesis incorporates prospective and non-technical literature into the synthesis as well as the usual retrospective and technical studies. We are applying this variation in a synthesis on Chemical Warfare issues.

We wish to acknowledge a number of individuals outside of GAO who contributed their time and expertise in careful review of an earlier draft of this paper. We appreciate the insightful comments of Carol Weiss, Robert Haveman, Mary Kennedy, David Cordray, Ernst Stromsdorfer, Peter Rossi, John Evans, Robert Orwin, Robert St. Pierre, and Dennis Deloria. Their comments were critical in helping to improve this paper.

A special acknowledgement is extended to Professor Richard Light of Harvard University who assisted us on this paper. His research knowledge, evaluative expertise and experience with the particular problems of synthesis development were invaluable to the paper's author.



Eleanor Chelimsky
Director, Institute for
Program Evaluation

Contents

FORWARD		i
CHAPTER		
1	DEFINING THE EVALUATION SYNTHESIS	1
2	DEVELOPING THE SYNTHESIS	5
	Identifying and negotiating the study topic and questions	5
	Collecting information	8
	Determining the types of studies to include	10
	Reviewing the studies	13
	Redetermining the appropriateness of the synthesis method	16
3	PERFORMING THE SYNTHESIS	22
	Quantitative approaches for evaluation synthesis	23
	Non-quantitative approaches in evaluation synthesis	27
	Merging quantitative and non-quantitative approaches	34
	Identifying gaps	37
4	PRESENTING THE FINDINGS	39
5	STRENGTHS AND LIMITATIONS	41
6	A CASE STUDY EXAMPLE	44
TABLE		
1	Example of table overviewing the data base	48
APPENDIX		
I	References	50

CHAPTER 1

DEFINING THE EVALUATION SYNTHESIS

The Congress has frequent recurring needs for evaluative information on Federal programs. During budget hearings, for example, or during the authorization and appropriations process, or at times of oversight activities, there is a surge in Congressional need for evaluative information to assist in programmatic or policy decision-making. A frequent question, particularly with regard to service delivery programs, is how well the program is working. This usually means both whether the program is operating as intended and whether it is having the desired effects. Many valuable approaches exist for providing comprehensive and rigorous evaluative information with regard to these questions. There are, for example, impact evaluations, process evaluations, and discrepancy evaluations--all approaches to finding out both how well a program is working and what can be done to improve its performance. With application of these approaches there have, however, been major problems.

The problems are two-fold. First, because designing and conducting good evaluations take a long time, evaluators, as a group, have had great difficulty in getting valid program evaluation information to legislative users rapidly enough to fit the time constraints of the congressional negotiation or decision-making process. Second, evaluation studies have tended to increase knowledge over time in a fragmented rather than an integrated fashion. Studies on services, target groups, or other program aspects are usually released incrementally over a number of years with agencies seldom making attempts to integrate the information unless there is a requirement for a yearly report. (Even in these cases, the yearly report will typically cover only agency-sponsored evaluation studies released that year.) Thus, legislative users may either receive only part of the total information available, or else they may receive it in a form that is so voluminous and yet so fragmented as to make access difficult.

To provide timely yet comprehensive and integrated information to the Congress on how Federal programs are working, the General Accounting Office's Institute for Program Evaluation (IPE) is applying a cluster of techniques known collectively as the evaluation synthesis. This approach does not seek to produce evaluations faster; instead it addresses the problem of timeliness by making use of existing evaluations. The evaluation synthesis is a methodology for addressing only those questions which can be satisfactorily answered without conducting primary data collection; it is not a replacement for original data collection.

What can the evaluation synthesis accomplish? In IPE we have used the evaluation synthesis to answer congressional questions about how programs are working--both how programs are operating and what their effects are. For example, we have found that the evaluation synthesis can be used to provide an estimate of how many people are actually receiving program services. Our report (GAO/IPE-81-1) on the Education for all Handicapped Children Act used fourteen existing studies and two data bases to estimate and describe the number of eligible handicapped children receiving special education services. This report was able to use different sources not only to provide an estimate of how many children are receiving services, but also to provide estimates of children's racial/ethnic background, and severity of handicap. No one 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. Our special education report again serves as an example. The studies enabled in-depth examination of this issue including estimates of particular handicapping conditions underrepresented and grade/age levels with particular underrepresentation.

As stated, in addition to answering these program operations questions, we have used the evaluation synthesis to answer questions about program effects. Our report on CETA (GAO/IPE-82-2), for example, examined the effects of CETA programs on disadvantaged adult enrollees. 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 IPE report (IPE-83-1) used the evaluation synthesis method to provide estimates of the effectiveness of expanded home health care services to the elderly. Estimates of effect were provided for institutional use, client outcomes, and cost. Twelve major studies were used in determining the estimates.

We also used the evaluation synthesis to examine the comparative performance of two or more programs. Our report on block grants, for example (GAO/IPE-82-8), examined the program operations question of whether the poor and other disadvantaged groups have been served equally under block grants and categorical programs. Eight basic evaluation studies--some comprising a series of reports--were used.

The evaluation synthesis, then, as the above examples show, brings together existing studies, assesses them, and uses them as a data base for answering specific congressional questions. It enables determining what is actually known about a particular topic, estimating 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 identifying gaps in evaluative research that remain, with regard to the congressional questions.

Designed to be performed in a short time period of approximately 3 to 9 months, 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 in this time frame. 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 both with the topic and the data base.

Additional advantages of the evaluation synthesis method are that:

- o 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;
- o it can integrate administrative data and findings from studies with either qualitative or quantitative emphasis;
- o it improves the use made of evaluative information since, in and of itself, it helps ensure the initial or secondary legislative use of evaluations that have already been completed.

What is unusual about the evaluation synthesis (as opposed to the many other efforts involving the review and analysis of evaluative literature) is that, as part of an overall IPE 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 congressional need--requested or anticipated--for certain information. This means that the work must always begin with a framework of questions which impart logical cohesion to the effort. Some of the questions may be answerable by the available information but others may not be. The latter serve to identify gaps in the desired array of information. The questions must, however, be tested and judged in advance to ensure that some questions have been included which are at least partially answerable via the evaluation synthesis.

The questions, together with the available information, drive the actual procedure used to synthesize the data. While there are some general steps, detailed in the next sections, for conducting an evaluation synthesis, there is no standard procedure for actually synthesizing the information. The question(s) as well as the nature and extent of information available dictate the specific synthesis procedures used.

For discussion purposes, the design for an evaluation synthesis generally consists of eight basic steps: (1) identifying and negotiating the study topic and questions; (2) collecting evaluation and other information; (3) determining the types of studies the synthesis should include; (4) reviewing the studies; (5) redetermining the appropriateness of the synthesis method; (6) synthesizing the information and determining confidence levels; (7) identifying gaps in the evaluative knowledge that remain; and (8) presenting the findings. This report describes these steps, examines the strengths and weaknesses of the method, and presents a case study example.

The first five steps, which chapter 2 describes, show how the synthesis develops through an iterative and refining process. Steps six and seven, the actual synthesis, are described in chapter 3, while chapter 4 discusses presentations of the evaluation synthesis findings. Strengths and limitations of the evaluation synthesis method comprise chapter 5, and chapter 6, the closing chapter, is a detailed outline of one IPE evaluation synthesis.

CHAPTER 2

DEVELOPING THE SYNTHESIS

The process of developing the synthesis is iterative. Through a series of five steps the synthesis topic and information base are defined and reexamined.

IDENTIFYING AND NEGOTIATING THE STUDY TOPIC AND QUESTIONS

Since the evaluation synthesis is performed either in anticipation of a congressional need or as a response to an actual congressional request, it is specifically designed to provide information for a particular legislative purpose. There should, therefore, be a clear indication that the Congress will need certain programmatic information for a specific purpose (e.g., an anticipated hearing, an oversight review, or a reauthorization debate) in the near future before this method is selected. The skill of the evaluator may be quite important in actually anticipating congressional information needs. The evaluator is likely to need some substantive knowledge to be able to identify fundamental issues and predict when they will surface.

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:

I. Program Operations

- o Who does the program serve and to what extent are the intended beneficiaries being served?

Our report on the handicapped, for example, asked not only who was receiving services, but also what groups were over-and under-represented with respect to receipt of special education services.

- o 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 our report on CETA, for example, we examined shifts in the mix of services over time in CETA Comprehensive Services and Public Service Employment 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?

- o What administrative processes and procedures are implemented? How is the program administered?

In the IPE study of lessons learned from past block grants, we investigated studies of costs of administering block grants and the effects of fixed percentage caps on administration.

II. Program Effects

- o What are the general outcomes for program recipients?

Our home health care study, for example, investigated studies of the effects of expanded home health care on client longevity, satisfaction, physical functioning, and mental health.

- o Do program outcomes vary by type of recipient and/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.

- o What is the program impact on other than program recipients?

A major question in IPE's study of expanded home health care services was the effect of expanded home health care on nursing home and hospital use.

- o How effective is the program in terms of costs, alternative programs, or different versions of the program?

Our CETA study, for example, investigated the effectiveness of CETA in terms of post-program 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 (e.g. on-the-job training versus classroom training) was more effective than another type.

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 not only provide a framework for conducting the evaluation synthesis, but also provide a framework for reporting the findings.

The process of selecting the precise topic and identifying the actual study questions drives the evaluation synthesis method. Hence, it is vital that up-front negotiation with congressional staff take place in order that the evaluation synthesis objectives mirror congressional needs and expectations. This is particularly important because the evaluation synthesis can only answer questions for which there already exists study information. Even when there is study information, it can only answer questions to the depth or extent that the evaluation studies have addressed them, and it can only be as current as the studies themselves. It is important during this step to conduct a preliminary review of the kinds of data available. Before negotiating 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 negotiated so that they are neither so broad that addressing all of the pertinent evaluation information is not possible in a short time frame, nor so narrow that little evaluation information is available for responding to them.

Identifying the study questions is usually a two-step process. After first negotiating the main topic, a preliminary review of available evidence is done to assess the appropriateness of using evaluation synthesis to answer these questions. This initial assessment implies (at least in some cases) a renegotiation with the Congress over study questions and method. The broad steps are:

- a. Find out what the congressional committee wants to know.
- b. Find out what evidence is available.
- c. Negotiate what questions can be answered given what evidence is out there.

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 generally is less applicable to new policies or programs, unless there already exists a body of relevant, usable information. This is because substantial time is typically needed for a base of evaluative knowledge to be built around a topical area. On the other hand, for programs with a long life it may be desirable to set a cut-off point for the time frame of program operations to be covered in the synthesis.

An important consideration in early negotiation with the Congress over study topic and questions is the degree of precision needed in the answers to be found. For instance, a congressional committee 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 provided by the evaluator will either be a formal confidence interval, if the analysis is based on surveys using probability samples, or, even if it is based on case studies or less rigorous methods, the answer will have the flavor of a confidence interval. Any synthesis will 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, 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 information-users and evaluator.

COLLECTING INFORMATION

Once the specific questions have been developed (and again, the questions can only be developed soundly 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 non-agency sponsored literature. Without including non-agency

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 related to the selected topic. 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. Such descriptive data should be incorporated into the synthesis if available. 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 frame and the focus on secondary data collection required by the method dictate that interviews of agency officials and others are kept to a minimum, interviews may be needed to complete 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.

Non-agency sponsored literature covers all evaluation studies other than those initiated by the Federal agency administering the policy or program. Thus, for example, this literature would include studies sponsored by other Federal agencies in the executive branch, studies sponsored by legislative agencies such as the General Accounting Office, Congressional Research Service, and Congressional Budget Office, studies undertaken independently by State or 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, would 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.

Several potential pitfalls exist in collecting the literature for the synthesis. First, as documented by White (1982), focusing only on published reports can lead to erroneous conclusions. White found that published reports tended to have more significant positive findings than other, unpublished research reports. Those with less significant findings were less "newsworthy" and, therefore, usually not published. Thus, just examining published reports might lead to an inflated view of program impact. This, however, is a problem of omission, and there is no obvious remedy for the

problem, given that evaluators are typically limited to published studies.

Being sure that no major published study has been omitted is usually a considerable problem in an evaluation synthesis. One approach useful in preventing such an omission is to ask the assistance of outside experts to help identify the literature and/or later review the literature collected.

DETERMINING THE TYPES OF STUDIES TO INCLUDE

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. 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.

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 randomly assigned to have medical rather than surgical treatment became more intensely ill and sought surgery. The surgical group became the more severely-ill group. By identifying such "cross-overs" from medical to surgical treatment, the synthesizers were able to account for a source of bias.

As the above example shows, whenever possible the evaluation synthesizer should seek studies that use a variety of methods. These variations in study types may control bias and prove helpful in accounting for discrepancies in study findings leading to more reliable answers to congressional questions. To illustrate again, suppose Congress wished to find out first, how many people have been victimized by violent crime in each of the past five 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 capturing 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.

If Congress 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 surveys concerning violent crime where one question asked those responding that they have been victims of violent crime is whether they received government aid. Again, the key point is that in conducting a synthesis one should include both kinds of studies, if 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.

Rather than viewing diversity in types of studies as frustrating, one may capitalize on the benefits that diversity can offer. While it seems sensible to exclude specific studies from the evaluation synthesis that fail to meet basic acceptability standards (a step discussed in the next section), it is also important to realize that different types of studies may produce different outcomes simply because they are designed to elicit different information. A randomized experiment to investigate preschool effectiveness produces a different type of evidence, for example, than a descriptive comparison through case study of existing preschool centers. The variation in study types may be viewed as an asset. In fact, confidence may be highest with respect to findings that are consistent across different study designs (McCall, 1977; Pillemer and Light, 1979).

Thus, in conducting an evaluation synthesis, a synthesizer should ensure that several major types of designs, if available, are represented. If the number of studies is large, the synthesizer can stratify studies by type of design and/or general type of outcome measure, and then randomly

select a number of studies from within each stratum. This would build in diversity.

When the goal of an evaluation synthesis is to enumerate a population for a congressional committee, one might ask a number of questions whose answers can provide information about the types of studies to include. These are:

1. Have surveys based on random samples from a defined population been conducted? The best case for the evaluation synthesizer occurs when one or more surveys, based on probability sampling, have been conducted sufficiently recently to answer the congressional question. In such an instance it may be that just a few studies provide a fully adequate basis for answering the research question. Survey research is perhaps the most sophisticated and reliable form of social science research available at present. However, such research requires a list of every element in a defined population (for example, individual children) or a list of clusters of those elements (for example, schools or households). Such instances may be rare in evaluation synthesis but when the conditions for such survey research exist, the synthesizer should strive to locate available studies.
2. Is the available research a "mixed bag?" Though there may be a single, well-designed survey which has just the answer the Congress needs, it may be the case instead that there are some very well done surveys, but only for certain localities or certain years. When surveys using probability samples provide an incomplete basis for answering questions, these surveys, in combination with other types of studies, may provide the needed answers. For instance, suppose well done surveys on violent crime are available for certain years in certain states, but that data on crimes reported to police are available for every state every year. Both sources should be included in a synthesis since, in combination, they may provide the basis for better answering questions about year by year changes in crime rates for every state. For those years and states where both sources are available, the more valid survey data can be used to estimate the bias in the crime report data. The estimates of this bias could then be used to extrapolate estimates of crime for states and years where only police reports are available.
3. Is the classical survey impossible given the research question? The nature of the research problem may preclude listing the population. Illegal aliens, for

instance, are absent from lists survey researchers use as a sampling frame. Other populations of interest are reluctant to divulge their identity. The incidence of drug addition, alcoholism, child abuse, or venereal disease becomes difficult to estimate since relevant respondents conceal their identity. However, a variety of methods have been developed to estimate the size of such hidden populations. Studies using these methods should be included in the synthesis, if possible.

4. Do population estimates diverge sharply? A final case is where several estimates of the population one wants to count, each based on seemingly sound methodology, diverge sharply. Three alternatives are then available: use the average of these estimates (which may be meaningless), report only the range (which may be so wide as to be useless for decision-making) or attempt systematically to account for the variations in the estimates. In most syntheses, this last approach seems advisable. However, one needs to generate hypotheses about the sources of these variations and also to have enough studies in the synthesis to allow a test of those hypotheses.

In general, formally testing hypotheses about why studies report different estimates requires a sample of studies large enough to enable an analysis of how study results vary depending upon study characteristics. The more hypotheses, the more studies are needed. As in other statistical applications, the synthesis results become increasingly reliable as the sample size gets larger. For this reason it makes sense to include studies widely varying in their methodology. Minimal standards will be necessary, but beyond these, the proposition that variations in methodology influence results needs to be tested empirically.

REVIEWING THE STUDIES

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 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 the appropriateness of the design.

While we have determined that it is important to include different types of studies in the evaluation syntheses, what does the synthesizer do with studies that vary in quality? This is a question which has provoked heated debate.

For example, Glass and Smith (Smith and Glass, 1977; Glass and Smith, 1978) argue against instituting rigorous inclusion standards for synthesis. They choose instead to include all studies that present sufficient statistical information for computing an effect size, and to look for differences in effect sizes that may be related to differences in study characteristics such as use of randomization. They explain this position in describing a synthesis of studies on the effectiveness of psychotherapy:

The mass of 'good, bad, and indifferent' reports show almost exactly the same results. Connoisseurs' distinctions about which studies are 'best' and which ought to be discarded would lead, in this instance, to a profligate dismissal of hundreds of findings. (Glass and Smith, 1978, p. 517)

On the other side, a critic of lenient inclusion standards is Eysenck (1978). He argues that leniency constitutes an "abandonment of scholarship," and believes, at least for psychotherapy, that no study has utilized sufficient methodological controls to provide useful information: "I would suggest that there is no single study in existence which does not show serious weaknesses, and until these are overcome I must regretfully restate my conclusion of 1952, namely that there is no acceptable evidence for the efficacy of psychotherapy" (1978, p. 517).

A critical issue in this debate is what constitutes a "good study." It seems reasonable to us 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 assessment of the overall soundness of each individual study. Major weaknesses of study design, conduct, analysis, or reporting which 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:

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

This list shows some of the issues which 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, when performing the synthesis, the judgment is taken into account in determining 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 recent 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 of the IPE evaluation syntheses have done, to describe the strengths and weaknesses of the study that led to a summary judgment of quality or utility. Our report synthesizing studies on special education, for example, included the actual review of each study as a technical appendix. Thus, the basis for the judgment was available for each reader to assess.

REDETERMINING THE APPROPRIATENESS OF THE SYNTHESIS METHOD

Is the available research sufficient to answer congressional questions? In developing the evaluation synthesis, it is useful to classify each study (and/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. (See page 48 for an illustration). 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 congressional question using evaluation synthesis. For example, we collected a number of studies attempting to estimate the size of the illegal alien population in the U.S. (GAO/IPE-82-9). 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 U.S.. 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 we concluded 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 any new methodology for solving the thorny problems of applied research will promise more than it can deliver. Evaluation synthesis is no exception. While we believe evaluation synthesis is an important contribution to answering many congressional questions, it is no panacea. By what criteria may the synthesizer soberly weigh the prospects of evaluation synthesis against the prospects of new research in answering congressional questions? How can an analyst help the information user form realistic expectations, early on, about the likely accuracy of evaluation synthesis results? We recommend explicit attention to these questions in the conduct of each synthesis. Specifically we recommend that, after a preliminary review of the available evidence, but before conducting a detailed synthesis, the analyst redetermine the appropriateness of the synthesis method. The purposes of such a redetermination are the following:

1. To clarify information-user expectations before the analyst becomes involved in the details of the synthesis itself.
2. To enlist the collaboration of the congressional client in addressing likely difficulties in the work. The substantive expertise of congressional staff, for example, may prove invaluable in the on-going work.
3. To prevent months of labor being wasted when synthesis cannot likely meet congressional 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.

Criteria for redetermining appropriateness

An analyst redetermining the appropriateness of the evaluation synthesis should ask the following questions:

1. Do all studies likely share a common bias of unknown direction or magnitude?

In the coronary bypass surgery example mentioned earlier, quasi-experimental studies, in conjunction with randomized experiments, made a contribution to knowledge about the effectiveness of surgery. But what if only quasi-experimental

studies had been available? Taken as a whole, these quasi-experiments systematically over-estimated the effectiveness of the surgery. We know this because we can compare their results to those of randomized trials. In retrospect, it is thus clear that if only the quasi-experimental studies were available, the sound policy would be to recommend new research, specifically, randomized trials. This conclusion also results from research on the Salk vaccine. In that case, early quasi-experimental evidence badly underestimated the true effect of the vaccine, a finding strongly confirmed by later research using randomized assignment (Gilbert, Light, and Mosteller (1975)).

On the enumerative side, while we might expect crimes reported to the police, for example, to under-estimate crime incidence, the intuition that this is the case can be confirmed, and the likely extent of unreporting estimated, only when more reliable survey research provides a standard of comparison.

Though some retrospective cases are illuminating, prospective judgments about the appropriateness of evaluation synthesis as opposed to new research are obviously more difficult. Sound assessments would seem to require two elements. The first is detailed methodological knowledge of the available research. A thorough review of available evidence will help clarify typical problems previous investigations have encountered in evaluating a program or estimating a population. Such detailed knowledge will also surface important strategies for accounting for discrepancies among study findings. The second is a combination of substantive and methodological expertise on the research for the proposed synthesis. Biases in study findings have both methodological and substantive roots. As mentioned, on the methodological side, quasi-experiments have been found to obtain systematically different results from randomized trials. Survey results will often depend on variations in sampling plans and instrument design. Sometimes, however, variations in methodology do not predict variations in findings. To some extent the influence of methodology on outcomes can be discovered only empirically, that is, retrospectively. However, substantive expertise may help a methodologist to assess the likelihood that available study results are strongly biased in one direction.

2. How variable are the reported findings of studies available for synthesis?

A thorough review of available evidence should estimate the variability in reported outcomes of studies.

This variability may then be compared against the precision needed by the Congress in the answers to their questions. Assuming no systematic bias exists (see previous question and discussion), if the variability of reported findings is within tolerable limits (as compared to the required precision), the synthesis would appear appropriate. If not, two strategies are available: a) attempt systematically to account for variability in study findings using a combination of methodological and substantive insights based on collaboration between the analyst and the congressional committee; b) recommend original research as an alternative to evaluation synthesis. This decision is based on the answer to the following question.

3. If the variability in findings is great, what are the prospects of accounting for their variability using knowledge of variations in study features?

Again a combination of detailed familiarity with the data and collaboration of methodologists and substantive experts can help to answer this question.

4. Are the findings of previous studies likely to be outdated?

Social science generalizations tend to decay over time. The extent of the decay varies radically depending on the substantive area. Previous research may enable an empirical test of the proposition that findings are time dependent.

5. Do available studies adequately cover the range of settings and populations to which the congressional client needs to generalize?

Again familiarity with available studies and active negotiation to clarify congressional information needs is crucial to answer the question.

6. Is original research feasible from a timeliness point of view for answering the congressional questions? If so, is it likely to do better than evaluation synthesis?

There are at least four alternative courses of action open to the information-user, given enough time:

- a. evaluation synthesis
- b. new original research
- c. a combination of synthesis and original research

- d. no further effort to obtain information of an empirical character; it may be perfectly reasonable to rely on expert judgment or argument as an alternative to empiricism.

If Congress requires information, one must assess the marginal return of evaluation synthesis versus original research or a combination in terms of new expenditures of time and money. Specifically one must ask:

- a. Is new research a viable way to provide information given the time available for answering the congressional need?
- b. How likely is new research to solve the problems that have plagued previous research?
- c. How much will new research cost?

Possible outcomes of the appropriateness determination

1. Recommend evaluation synthesis. If the appropriateness assessment redetermines the applicability of synthesis, the main accomplishments of the assessment will have been:
 - a. To clarify the goals of evaluation synthesis. Specifically, the assessment will determine whether the main feature of synthesis with respect to each research question will be to estimate an "on average" answer or to use knowledge about variations in study methodology, setting, and content to account for variability in findings.
 - b. To clarify user expectations for the information value of the synthesis.
 - c. To enlist the active and continuing collaboration of the congressional committee in the purpose and structure of the synthesis.
2. Recommend renegotiating the question. Another possible outcome of the appropriateness assessment is that once again the exact questions to be answered by the synthesis must be renegotiated. The indepth review of the available evidence, which is complete at this stage, may indicate that synthesis is still applicable, but only if the questions are revised. The synthesizer and congressional staff will need to work closely to maximize the likelihood that the

specific modifications will result in answerable questions that still meet the congressional committee's information needs.

3. Recommend new research. New research may be recommended as an alternative to synthesis if time is available. Or the assessment may recommend a combination of synthesis and new research. For example, the synthesis can be merged with small scale interview efforts as was demonstrated in the IPE report on lessons learned from past block grants (GAO/IPE-82-8). In either case a major value of the developmental synthesis work and the appropriateness assessment is to give the recommendation of new research a rigorous, systematic character it has sometimes lacked, especially in applied research. By assessing the entire body or a representative sample of available studies, assessing the biases and heterogeneity of its findings, and assessing its characteristic weaknesses, the act of recommending new research has a solid foundation that should contribute to future utility of the findings for decision-making. A report should be developed, similar to IPE's report on the size of the illegal U.S. population (GAO/IPE-82-9), which assesses the available studies individually and as a whole, identifies the research needed, and specifies features requisite for the design of such research.

CHAPTER 3

PERFORMING THE SYNTHESIS

Given a set of studies which have been individually assessed and deemed usable for the synthesis, the next question is: How are the different studies compared? The answer is that there is no standard approach. 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, in the evaluative synthesis 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: (1) who does the program serve under ideal circumstances?, or alternately, (2) who does the program serve on the average? In the first instance, the analyst might want to investigate a number of case studies and provide a narrative description of the findings. In the second instance, the analyst might take the arithmetic average of the answers given by the individual studies available or the analyst might express the answer as the range between the highest and lowest estimates. This analysis would, however, ideally be quantitative in contrast to the former example. Survey data would be appropriate, if they were available. A problem here is that, since the evaluation synthesis is employed to answer congressional questions, rather than to produce new knowledge given existing information, there is little likelihood that in performing an evaluation synthesis for the Congress, the ideal quantitative analysis will be 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.

This chapter discusses both quantitative and non-quantitative approaches to evaluation synthesis. The first are ideal for certain questions, but are often feasible only for the researcher who selects a topic for synthesis based on its new knowledge potential. The second are what most synthesizers will have to wrestle with when responding to congressional committees' policy driven rather than knowledge driven types of questions.

QUANTITATIVE APPROACHES FOR EVALUATION SYNTHESIS

The literature describes four basic quantitative or statistical approaches for synthesizing the findings of experimental or quasi-experimental studies. These approaches, detailed in the following sections, are (1) conducting a combined significance test, (2) computing an average effect size, (3) blocking, and (4) the cluster approach. Because the basic assumptions needing to be met are quite stringent, however, the IPE syntheses to date have not been able to use these quantitative approaches. Indeed, it is not expected that there will be many occasions for their use 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, and we present them here as ideal methods for use when possible.

1) Conducting a combined significance test. When multiple independent studies all compare two treatments, the treatments 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 groups have the same population mean.

A number of procedures have been suggested using this idea. Rosenthal (1978) has summarized many of them, and has 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 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. Other conceptually similar techniques include adding weighted Z's, adding t's, adding logs and adding probabilities (see Rosenthal, 1978, for a detailed explanation and computational examples).

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 adds the caveat that the studies should have tested the same directional hypothesis. To illustrate this approach, assume 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. On the other hand, if the studies are combined (e.g., 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.

2) Computing an average effect size. The techniques just discussed focus on statistical significance of results. Glass (1977), while recognizing the value of significance testing, argues for a shift in emphasis. He points out that sometimes the results of a combined statistical test are not particularly illuminating:

For most problems of meta-analysis, however, the number of studies will be so large and will encompass so many hundreds of subjects that the null hypothesis will be rejected routinely. Perhaps it is more realistic to think of the typical meta-analysis problem as residing in that vicinity the statistician calls the limit, where all null hypotheses are false and inferential questions disappear. The statistical integration of studies probably ought to fulfill descriptive purposes more than

inferential ones, though obviously it may fulfill both (p. 361).

The key descriptive statistic that Glass (1981) has employed in his pioneering synthesis is the effect size. When comparing 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 a study included 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 changes over time. The above example provides a working definition of effect size that is congruent with Glass' extensive work.

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 conclude 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.

3) The blocking technique. The two procedures discussed so far emphasize a "pulling together " of information. In

increasing sample size by combining across studies, or in computing a broad average result, evaluation synthesis takes the view that the best use of several small studies is to treat them as smaller proxies of a much larger study. This larger study is not available, but if it were, and if it were well done, it would get at the "truth".

A different emphasis may sometimes be useful in the evaluation synthesis. Instead of focusing on how to most effectively pull together several results into a single grand finding, a synthesis might actually try to do almost the opposite: search for variation, or particular discrepancies, among study findings. If there is a large number of studies, the opportunity exists to search for unexpectedly large variation in findings, and to try to explain it.

Rosenthal (1978) presents a procedure for doing this. We assume several studies each compared programs A and B, and an analyst wants to combine their results. The blocking technique involves comparing the outcomes by formatting the results into an overall analysis of variance (ANOVA), with studies regarded as a blocking variable. If the means, sample sizes and standard deviations are available from each study, means squares can be constructed and a two-way ANOVA (treatments by studies) can be performed (see Rosenthal, 1978, for additional details).

Studying the "main effect" of treatments from the ANOVA provides an average measure of their differential effectiveness. This fulfills one purpose of synthesis. As with the procedures for conducting a combined test, the blocking technique can dramatically increase the power of the statistical test.

But the key payoff of including studies as a blocking variable is that it helps to identify unusual variation in outcomes. If the size of the effects in the separate studies differs sharply, the studies-by-treatment interaction term in the analysis of variance will turn up significant.

4) The cluster approach. A "cluster technique" is also available for dealing with the question of aggregation of experimental or quasi-experimental studies. The underlying idea is that subgroups within any broad treatment group must be compared before they can be merged for the purposes of a overall test. A series of "hurdles" must be passed before combining. Suppose that ten studies each compare treatments A and B. Before collapsing all of the A subgroups or B subgroups together, the synthesizer must first determine that the

means, variances, relationships between dependent variables and co-variates, subject-by-treatment interactions, and contextual effects are similar across subgroups. Data may be combined only if the subgroups are similar or if an explanation for the differences is uncovered, so subgroups can be statistically adjusted prior to combining. After combining, an overall test for the difference between A and B is performed.

The cluster approach requires studies to meet even more stringent requirements than other quantitative approaches. Participants in each study must come from a precisely definable population. Outcome measures must be comparable across studies. In addition, access to raw data from each study may be necessary. If the hurdles for collapsing across subgroups can be passed without the need for raw data, the main effect of programs or treatments can be studied using summary statistics. This involves carrying out an ANOVA in a fashion similar to the blocking technique.

NON-QUANTITATIVE APPROACHES IN EVALUATION SYNTHESIS

IPE's experience to date is that the evaluation studies available for answering congressional questions have not met the assumptions or contained sufficient information to allow use of the powerful statistical approaches described in the previous section. Additionally, while the number of quantitative studies usually has been extremely limited, 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. This section details these five types of information, describes general situations in which this information should be synthesized, and outlines some guidelines for incorporating such information.

The five types of information potentially valuable for the evaluation synthesis which are not suitable for statistical analysis are (1) single case designs, (2) non-quantitative aggregate studies, (3) non-quantitative 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.

1. Single case design. Detailed studies of single cases are common, and techniques for analyzing such information are being 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) present many examples of "N = 1" research in psychology. Case studies are also frequently used in public policy analysis to examine the effects of non-experimental 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 disaggregated 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 IPE synthesis on special education were case studies of local school districts.

2. Non-quantitative 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 feel 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 non-quantitative insights.

As Zimiles (1980) points 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 (p. 7).

Here an evaluator is faced with a trade-off between precision and meaning. Organizing a synthesis forces us to confront a similar dilemma. Which outcomes appearing in the studies should be included in a synthesis? If we decide not to rely exclusively on quantitative measures, we must figure out how to incorporate non-quantitative 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 IPE syntheses fall into this category. For example, IPE's block grant synthesis identified about 10 reports focusing on administrative costs before and after program consolidation. The calculation of comprehensive and reliable estimates of effect was hindered, however, 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 non-quantitative manner.

3. Non-quantitative 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 painstakingly laid out. Often the effort put into these non-quantitative descriptions far surpasses that involving the numerical information. It is not always either appropriate or desirable to reduce a study to one or at most several numerical indices. The one number 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 which become major study limitations. Most synthesizers will need to include information in the evaluation synthesis that goes beyond numerical outcomes.

4. Expert judgment. A reviewer may choose to include expert opinion at early stages of the synthesis, such as in evaluating individual studies. Or, he or she may want to systematically compare studies relying on expert judgments about program effectiveness. Syntheses should be able to incorporate these inputs.

5. Narrative reviews of collections of research studies. As Cook and Leviton (1980) have pointed out, a carefully done narrative review, explicit about its analytic procedures, can be extremely valuable. Narrative reviews of collections of

research studies frequently, for example, may identify methodological weaknesses of certain broad types or groups of studies in a particular topic area. The synthesizer will need to consider these points in considering whether or not to include these studies in the synthesis and, if included, in interpreting findings from these studies.

Indications of the need for non-quantitative approaches

There are special circumstances when non-quantitative approaches to the evaluation synthesis are particularly appropriate. Four of these situations are when (1) treatments may be individual and/or more concerned with process than outcomes, (2) program effects are assessed across multiple levels of impact, (3) uncontrolled treatment groups are compared with the treated control group, (4) the "wrong" treatment is studied. The following sections further explain these situations.

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 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.

Non-quantitative 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 help develop their child's individualized education 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 policy maker who wants to know if P.L. 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.

2. Assessing program effects across multiple levels of impact. Quantitative approaches can be employed when all the studies have assessed program effects at the same "level" or unit of impact. This level often is the individual participant. For example, most day care studies examine behaviors of participating children. But programs can have impact 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 aggregating 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 receiving a treatment to others who do not, subtle differences between similarly labeled treatments are common. Non-quantitative information can offer valuable guidance in helping a reviewer to decide how similar the treatments are.

A recent example of this comes from a study by Fosburg, Glantz, et al., (1980). They reviewed a series of studies of children's nutrition programs 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 non-quantitative 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 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.

4. Studying the "wrong" treatment. Occasionally when synthesizing outcomes, in those cases where quantitative approaches have proved feasible, one finds that a relationship between a program and an outcome is not as strong as was originally hoped, but that outcomes are sometimes successful. This may lead to research for features of a program other than the originally planned treatment that might explain the differential success. Here, descriptive or non-quantitative 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 a reviewer 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.

IPE's approach to evaluation synthesis

As can be seen from these situations, quantitative and non-quantitative approaches can go hand-in-hand or stand alone. There are few guides in the literature, however, to non-quantitative approaches to evaluation synthesis. The general IPE 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.

The IPE approach begins with the review of the individual study, or study type, to identify strengths and weaknesses of the study which will affect confidence in the findings. If there is major weakness in the design and conduct of the study, low confidence in the individual study findings will, of course, be indicated. For example, the IPE synthesis on home health care 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, those studies which considered only the postprogram experiences of CETA trainees without regard to participants' preprogram experiences or without comparison groups were omitted from the synthesis.

Weak studies are not always omitted, however. For example, the IPE 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 studies were examined to see if any general patterns were discernible across the entire set of estimates. Given the weaknesses of the data, however, 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 non-quantitative approach. For example, the IPE 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 impact evaluations.

In brief, the non-quantitative approach used in the IPE synthesis generally requires 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 NON-QUANTITATIVE APPROACHES

Ideally, the non-quantitative approaches to evaluation syntheses should complement the quantitative approaches. Several of the types of information discussed in the previous section on non-quantitative approaches--single case designs and non-quantitative information on quantitative studies--illustrate how non-quantitative information can supplement the quantitative, when it is in fact feasible to implement quantitative approaches. We also described situations, such as the uncontrolled treatment groups and treated control groups, where without the insights provided by the non-quantitative information the quantitative analysis would be, at best, misleading.

Non-quantitative 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 for setting-by-treatment interactions, to determine if case studies show a similar program was implemented at different sites, or to examine variation across studies in relation to design characteristics and analysis strategies. Taking this perspective, variation among study findings uncovered through one approach to synthesis and investigated through another can be a useful, constructive, information-laden occurrence.

Exploiting differences in study findings

To benefit from discrepancies among studies, whether uncovered through a quantitative or non-quantitative 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 six specific ways to seek out and confirm explanations for conflicting findings.

1. Determining if similarly labeled treatments and programs differ in important ways. Just because several research reports describe a program such as Head Start, or Follow Through or Upward Bound, one should not assume these are in fact the same program. An initial step in synthesizing findings across studies should be to see whether a set of programs with the same name are in fact providing the same services. In the process, we may discover program variations that are especially effective or ineffective, leading to valuable substantive insights. For example, one analysis of public school effectiveness indicated that the most effective schools had better paid and more experienced teachers, and smaller classes than the average school (Klitgaard, 1975).

2. Looking for setting-by-treatment interactions. Assurance that programs with the same title are in fact similar will not eliminate conflicts. A treatment may be more or less effective depending on who participates in it, where it is administered, or some other situational factor. These setting-by-treatment interactions are quite common in social and educational programs. To provide one example, offering rewards to children contingent upon correct performance on IQ tests is effective in raising the scores of initially low scoring subjects, but has little effect upon the performance of higher scorers (Clingman and Fowler, 1976). So this "reward" treatment is not universally effective, but rather depends to some degree upon the characteristics of children receiving it.

3. Investigating different research designs used across studies. One study characteristic that has received particular attention is how people are assigned to treatment and control groups (Campbell and Boruch, 1975). For example, Gilbert, Light, and Mosteller (1975) report a strong relationship between the use of randomization and reported outcomes in studies of the effects of shunt operations in medicine.

One design characteristic, such as the length of time a treatment is implemented, can influence findings. For

example, there are a number of short-term experiments and several longer-term studies investigating the effects of TV violence on children's attitudes and behavior. The short-term studies generally show that viewing violence increases children's aggression, while some longer-term studies demonstrate increased aggressiveness in children assigned to a nonviolent TV diet (Leifer, Gordon and Graves, 1974).

4. Examining different analysis strategies used in different studies. Even if all the analyses are done correctly, the particular analysis procedures that are used may be related to findings and create artificial conflicts. For example, as we have previously pointed out, the unit or level of analysis may differ across studies. Whether an analysis is conducted at the pupil, class, or school level dramatically influences estimates of the strength of relationships in evaluations such as Project Follow Through (Haney, 1974).

5. Relating background variables to findings. One strategy involves coding information about participants' background characteristics (e.g., SES) and design characteristics of the research (e.g., 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 (1978) 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 curricula.

With a few key explanatory factors identified, an 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 1978, 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 ten Head Start programs produce relatively consistent results. Suppose also that while the program curricula 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 to limit the field for first-cut analysis and study designs.

IDENTIFYING GAPS

Little needs to be said about the next step in performing the synthesis--identifying gaps in knowledge about a program. The evaluation synthesis methodology clearly points out two types of information gaps. The first type has to do with gaps because some questions were not addressed. In the second type, there is lack of information because questions, although

addressed, were inadequately or poorly answered. The first gap indicates studies that may need to be conducted; the second may show studies that need to be redone or perhaps reconceptualized.

Documenting the fact that answers are not available for questions posed by the Congress is of major importance in the evaluation synthesis. In addition to pointing out areas where evaluative efforts may be needed, it also provides understanding of the completeness and reliability of the information on which potential policy decisions may be based.

CHAPTER 4

PRESENTING THE FINDINGS

The information generated through the evaluation synthesis process is brought together in a report that is carefully formatted to respond to the questions which were negotiated with congressional staff. An introductory chapter is recommended which briefly describes the history both of the study and the particular Federal program(s) involved, and presents the study objectives, scope, and methodology. The latter section might include a framework showing the evaluation questions and subquestions, a description of the evaluation studies and data bases, a table showing the relation 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 been conducted within the last three years or that they have experimental designs). The section should answer the following types of questions: 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 congressional questions. While the body of the report includes discussion of the adequacy of the data available for response to a particular question, we suggest including the actual technical reviews of the studies as an appendix. The technical appendix should provide information enabling the reader to judge the validity of both the reviewer's conclusion about each study and the reviewer's general use of the study data in the report. The technical review should systematically describe each study across such dimensions as title, report reference, study purpose, data collection period, sample selection, data collection, data analysis, and usefulness (previous section on reviewing the studies indicated the types of questions leading to judgments of usefulness). Data bases also should be described, although not all the same dimensions will be appropriate.

For several reasons, we suggest caution be exercised in drawing conclusions from the synthesised data and 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.

A briefing for Congressional members is a recommended part of presenting evaluation synthesis findings. Following agency review of the findings, the briefing is a useful vehicle for highlighting findings and drawing implications for planning oversight hearings and/or considering legislative changes.

CHAPTER 5

STRENGTHS AND LIMITATIONS

As with all methods, the evaluation synthesis, even when used appropriately, is no panacea. It does some things better than others. This chapter discusses the strengths and limitations of the method.

STRENGTHS

The major advantage of the evaluation synthesis, given that the preconditions for its use are met, is its ability to provide relatively inexpensive, comprehensive, and timely information to the Congress. It is designed to be performed by one or two persons, with methodological expertise, over a time period of approximately 3 to 9 months. By integrating findings from already completed studies, the evaluation synthesis can potentially serve congressional needs for relatively short-term evaluative information. The focus of the evaluation synthesis is tailored to specific congressional concerns.

The evaluation synthesis directs what can be large amounts of existing and possibly conflicting information from a variety of sources and time periods to the answering of specific questions. It enables the initial secondary legislative use of evaluations which have already been completed. In fact, it also enables use of evaluations even though the question at hand may have been a secondary rather than a primary focus of the study.

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.

The evaluation synthesis, by drawing together information about a specific question from a disparate number of completed evaluation studies, 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 identification of remaining unanswered questions. IPE evaluation syntheses have previously identified studies that are needed. Our evaluation synthesis of CETA, for example, pointed out that studies are needed of

the predictive power of short-term performance indicators like job placement and also of the factors that govern service mix for the individual operating agency.

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 which influence the usefulness of the study, and this can also influence a sponsor's posture with regard to future studies undertaken. Our synthesis of special education studies, for example, found that many study reports did not adequately describe the methodology they employed. The Department of Education indicated in their comments on our report that they had reviewed the studies the report used and agreed that the criticism was valid. Since most of the studies were conducted under contract to the Department, 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.

LIMITATIONS

The main limitations of the evaluation synthesis methodology stem from its reliance on extant data. The methodology is best applied to those areas where 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 only answer questions to the extent that the existing studies have addressed them. Thus, for example, evaluation synthesis 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 and their quality.

Poor reporting also limits the evaluation synthesis. As discussed in the previous section, procedures may be 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 be so minimally described that judgments cannot be made about the similarities and differences across studies. Another problem that can be caused by poor reporting is that variables of interest

may not be 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 which program managers have already taken steps to address and which are no longer characteristic of the program. The methodology is thus 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.

CHAPTER 6

A CASE STUDY EXAMPLE

The General Accounting Office's Institute for Program Evaluation (IPE) completed its first evaluation synthesis in three months, applying the methodology outlined above (GAO/IPE-81-1). The policy area selected was the Education for All Handicapped Children Act of 1975, Public Law 94-142. It might be useful to describe briefly how an actual study conformed to the outline presented.

Identifying and negotiating the study topic and questions

Special education was an area in which IPE had a staff member who was already familiar with the general issues and the evaluation literature. No start-up time was necessary. Further, congressional hearings were anticipated on Public Law 94-142 within a 6-month period, congressional staff were interested in obtaining additional evaluative information about the more than 900 million dollar program, and a recent General Accounting Office report on the program's implementation provided potential topic areas for more in-depth investigation (GAO/HRD-84-53). An additional factor in selecting special education was that we knew that the Federal plan for the evaluation of Public Law 94-142 (Kennedy, 1978), which calls for multiple studies to address a series of evaluation questions over time, had resulted in the availability of a considerable number of evaluation studies sponsored by the Office of Special Education.^{3/}

Given these factors, and congressional interest in the synthesis, the topic of access to special education (or who the program serves) was judged suitable for the synthesis. Working closely together, congressional staff and IPE staff identified four questions which provided the synthesis framework:

- What are the numbers and characteristics of children receiving special education?
- Are there eligible children underrepresented in special education?
- Are certain types of children overrepresented in special education?
- What factors influence who gets special education?

Although subquestions were further identified for each of the four questions, the four questions were the major organizers.

IPE staff knew the evaluation literature well enough to determine that there was at least some information available on each question. We were also able to judge that the questions were neither so broad that we could not address them in the committee's time frame, nor so narrow that we would have little to say. Given the relatively young age of P.L. 94-142, no cut-off date for the period we would study was needed.

Collecting information

Again, because of our in-house expertise, collecting the relevant evaluation information was not the challenge it sometimes can be. No outside experts were needed to help identify the literature; interviews of agency officials and site visits were not conducted.

IPE staff identified 15 relevant evaluation studies; seven of these studies were sponsored by the Office of Special Education as part of their evaluation plan for Public Law 94-142. Others were Office of Special Education-sponsored, but field-initiated research studies, evaluation studies sponsored by other parts of the former Office of Education, General Accounting Office reports, or for example, independent association studies. Use was also made of two relevant data bases, Public Law 94-142 State Child Count data and Elementary and Secondary Civil Rights Survey data. Available data summaries provided a basis for further analyses.

Determining the studies to include

Given the limited number of relevant studies, we did not have a logical need to reduce the quantity of studies. Additionally, we viewed variation in study types as an asset and, therefore, we did not focus on studies that used one particular method versus some other method. Basically, we included all studies available for the synthesis.

Reviewing the studies

A cross-check of studies by evaluation question and subquestion indicated that not only did no one study address all four questions, but also the studies did not distribute themselves evenly across questions. Some questions had more potential information available than others to address them and relative information gaps were immediately apparent. We have included a portion of the table to show its utility in presenting an overview of the nature and extent of the data base available to answer the study questions. (See page 48.)

Next, IPE staff reviewed and assessed each study--whether case study or survey or content analysis. Each technical review described the study purpose, data collection period, sample and selection procedures, data collection and analysis, and general usefulness. Criteria for determining usefulness, which were presented in the report itself, were indicators of sound methodology such as a clear and appropriate study design and procedures to ensure reliability across data collectors.

No formal tests of the consistency or reliability of the judgments of technical adequacy were made. Each review was, however, included in the report as part of a technical appendix. Thus, the basis for the judgments of technical adequacy was there for each reader to assess.

Redetermining the appropriateness of the synthesis method

We did not formally go through this step in performing this first IPE synthesis effort.

Synthesizing the information and determining confidence levels

Quantitative techniques (i.e., conducting a combined significance test, computing an average effect size, blocking or the cluster approach) were not appropriate either for the question being answered or for the studies available. Our approach was non-quantitative.

Our first step was to determine the best source (or sources) of information available for answering the questions. With the information sources for each major question already identified, we used the appropriate technical reviews. We then analyzed the similarities and the differences in the findings of the studies. If findings were similar across studies, we still looked for alternative

explanations for the findings. If an exception stood out, we examined whether there were methodological reasons for the findings or if a significant program variation could be suggested. Thus, both the rigor or methodological strength of the study (or studies) and identification of alternate explanations for the findings influenced the confidence we placed in the findings.

We used case study data to expand upon survey data. For example, survey data showed convincingly that males are numerically overrepresented in special education classes; case studies of local school districts suggested that at least one contributing explanation for the phenomenon might be bias in the child referral and assessment practices.

We placed increased confidence in findings that were similar across disaggregated studies of similar events. For example, three large case studies of multiple school districts each found that access to special education is dependent on school district resources. The case studies were conducted by different groups, at different times, and with different site selection criteria. That the findings concerning local resources stood out in each study resulted in our "weighing" the findings across studies more strongly than we did the individual studies.

Identifying gaps in the evaluative knowledge

Evaluation gaps were readily apparent. In the special education synthesis, some gaps resulted from lack of information because the question, although addressed, was inadequately addressed. For example, we pointed out that some studies tried to answer questions about the program before it had been implemented. Confidence in the findings was low as it was reasonable to assume that as implementation progressed, a very different picture would have emerged. Other gaps in the evaluation information resulted from lack of studies. For example, we found little investigation of relationships between high school drop-outs and children potentially in need of special education.

Presenting the findings

The four questions which focused our special education synthesis each headed a chapter in the report. No recommendations were made, although we did include some general observations on the overall findings.

Table 1
EXAMPLE OF TABLE OVERVIEWING THE DATA BASE
Table 1.2

SUMMARY OF RELEVANT STUDIES AND DATA BASE

<u>Name</u>	<u>Source/ contractor</u>	<u>Evaluation question/ subquestion</u>	<u>Methodology</u>	<u>Data collection period</u>
I. <u>Studies</u>				
A National Survey of Individualized Education Programs for Handicapped Children*	Research Triangle Institute	1.0/1.1-1.6 2.0/2.2,2.3 3.0/3.1,3.2 3.3	National Survey	2/79-5/79
A Study of the Implementation of Public Law 94-142 for Handicapped Migrant Children	Research Triangle Institute	2.0/2.4	Survey	3/80-5/80
Case Study of the Implementation of Public Law 94-142*	Education Turnkey Systems	4.0/4.1,4.3	Case Study	Fall 1977- Winter 1979
Local Implementation of Public Law 94-142: First Year Report of a Longitudinal Study*	SRI International	2.0/2.1,2.2 4.0/4.1,4.2 4.3	Case Study	9/78-6/79

*Indicates studies conducted for the Office of Special Education in response to the Federal plan for evaluation of Public Law 94-142.

<u>Name</u>	<u>Source/ contractor</u>	<u>Evaluation question/ subquestion</u>	<u>Methodology</u>	<u>Data collection period</u>
An Analysis of Categorical Definitions, Diagnostic Criteria and Personnel Utilization in the Classification of Handicapped Children*	The Council for Exceptional Children	4.0/4.2	Document Review	State documents in effect July 1977
Validation of State Counts of Handicapped Children: Volume II- Estimation of the Number of Handicapped Children in Each State*	Stanford Research Institute	2.0/2.1	Review of Studies	Studies conducted prior to 1977
Service Delivery Assessment: Education for the Handicapped	Office of the Inspector General DHEW	2.0/2.1,2.2 4.0/4.1,4.2 4.3	Case Study	Report issued 5/79
Unanswered Questions on Educating Handicapped Children in Local Public Schools	Comptroller General, GAO	2.0/2.1 3.0/3.1 4.0/4.2,4.3	Case Study	1977-1979
Case Studies Of Overlap Between Title I and Public Law 94-142 Services for Handicapped Students	SRI International	4.0/4.4	Case Study	Report issued 8/79

REFERENCES

- BELSKY, J. and L.D. Steinberg. The Effects of Day Care: A Critical Review. Child Development, 1978, 49, pp. 929-949.
- 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. Chapter in Evaluation and Experiment, C.A. Bennett and A.A. Lumsdaine (ed.). New York, New York: Academic Press, 1975.
- COOK, T.D. and Leviton, L.C. Reviewing the Literature: A comparison of Traditional Methods with Meta-Analysis. Journal of Personality, 1980, 48, pp.449-472.
- DUKES, W.F. N=1. Psychological Bulletin, 1963, 64, 74-79.
- EYSENCK, H.J. An Exercise in Mega-Silliness. American Psychologist, 1978, 33, pp. 517.
- FOSBURG, S. and Glantz, F. 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, 1978, 33, pp. 515-517.
- GILBERT, J.P., R.J. Light, and F. Mosteller. Assessing Social Innovations: An Empirical Base for Policy. Chapter in Evaluation and Experiment, op. cit.
- GLASS, G.V., Bibliography of Writings on the Integration of Research Findings. Unpublished Manuscript. Denver; University of Colorado. Laboratory of Educational Research. 1978.
- Intergrating Findings; the Meta-Analysis of Research. Chapter in Review of Research in Education, 1977, 5, pp. 351-379.
- Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, 5, November 1976, pp.3-8.
- GLASS, G.V., McGaw, B. and Smith, M.L. Meta-Analysis in Research. Beverly Hills, California. Sage Publications, 1981.

- GLASS, G.V. and Smith, (1978). Reply to Eysenck. American Psychologist. 1978, 33, p. 517.
- HALL, J.A. Gender Effects in Decoding Nonverbal Cues. Psychological Bulletin, 85, 1978, pp. 845-857.
- HANEY, W. Units of Analysis Issues in the Evaluation of Project Follow Through. Document prepared for U.S. Office of Education, Contact No. OEC-0-74-03-94. Cambridge, Mass. The Huron Institute, 1974.
- HERSON, M. and Barlow, D.H. Single-Case Experimental Designs: Strategies for Studying Behavior Change. New York: Pergamon Press, 1976.
- HUNTER, J.E., Schmidt, F.L. and Jackson, G.B. Meta-Analysis: Cumulating Research Findings Across Studies. Beverly Hills, California: Sage Publications, 1982.
- KENNEDY, M.M. Generalizing From Single Case Studies. Evaluation Quarterly, 1979, 3, pp.661-678.
- Developing an Evaluation Plan for Public Law 94-142. New Directions for Program Evaluation, 1978, 2, pp. 19-38.
- KLITGAARD, R. Going Beyond the Mean in Educational Evaluation. Public Policy, 1975, 23, pp.59-79.
- KRATOCHWILL, T.R. Single Subject Research New York: Academic Press, 1978.
- N=1: An Alternative Research Strategy for School Psychologists. Journal of School Psychology, 1977, 15, pp.239-249.
- LEIFER, A., N. Gordon, and S. Graves. Children's Television More Than Mere Entertainment. Harvard Educational Review, 1974, 44.
- LIGHT, R.J. and P.V. Smith, Accumulating Evidence: Procedures for Resolving Contradictions Among Different Studies. Harvard Educational Review, November 1971, 41, pp. 429-471.
- MCCALL, R. Challenges To A Science Of Developmental Psychology. Child Development, 1977, 48, pp.333-334.
- 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, California: Sage Publishing Company, 1979.

- ROSENTHAL, R. Combining Results of Independent Studies. Psychological Bulletin, January 1978, 85, pp. 185-193.
- SALTER, W.J. Conducting Social Program Evaluations. Essay Prepared for Bolt, Beranek, and Newman, Inc., Cambridge, Mass., August, 1980.
- SMITH, M.L. and G.V. Glass. Meta-analysis of Psychotherapy Outcome Studies. American Psychologist, 32, 1977, pp.752-760.
- STOCK, W. et. al. Rigor in Data Synthesis: A Case Study of Reliability in Meta-analysis. Educational Researcher, 1982, 11, pp. 10-14.
- U.S. GENERAL ACCOUNTING OFFICE. IPE-83-1. The Elderly Should Benefit from Expanded Home Health Care but Increasing These Services Will Not Insure Cost Reductions, December 7, 1982.
- IPE-82-9. Problems and Options in Estimating the Size of the Illegal Alien Population, September 24, 1982.
- IPE-82-8. Lessons Learned From Past Block Grants: Implications for Congressional Oversight. September 23, 1982.
- IPE-82-2. CETA Programs for Disadvantaged Adults--What Do We Know About Their Enrollees, Services, and Effectiveness? June 14, 1982.
- IPE-81-1. Disparities Still Exist in Who Gets Special Education, September 30, 1981.
- HRD-81-43. Unanswered Questions on Educating Handicapped Children in Local Public Schools, February 5, 1981.
- WHITE, K.R. The Relation Between Social Economic Status and Academic Achievement. Psychological Bulletin, 1982, 91, pp. 461-481.
- WORTMAN, P.M. and Yeaton, W.H. Synthesis of Results in Controlled Trials of Coronary Artery Bypass Graft Surgery. Evaluation Studies Review Annual, 8, 1983. Beverly Hills, California: Sage Publications Inc.
- YIN, R.K. and Heald, K.A. Using the Case Survey Method to Analyze Policy Studies. Administrative Science Quarterly, 1975, 20, pp. 371-381.
- ZIMILES, H.M. Generalizing From Single Case Studies. Evaluation Quarterly, 1979, 3, pp. 661-678.

IE N D