

GAO

**Program Evaluation and Methodology
Division**

July 1989

**Prospective
Evaluation Methods:
The Prospective
Evaluation Synthesis**

046125/139231

Foreword

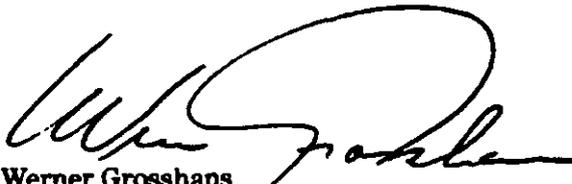
Frequently, GAO is asked to prospectively assess the implications of various policy initiatives facing the Congress. GAO, to the extent possible, assists congressional decisionmakers in their deliberative process by providing analytical information on the options under consideration. As the nature of GAO's work evolves and becomes more diverse and complex, evaluators must have the necessary tools to effectively answer and develop useful and timely responses to various type questions posed by congressional leaders.

GAO's policy guidance includes such items as methodology transfer papers and technical guidelines to provide evaluators with the tools to make informed decisions. This methodology transfer paper on "Prospective Evaluation Synthesis" provides a systematic methodology for those evaluators who may be faced with assessing future implications or outcomes for policies under consideration.

In preparing products using this methodology, evaluators must clearly identify the

- assumptions considered and data used to arrive at the information presented,
- supporting analyses to assess the options under consideration,
- external sources of information used as a basis for our findings and the reliance that a reader may place on the evidence presented, and
- analytical processes used to lead us to rank the options in the manner that we did.

Staff using this methodology must be especially careful to maintain independence and objectivity, since the reported options for projected future outcomes may subject GAO to criticism for supporting what may be perceived as partisan views.



Werner Grosshans
Director, Office of Policy

Preface

Why should a GAO evaluator read a paper on the prospective evaluation synthesis (PES)? GAO evaluators must know about methods such as the PES because the changing nature of our work requires us to be familiar with the strengths and limitations, and the applicability, of ways to answer questions about the future. The PES is one these of methods.

Prospective Methods

GAO is increasingly being asked to answer questions about the future that involve analyses of alternative proposals and projections of various kinds. To support GAO's capacity to answer these questions well, our policy and project manuals have been expanded to discuss, for example, different types of forecasting and formal modeling approaches and our standards for carrying these out. This is because systematic methods for dealing with questions about the future can be more efficient and yield sounder, better-documented answers than more informal methods do.

Many methods exist to deal with forward-looking, future-oriented questions. Collectively, they are referred to as prospective methods to distinguish them from approaches designed to answer questions about what is happening now or what has happened in the past—that is, retrospective methods.

The PES

Among the prospective methods, we have chosen to focus here on the prospective evaluation synthesis. GAO developed the PES as a systematic method for meeting congressional requests for analyzing proposed legislation and helping identify top-priority problems. Other applications of the PES might be in the analysis of recommendations in draft GAO reports and in assessing the adequacy of proposed regulations.

This paper shows how the tools of evaluation methodology can be applied in order to provide the best possible information prospectively on the likely outcomes of proposed programs. A PES may be conducted through the comparison of policy or program alternatives, although it is also useful when focused on a single policy or program. It is easiest to perform when an adequate data base already exists. Fortunately, data bases concerning proposed programs frequently do exist, primarily because problems are rarely new. Often they have been addressed by past programs whose experiences can be drawn upon for the PES.

In essence, a PES is a combination of the following activities: (1) a careful, skilled textual analysis of a proposed program, designed to clarify the implied goals of that program and what is assumed to get results, (2)

a review and synthesis of evaluation studies from similar programs, and (3) summary judgments of likely success, given a future context that is not too different from the past. In this respect, the PES resembles the evaluation synthesis approach, except that the focus of the PES is on how evaluation studies cast light on the potential for success of the proposed programs, as opposed to reaching conclusions about the actual performance of existing programs.

Three other points emerge from the experience with PES. First, the PES may call for a greater selectivity than the evaluation synthesis. The latter involves a comprehensive review of all existing studies, which can allow us to generalize quite broadly. The time-driven nature of PES may restrict it to a narrower focus and the use of strategies, such as sampling, to balance resources and the need for external validity. Second, legislators and congressional staff who have received a PES view it as a useful tool. From the congressional perspective, a PES means that expert design assistance is available for a new program at the point when it is most needed and when it can help convince others of the basic logic and likely success of the program. Third, from a public policy perspective, providing understanding ahead of time about how a program is likely to work renders an important service by validating the basic soundness of what is to be undertaken and thereby increasing its chances for success.

The Background of This Paper

This paper is based on the work of David Cordray and Stephanie Shipman on teenage pregnancy and children's programs, as well as on the work of James Solomon and Gerald Dillingham on catastrophic health care. It also follows the general lines of a paper Peter Rossi and I prepared on the prospective evaluation synthesis. It has been reviewed by all major offices within GAO and by Peter Rossi, Michael Quinn Patton, Lee Sechrest, and Joseph Wholey. Adapted from these materials by Lois-ellin Datta, it is one among the series of transfer papers PEMD issues that give GAO evaluators handy guides to various aspects of evaluation methodology and that explain specific procedures.



Eleanor Chelimsky
Assistant Comptroller General
Program Evaluation and Methodology Division

Contents

Foreword		1
Preface		2
	Prospective Methods	2
	The PES	2
	The Background of This Paper	3
Chapter 1		8
What Is a Prospective Question?		
Chapter 2		13
The Need for Systematic Methods for Answering Forward-Looking Questions		
Chapter 3		15
Prospective Methods and the Prospective Evaluation Synthesis Broadly Defined		18
	When the PES Is and Is Not Appropriate	18
	The PES and the Recommendations GAO Makes	19
Chapter 4		22
The PES: Initial Steps		23
	1. Defining the Problem	23
	2. Selecting Alternatives to Evaluate	25
Chapter 5		30
The PES: Middle and Final Steps		30
	3. The Conceptual Analysis	30
	4. The Operational Analysis	34
	5. Testing the Model	36
	6. Presenting the Results	43

<hr/>		
Chapter 6		49
Variants of the PES	Targeted PES	49
	Variants Using Other Sources of Information	53
<hr/>		
Appendixes	Appendix I: A Brief History of the PES and Some Other Prospective Methods	58
	Appendix II: Data Quality Judgment Models	61
	Appendix III: A Project Evaluation Profile	67
<hr/>		
References		72
<hr/>		
Tables	Table 1.1: Types of Forward-Looking Questions and What We Are Asked to Do	9
	Table 1.2: Features of Retrospective and Prospective Methods	12
	Table 3.1: Some Prospective Methods	16
	Table 3.2: Illustrations of Where a PES Might Strengthen Our Recommendations	20
	Table 3.3: Situations in Which a PES Should and Should Not Be Considered	21
	Table 4.1: Steps in the Basic PES Approach and Persons Involved	22
	Table 4.2: Step 1: Defining the Problem	23
	Table 4.3: Step 2: Selecting Alternatives to Evaluate	26
	Table 5.1: Step 3: Conceptual Analysis	31
	Table 5.2: Step 4: Operational Analysis	35
	Table 5.3: Step 5: Testing Key Assumptions Against Existing Evidence	37
	Table 5.4: Step 6: Presenting Results	44
	Table 5.5: Example of Presenting PES Findings	46
	Table 6.1: Targeted PES and Related Critical Issues	49
	Table II.1: Advantages and Disadvantages of Four Data Quality Judgment Models	62
	Table II.2: Example of a Fatal Flaws Analysis	65
<hr/>		
Figures	Figure 3.1: The Triad of Analysis	17
	Figure 5.1: Underlying Conceptual Model of the First Bill	32
	Figure 5.2: Underlying Conceptual Model of Program A in the Second Bill	33

Contents

Figure 5.3: Underlying Conceptual Model of Program B in the Second Bill	34
Figure 5.4: Underlying Operational Model of Program B in the Second Bill	36

Abbreviations

GAO	U.S. General Accounting Office
PES	Prospective evaluation synthesis

What Is a Prospective Question?

To understand prospective questions, it can be helpful to begin with some examples of GAO reports. GAO reported that the passage of a proposed bill, S. 581, would probably open some jobs to women that were currently closed and that might otherwise remain closed after the review required by the secretary of the Department of Defense was finished.¹ GAO also informed the Congress about difficulties with specific Food and Drug Administration forecasts. These forecasts predicted the increase in the number of medical-device problems that would be reported by hospitals and the number of agency staff that would be necessary to analyze the reports of those problems under the proposed Medical Devices Improvement Act of 1988. We concluded that these forecasts were biased and not representative of what would be generated from data obtained from U.S. hospitals in general.² And GAO found in yet another study that the Internal Revenue Service needed to review its entire revenue-estimating process in order to validate the assumptions used to better reflect actual historical trends.³

These reports illustrate the prospective, or forward-looking, questions that GAO is often asked to deal with.⁴ As table 1.1 shows, at least four kinds of forward-looking questions can be identified in reports we have issued already, requests that have been met in ways other than through reports, and our own policies regarding our recommendations.

¹U.S. General Accounting Office, Women in the Military: Impact of Proposed Legislation to Open More Combat Support Positions and Units to Women, GAO/NSIAD-88-197BR (Washington, D.C.: July 1988).

²U.S. General Accounting Office, Medical Devices: FDA's Forecast of Problem Reports and FTEs Under H.R. 4640, GAO/PEMD-88-30 (Washington, D.C.: July 1988).

³U.S. General Accounting Office, Tax Administration: Difficulties in Accurately Estimating Tax Examination Yield, GAO/GGD-88-119 (Washington, D.C.: August 1988).

⁴GAO does not normally make forecasts, although we have done so on special request (for example, in response to our assigned duties under requests related to Gramm-Rudman-Hollings). We do often evaluate the forecasting process and the methodology used. Our past work has indicated, for example, that agencies can improve forecast accuracy by using better techniques and validating predictions. The same points apply to modeling. It should also be noted that other agencies are frequently called upon for forward-looking analysis. The Office of Management and Budget requires regulatory impact analysis before any major new regulation is put into effect. And the Congressional Budget Office is required to "price out" all new legislation. Thus, there are many applications and methods in this prospective area.

Table 1.1: Types of Forward-Looking Questions and What We Are Asked to Do

Question type	What we are asked to do	
	Critique others' analyses	Do analyses ourselves
Anticipate the future	1 How well has the administration projected future needs, costs, and consequences?	3 What are future needs, costs, and consequences?
Improve the future	2 What is the potential success of an administration or congressional proposal?	4. What course of action has the best potential for success and is the most appropriate for GAO to recommend?

The use of the PES described in this paper is consistent with GAO's policy on forward-looking questions and on the methodology to be used in developing recommendations. This policy is set forth in the General Policy Manual, chapter 10.0, and in chapters 12.10 and 12.18 of the Communications Manual. These latter chapters specify, for example, the procedures that are to be followed when dealing with programs and policies under legislative consideration or recommendations asserting the possibility of budgetary savings. Particularly relevant in the General Policy Manual are the sections on formal modeling, economic optimizing, and forecasting.

1. How well has the administration projected or estimated the future needs, costs, and consequences? In responding to such a forward-looking question, GAO may need to address issues such as the following:

- How well has it anticipated, for example, revenues or staff needs or emerging problems?
- Are the methods for projection sound?
- Are the data bases reliable and adequate?
- Are the assumptions explicit?
- Are they reasonable?
- Have the projections been overgeneralized?
- Are there feasible improvements to the procedures or the reporting?
- Are better estimates, or better-reported estimates, available?

In the case of repeated or regular forecasts, we may have to examine whether the relevant agency systematically evaluates their accuracy and, if so, whether the error rates are acceptable and without bias. Further, when the administration publishes claims about the likely consequences of its own proposed activities, we may examine whether claims are methodologically sound and properly presented. And, when the administration has sought to block or prevent action, using projections

or estimates of future costs or consequences, we may determine whether these projections, too, are sound and accurately reported.

2. What is the potential for the success of a congressional or administration proposal? In answering this type of inquiry, GAO could look at the following questions:

- Given the characteristics of new or amended legislation being considered by the Congress, how likely is it that a bill will achieve its stated objectives?
- What features might be modified to improve its chances of success?
- Are there side effects or pitfalls known from past experience that could be remedied prospectively?
- When the administration initiates a new policy or new legislation by proposing a set of activities, how likely is it that these will work?
- What changes that might be made before the proposal is put into effect would better achieve the intended results?
- What unidentified dangers should be considered before action is taken?

3. What are future needs, costs, and consequences? In many areas, GAO is asked to anticipate the future in analyses such as the costs of future illegal immigration, the flow of future legal immigrants, the future costs of the AIDS epidemic, military personnel needs, and the adequacy of stockpiles of materials critical to the national defense. According to our policies, we are expected to use state-of-the-art methods for making any quantitatively based forecasts or projections and to use due professional care in applying qualitative approaches, such as expert panels. We could check on whether we have used the technically most solid procedures, fully considered alternative methods, and applied and reported properly the ranges of uncertainty inevitable in any prediction, using approaches such as sensitivity analyses to test systematically the effects of different assumptions.

4. What course of action should we recommend as most likely to succeed in addressing the problems we identify? Our policies require us to carefully consider alternative actions resulting from our findings and to weigh the costs of these alternatives and their likelihood of success before we present them as matters for consideration or as recommendations. This requirement distinguishes GAO from other congressional support agencies. They follow the policy analysis approach of presenting options but do not make recommendations. GAO goes through the analytic steps and makes its choice of the preferred solution. Further, GAO

systematically follows up and reports on the acceptance of the recommendations it makes in its reports. In this context, procedures for developing alternatives and selecting recommendations can be seen as the most crucial part of our work. Have we used the most methodologically sound procedures for identifying alternative actions and for making and documenting the analyses required in our policy and procedures manuals?

While these illustrations do not exhaust the range of prospective questions, what they say is that we are effectively in the futures business, both through the implications of our own policies and because the Congress is asking us to make or examine estimates of and projections about the future.³ This may be expected to continue (1) as the effort required for members of the Congress to push new legislation through the Congress and to amend existing legislation becomes greater, (2) as evaluations of past programs demonstrate problems that could have been prevented in existing programs, and (3) as the methodology and the motivation to get smarter about the future improve and increase. That is, we have an important role in helping prevent future problems and in helping promote greater success before action is taken and before program actors and stakeholders become entrenched.

This role complements our mission to report objectively, but in retrospect, on what is happening now and on what has occurred in the past. It is quite a different one, with distinctive methods of its own. As table 1.2 indicates, retrospective and prospective methods differ on such features as the source of the evaluation questions, where we get our information, and techniques for analyzing the evidence. Each method has its own requirements and its own strengths and limitations for our work. Those of the PES will be discussed in detail in this transfer paper. The requirements of retrospective methods have been presented in earlier transfer papers.

³The Kansas City Regional Office maintains a comprehensive review and bibliography of all GAO reports involving relatively innovative methodologies, providing easy access to these earlier applications, for job planning purposes. This list includes many reports dealing with forward-looking questions, some of which are included in our references to help illustrate further the range and history of this aspect of our work.

Chapter 1
What Is a Prospective Question?

Table 1.2: Features of Retrospective and Prospective Methods

Feature	Retrospective	Prospective
Source of questions	Criteria and issues in existing programs, regulations, and policies	Ideas and assumptions about problems, probable causes, and possible solutions
Primary sources of information	Documents, administrative data, interviews, observations, opinion surveys	Prior research, theory and evaluations, pilot or experimental tests of proposed approach; expert opinion
Primary types of analysis	Qualitative approaches to empirical data, quantitative approaches to empirical data, information synthesis in relation to program criteria and issues	Simulations, modeling, and information syntheses in relation to conceptual and operational assumptions of proposals (PES); Delphi techniques; analyses of likely impacts

We have already discussed the nature of forward-looking questions, described the types of methodological issues they raise, and summarized when a PES would and would not be appropriate. Subsequent chapters present a definition of the prospective evaluation synthesis, a detailed example of how to carry it out, and some of its variants. Special attention is given to the crucial issues of judging the quality of the information being synthesized and models for aggregating results across many prior studies.

The Need for Systematic Methods for Answering Forward-Looking Questions

In doing our work, we should use the methodology appropriate to the complexity of the question and to the level of effort required by the situation. Either overkill or underkill in design would be a mistake in job management. The first wastes scarce resources; the second fails to meet the need adequately.

For some questions and some circumstances, the use of highly systematic methods of dealing with forward-looking questions would be overkill. For example, we may be asked about one provision of proposed legislation in an area in which we have had many years of experience and in which we have published reports whose recommendations bear directly on the provision. Further, the idea may be one among several at early stages of consideration and it may be unclear that the legislation will move forward in the current session. Here, the evaluator might adequately satisfy methodological and customer concerns by drawing on our cumulative experience to discuss the issue as we have already seen it and, subject to our usual reviews for bill comments, comment informally on it. That is, we may use professional judgment and opinion.

Where the questions are controversial, far-reaching, and sensitive, more systematic methods may be called for. For example, our analyses of the savings and loan problems, and of various bailout proposals, called for more than informal methods, because of the sensitivity and long-term consequences of how this issue is resolved.

Among the advantages of using systematic methods are the following.

1. The full range of existing information may be efficiently brought to bear on the question. Rather than relying, in a somewhat happenstance way, on an individual's memory, we identify, consider, and apply the body of available knowledge to answering the question. Data that were costly to collect in the past and are still relevant but that might otherwise be neglected can be used. The risk of overlooking contradictory evidence may be notably reduced.
2. The degree of confidence we have in our own answers—whether analyses of other people's forecasts, conclusions regarding the success of proposed legislation, or our own recommendations—can be stated more precisely than less-formal methods permit. When we deal with the future, uncertainty is part of any analysis, no matter how sound, but the more precisely we state the degree of uncertainty, the more complete, and the more useful, our prediction will be. Saying, "We are 95-percent confident that the number of competitively awarded contracts will

increase by between 10,000 and 15,000 for each of the next 4 years" provides more precise information to a decisionmaker about likelihood than does the statement "More contracts will be awarded competitively in the future."

3. One method for promoting the quality of prospective work is independent replications. When we use systematic methods to review other people's projections or to make our own, we are better able to replicate the analyses and thus promote quality. That is, when independent analysts obtain the same results, confidence in findings rises. In the physical sciences, such replication in independent laboratories is often required before a result is accepted as sound. However, replication requires precision in describing and carrying out the analytic procedures. Similarly, in the social sciences, of which program evaluation is a part, using systematic methods permits replication and helps distinguish robust findings from artifacts of differences in technique.

4. Systematic methods can help us follow high-quality standards of evidence and analysis in documenting the basis for answers about the future. Much of our work requires an element of judgment. Prospective jobs inherently involve a greater degree of uncertainty than retrospective questions and, consequently, a greater element of judgment. In all such jobs, we must be scrupulous in identifying sources of uncertainty and, consequently, the need for alternatives and options. However, using systematic prospective methods can reduce the qualifications we have to add. Fewer caveats may be necessary if we apply state-of-the-art methodology.

In short, systematic prospective methods hold great promise for strengthening our ability to speak well to emerging issues.

Prospective Methods and the Prospective Evaluation Synthesis Broadly Defined

Prospective questions deal primarily with what will happen in the future. However, most prospective methods rely heavily on information about what has happened in the past, primarily empirical and evaluative data. Judgments—that is, assumptions and interpretations—enter in, particularly when we speculate on future conditions or alternative scenarios. Methodologically, answers to these questions require approaches that meet special challenges, compared with retrospective methods.

For example, almost all evaluations have to take context into account if the ability to generalize is an issue. In retrospective methods, one approach that permits generalization is simple random sampling from a properly defined population. Another such approach is stratified random sampling, in which relevant subgroups are considered, such as urban and rural or rich and poor states. Where there is reason to expect that the results of a program will depend on different circumstances—the economy, the culture, human resources—stratified random sampling is typically used. For retrospective studies, what is relevant is usually clear, and how the characteristics of entities we could sample vary is usually known.

Not so for prospective studies. What the relevant characteristics of the future will be, and how entities will vary, encompasses a wide range of possibilities. For example, whether participants in a proposed job-training program will be likely to find employment in a given period may be influenced more by overall trends in the economy than by instructional or targeting nuances. But perhaps economic conditions will be relatively unchanged, so that other characteristics of the context will be more important to consider.

Putting this distinction somewhat more technically, generalizations in retrospective studies are fairly straightforward, empirically based statements in which one moves logically from a sample to a population. Extrapolations in prospective analyses, in contrast, require one to move logically and conceptually, as well as empirically, by taking into account how a particular finding might operate under varying conditions and situations. We thus have to make economic and other assumptions explicitly; otherwise, we are implicitly accepting the continuation of the present unchanged into the future. (See Cronbach for a more detailed discussion.¹)

¹ Lee Cronbach, *Designing Evaluations of Educational and Social Programs* (San Francisco: Jossey-Bass, 1982).

**Chapter 3
Prospective Methods and the Prospective
Evaluation Synthesis Broadly Defined**

Despite this and other challenges, a set of prospective evaluation methods has been developed. As table 3.1 illustrates, these include actual, empirical, logical, judgmental, and mixed approaches.²

Table 3.1: Some Prospective Methods

Type	Illustrative technique
Actual	Experimental tests, Demonstration programs
Empirical	Simulation; Forecasting
Logical	Front-end analysis; Risk assessment; Systems analysis; Scenario building; Anticipatory analysis
Judgmental	Delphi techniques; Expert opinion
Mixed	Prospective evaluation synthesis

The prospective evaluation synthesis, or PES, is a new member of the class of prospective methods.³ It was adapted by GAO from the evaluation synthesis in order to answer questions about the future more systematically than informal methods and more rapidly than some other prospective methods such as experimental programs.⁴ (Appendix I also gives a brief history of the PES.)

Conceptually, the PES provides a way in which the logic of evaluation methodology and its procedures can be appropriately used in assessing the potential consequences either of an individual proposal or of alternative and competing policy proposals. It combines (1) the construction of underlying models of proposed programs or actions as developed by

²Economists have developed many quantitative methods for projecting the future, particularly those involving economic forecasting, modeling, and simulations. These have in common the specification of a theory (conceptual model in PES terms) of what is influencing relevant outcomes, the identification of key assumptions, quantification—on the bases of theory and past experience—of these assumptions, and running often very complex quantitative analyses of most likely outcomes under different assumptions about how the future will be similar to and different from the present and the past. For example, the Social Security Trustees Report is based on quantitative models whose key assumptions include more and less optimistic estimates of economic conditions. Our policy manuals describe some of these techniques and suggest appropriate uses. The PES can include the results of these modeling and simulation studies but differs from them in its greater reliance on prior empirical work on related programs in the past or on basic and applied research.

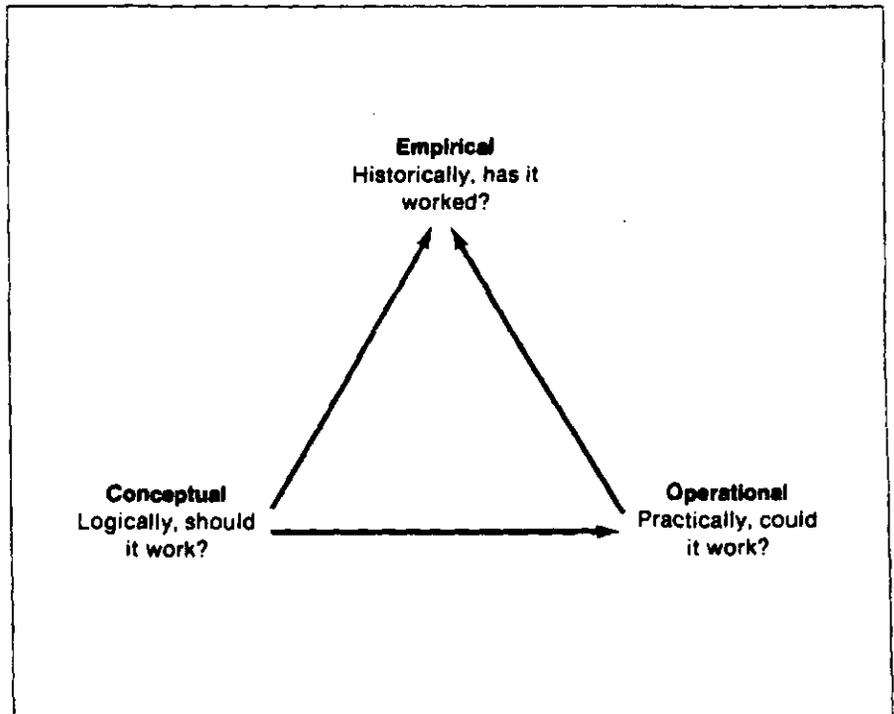
³Eleanor Chelmsky, "Federal Evaluation in a Legislative Environment: Producing on a Faster Track," pp. 73-86, in C. G. Wye and H. P. Hatry (eds.), *Timely, Low-Cost Evaluation in the Public Sector*. New Directions for Program Evaluation, No. 38 (San Francisco: Jossey-Bass, Summer 1988).

⁴U.S. General Accounting Office, *The Evaluation Synthesis*, Methods Paper 1 (Washington, D.C.: April 1983).

Wholey for evaluability assessment¹ with (2) the systematic application of existing knowledge as developed in the evaluation synthesis methodology. That is, a PES is a prospective analysis anchored in evaluation concepts. It involves logical, conceptual, and empirical analyses, taken in the context of the future.

As figure 3.1 illustrates, the conceptual analyses results help focus the operational analyses and answer the question, "Logically, should the proposal work?" The operational analyses further scope the search for empirical findings and answer the question, "Practically, could the proposal work?" The empirical analyses can open both new conceptual and operational possibilities and answer the question, "Historically, have activities conceptually and operationally similar to the proposal worked in the past?" Finally, the PES takes into account ways in which the past is and is not likely to be similar to plausible future conditions.

Figure 3.1: The Triad of Analysis



¹Joseph Wholey, "Evaluability Assessment," *Evaluation Research Methods*, L. Rutman (ed.) (Beverly Hills, Calif.: Sage Publications, 1977).

When the PES Is and Is Not Appropriate

As noted, the PES can be used either for examining an individual proposal or for comparing two or more policy alternatives. In examining an individual proposal, the PES requires a criterion, or a hoped-for good that needs to be made explicit. Developing explicit criteria is a task familiar to GAO evaluators. Nonetheless, it is often difficult, since legislative proposals can result from greater agreement on actions than on aims or goals. Assessing two or more proposals may be somewhat easier, because the points of "common cause" can serve as a proxy for the hoped-for good. Further, it is generally simpler to make comparative judgments ("Which is better?") than absolute ones ("Is it good at all? How good?").

The PES and Timeliness

Additional conditions affect the use of the PES. Although the PES has the promise of being among the most timely evaluation methods, obviously it cannot operate instantaneously. While times vary, an analysis of two or more bills might require about 3 months on the part of at least two evaluators in order to provide for adequate reviews of published and unpublished literature, consultation with technical experts, and the thorough assessment of the resulting information. However, a PES may take longer than 3 months, especially when the competing legislative proposals are quite complex, when there is little prior experience with issues, or when most of the literature is unpublished."

This time constraint indicates that a PES should be started as soon as possible after a customer's inquiry, in order to ensure that the assigned evaluators have the requisite time for their work. For less-complex issues, or situations such as analyses of possible GAO recommendations, where a separate report does not have to be written, less time may be required. As noted earlier, a greater level of effort would be allocated to controversial, sensitive, and far-reaching questions.

The PES and Data Availability

Another point affecting timeliness is that when an issue becomes extremely popular or extremely controversial in the legislature, it may happen that many different bills on the same subject are introduced within a short time. This can cause such logistical and other problems that a PES may not be the appropriate method. But if this situation

"The unpublished literature can include reports prepared under contract to the government, work in progress that has been presented as draft material or in speeches, and other relevant material that may not have appeared yet in print. Searching for these materials usually involves reviewing federal contracts and grants, contacting project managers and principal investigators, and canvassing other experts in the field.

should develop in the middle of the PES effort, then the evaluator would either have to resist expanding the scope of the study or obtain an extension of time.

As indicated above, the PES relies heavily on the knowledge—basic and applied—already produced by evaluators and researchers. The PES can be used effectively on topics for which a body of relevant literature exists. For some mature issues that have long attracted the attention of evaluators and researchers, the existing literature may be abundant, containing many studies and theories concerning the basic mechanisms involved. For others that are new or have not yet stimulated much investigation and scholarship, PES evaluators may not be able to find a great deal that is relevant.

As mentioned earlier, this outcome tells the policymakers that there is little empirical basis for their decisions. They can then judge the merits of moving ahead, not moving forward, or limiting the types of actions they take (targeting, demonstrations, and so on). It may also be an important opportunity to present to policymakers the research and data needs that would have to be filled in order to make firm judgments. The case of the PES that includes recommendations for demonstration, experimental, or pilot projects may, therefore, be relatively frequent, since such approaches can be useful alternatives to across-the-board changes in national policies.

The PES and the Recommendations GAO Makes

In many situations, a full PES would be overkill as we prepare recommendations. For example, finding a lack of accepted internal controls or finding a failure to report honestly information unfavorable to costly weapon systems leads quite directly to well-supported recommendations. In other circumstances, however, our findings are more complex, our sense of alternatives is broader, the results are more uncertain. In

Chapter 3
**Prospective Methods and the Prospective
 Evaluation Synthesis Broadly Defined**

Table 3.2: Illustrations of Where a PES Might Strengthen Our Recommendations

General circumstance	Specific example
Complex federal, state, and local relationships	<p>What is the best way for the federal government to encourage state and local governments to serve handicapped persons who are older and younger than regular school age?</p> <p>What would be the best strategy to strengthen results from federal funds in child abuse prevention?</p>
Nontrivial costs or burdens	How many Internal Revenue Service agents should be added to current staff or redirected from current tasks to go after unreported income not caught by computer matching?
Major structural or management changes	How should the responsibilities and roles of the Office of Management and Budget and other agencies be restructured to better identify low-quality surveys?
Very high national stakes are involved	What are the optimum ways of dealing with the savings-and-loan crisis?

some cases, these could be presented as matters for consideration. In others, particularly those involving controversial, sensitive, or far-reaching conclusions, our recommendations—derived perhaps through other methods—could themselves properly be the subject of a PES.

Table 3.2 illustrates some of these circumstances, which include, for example, situations in which the federal role may be relatively complex, our recommendations would pose notable costs or burdens, and major structural or management changes might be involved. In such circumstances, investing some time in a PES might permit us to be even more hard-hitting and convincing and to have a solid effect, leading in turn to greater savings and nonmonetary benefits. These and other considerations about when an evaluator should consider a PES are summarized in table 3.3.

Chapter 3
 Prospective Methods and the Prospective
 Evaluation Synthesis Broadly Defined

Table 3.3: Situations in Which a PES Should and Should Not Be Considered^a

Situation	Consideration of PES as a method	
	Probably should	Probably should not
Technical		
Data base quality	High, moderate	Low
Proposal complexity relative to time available	Complexity low or moderate and time short or moderate; or, complexity high and time long	High complexity, little time
Proposal stability	High, moderate	Low
Contextual		
Degree of federal leverage (regulations, funds)	Moderate, high	Low
National stakes	Moderate, high	Low
Consequences of our recommendations	Far-reaching	Restricted in scope

^aThese considerations apply to the PES. Other prospective methods could be useful when it would not be appropriate to do a PES.

The PES: Initial Steps

As table 4.1 shows, there are six steps in the basic PES approach, three of which closely involve the persons who request the job or are likely to use the results to make decisions—the customer. The six steps are defining the problem, selecting the options or alternatives to evaluate, analyzing the conceptual underpinnings of the selected alternatives, analyzing the operational logic of the selected alternatives, testing the key conceptual and operational assumptions against existing evidence, and presenting the results in relation to the key assumptions.

Table 4.1: Steps in the Basic PES Approach and Persons Involved

Step	Persons Involved
Defining the problem	Customer, evaluator ^a
Selecting alternatives to evaluate	Customer, evaluator
Conceptual analysis	Evaluator
Operational analysis	Evaluator
Testing key assumptions	
Check on assumption centrality	Customer, evaluator
Test against existing evidence	Evaluator
Presenting results	Evaluator

^aFor GAO, the customer is the congressional requester for the job. Other persons helpful at this step might include stakeholders and experts in the field. In the catastrophic health insurance PES, for example, health provider and consumer organizations provided useful input in defining the problem. Input is, of course, received in the context of the usual GAO guidance on ensuring our independence and objectivity.

While these steps are essential in using the PES for commenting on proposed congressional or administration actions, they also apply to the analysis of possible recommendations, with two modifications. First, generating alternative recommendations involves either usual GAO procedures or the application of techniques such as forecasts, assessment of likely impacts, and scenario-building. Second, we need to use judgment with regard to how extensively we can involve the customer in selecting options and in checking assumption centrality while maintaining our essential independence at this stage of our work.

In this chapter, we discuss the first two steps shown in table 4.1. The others are described in chapter 5. For each step, we first present what that step means, why it is important, what its role is, and the kind of activities that would fulfill the requirements. Then we illustrate how to do the step through its application in a GAO report. The applications in

both chapters center on a specific example, a PES conducted on competing legislative proposals dealing with the problem of teenage pregnancies.¹

1. Defining the Problem

Detailed Specification

Table 4.2 shows the key elements of this important first step. Here the evaluator works with the client to draw the target that the proposal is to hit, trying to be as clear as possible on the size and nature of the concerns that the proposal is intended to solve. In the PES, the evaluator is trying to see if the proposed program will work to solve not a generic problem, necessarily, but a specific one. A program that may be well-aimed at one target may miss another widely. For example, many programs can involve providing food supplements, nutrition education, and health screening. Some, however, may be aimed at solving the problem of low birth weight babies among low-income women and teenage mothers; others may be aimed at promoting age-appropriate progress in height and weight among preschoolers. Hence, the pivotal question of this first step: What's the target?

Table 4.2: Step 1: Defining the Problem

Aspect	Definition
What "defining the problem" means	Detailed specification of the concern that rules in and rules out what will be considered as part of the problem. This creates the "target" to be "hit" successfully by the proposal.
Why this step is important	Different people may define an apparently "clear" problem broadly or narrowly. Unless customer and evaluator agree on what is to be considered part of the problem, analyses aimed at determining whether proposals will work can themselves be off-target.
The role of this step	At the start of the PES, it helps determine the scope of the work and lays the foundation for the use of the results.
Activities that fulfill the requirements for this step	(1) Discussions with the customer and review of hearings (if any) on the proposal with regard to the size and nature of the problem. (2) Independent analysis of the evidence regarding the size and nature of the problem. (3) Identification of points that require agreement and decisions. (4) Discussions with the customer and others as necessary to reach closure on the definition of the problem.

¹U.S. General Accounting Office, *Teenage Pregnancy: 500,000 Births a Year but Few Tested Programs*, GAO/PEMD-88-16BR (Washington, D.C.: July 1988.)

Illustration

In 1984, there were about a million pregnancies and 500,000 births to women under 20. In response came bipartisan congressional efforts to increase the federal effort in this area. More than a score of bills were introduced into the Congress in 1986. Concerned about the best way to assess the proposed legislation, a congressional requester asked us two questions: (1) How effective had prior efforts been to address the problem? (2) What implications for structuring future legislation might be drawn from existing knowledge about teenage pregnancy?

The first step of the PES was to clarify the problem in order to focus the scope of the PES properly. In this example, the GAO staff determined that "teenage pregnancy" per se was not the problem, because policymakers were not concerned about births to married women under 20. Rather, two problems were posed in debates: (1) births to teenagers without the resources to support themselves or their children and (2) the negative health and social consequences for both mothers and infants associated with births to unwed and poor teenagers.

Faced with a subject that has been defined in more than one way, one can, of course, decide to restrict the focus of the PES to one definition or another. Following discussion with the customer, we chose to deal with both problems. In effect, this decision meant enlarging the scope of the PES to a review of the literature addressing both the prevalence of teenage motherhood and the consequences of that prevalence. Fortunately, the literature on teenage pregnancies was not ordinarily restricted to one or the other issue: most sources contained information relevant to both.

Certain topics that could have been included with the teenage pregnancy problem had received little or no attention. The excluded topics also helped define the policy space.² For example, congressional concern was expressed not about all pregnancies but only about those resulting in live births. Ignored in the discussion were the estimated 50 percent of the teenage pregnancies terminated by spontaneous or induced abortion.³ Furthermore, interest centered largely on the pregnant women and

²"Policy space" is within the boundaries of politically acceptable policies. Thus, the set of policies enclosed within the policy space of any given period consists of all the policies that are acceptable to one or another of the principal political partisans.

³It seemed obvious that a policy of promoting induced abortions as a solution to adolescent pregnancies was clearly outside the 1986 policy space.

not on the presumably teenage males who had impregnated them.⁴ Whether correct or not, the implicit legislative definition of teenage pregnancy in 1986 was as a problem primarily affecting the young women and their children.

Another aspect of defining the problem centered on who is to be considered a teenager. Clearly, women 18 or younger were included by everyone. But some discussions included all women under 25, while others restricted the definition to persons under 20. By agreement with the customer, we focused primarily on women 20 or younger.

2. Selecting Alternatives to Evaluate

The PES does not generate proposals at the beginning; that is, a proposal has already been made, and the issue is whether it is likely to hit the target, as we said earlier. Not all proposals are good or equally good candidates for a PES, however. This step does two things. First, it screens out proposals in which a PES is not the right evaluation tool. Perhaps, for example, the proposal seems to change daily or perhaps we have already reviewed similar proposals and can quickly draw on our corporate knowledge to provide comments on likely success.

Second, of the proposals for which the PES is the right evaluation tool, this step selects the optimum ideas for review. "Optimum" can include the consideration of a variety of factors. One is, of course, the specific interest of the customer. Others may include variations among proposals in cost, target groups, or the governmental means proposed—regulatory, categorical, tax policy, block grant. For example, proposals to provide long-term nursing care to the elderly could vary notably in cost, depending on such factors as the copayments required, the conditions covered, and the duration of care authorized. Some proposals could cost millions annually; others, billions. Selection on the basis of variation among the proposals could in turn reflect such factors as maximum ranges, special interests, and similarity to existing pilot work. The PES should be explicit about the basis to be used, because the choice made at the end of this step notably affects the scope of the work and the utility of the results. Table 4.3 describes this step.

⁴There was some concern in one proposal with teenage fathers, but this was never an important center of attention, although the problem could also be phrased as lack of family formation or of responsibility on the part of the young men. A PES could, at this stage, compare alternative target definitions in terms of precision, efficiency, and so on.

Table 4.3: Step 2: Selecting Alternatives to Evaluate

Aspect	Definition
What "selecting alternatives to evaluate" means	A PES usually begins with a specific proposal whose likely success is to be evaluated. What is actually evaluated may differ, however, as a result of activities conducted during this step. "Selection" means that at the end of the step, the proposal to be assessed will have been determined and alternatives, if any, will have been selected.
Why this step is important	Not all proposals are good candidates for a PES. And among the good candidates, not all may be equal in optimum use of time: it may be more useful to policy to analyze some proposals rather than others.
The role of this step	It helps ensure that the evaluator will not be wasting time, and it gives the analyses optimum value.
Activities that fulfill the requirements for this step	(1) Identification of the politically viable alternatives. (2) Screening to be sure there are no reasons, such as rapidly moving changes or an adequate body of analyses of similar prior proposals, to reject these as PES candidates. (3) Examination of the proposals that would be optimum to review in depth through the PES, according to criteria such as maximum differences in proposal characteristics. (4) Selection of the PES proposals.

Why the PES Begins With Existing Options

For any problem, a large number of potential policies and programs may be relevant. However, assessing the full range of possible alternative policies is not the concern of a PES. The PES task is constrained by two principles. (1) The task must be restricted to one that can be examined by posing the evaluation question, "Is there evidence that a particular program or policy will or will not be likely to meet its stated objectives?" (2) The PES begins with the options that policymakers are already considering in order for PES findings to be useful to them. Thus, this is a process that starts with the alternatives under consideration, then looks for any evidence concerning the potential efficacy of those alternatives, and, only if necessary, generates other options.

It is important to understand the implications of centering the PES on existing alternative policies. Another way to proceed would be to make a comprehensive review of all the research and evaluation literature relevant to the problem in question, attempting to infer the implications it has for policy and designing alternatives ourselves. However, this alternative is rejected in the PES method for two main reasons.

First, there may be only a loose fit between research findings and policy. It is possible for two reviewers to draw different policy implications

from the same research evidence.⁵ Unless some obvious logical error has been made, neither reviewer would be correct and neither would be incorrect in his or her projection of policy implications. But contradictory or even equivocal recommendations are difficult to use in decisionmaking.

Second, the PES approach allows the reviewer to make definite statements that are subject to verification. The outcome of a PES review is an assessment of whether the policy or policies under consideration are supported or not supported by the existing evidence. If a PES concludes that proposal A is justified by the evidence and some other commentator asserts that it is not, then it is possible to compare the analytic procedures used by each of the disagreeing parties to determine the position that is justified by the research evidence.

What about a situation in which none of the options already on the table is likely to work? To be maximally helpful, the PES relies on prior research and evidence as a way of refining the policy options. If the prior research did not support the options under consideration, then the PES would try to identify the policy options that were within the most realistic range of the research, when the questions were considered at appropriate levels of complexity. For example, proposed legislation on housing for physically handicapped adults might focus on increasing independence for single persons, but the literature might consistently place greater emphasis on group homes or family units.⁶

Illustration

As stated earlier, the PES is intended to weigh how closely the research and evaluation evidence supports a proposed policy or one or another of several alternative policies. In the case of the teenage pregnancy project in 1986, several alternatives could be compared. Twenty-two separate bills regarding teenage pregnancy had been introduced in the Congress, twice the number proposed the year before. For the PES, which had to be

⁵For example, given the existence of a large number of teenage pregnancies, one policy alternative would be to conduct campaigns to convince teenagers to have abortions. Another policy that fits the data is to conduct campaigns stressing sexual abstinence among teenagers. Still a third would be to provide cash bonuses and ongoing subsidies to men who would marry and support pregnant teenage women, since the underlying problem could be conceptualized as lack of family formation. None of these policies is "incorrect" in the sense of misinterpreting the basic finding of the existence of a widespread problem, but, also, none would have been relevant to the policy formation process in 1986.

⁶Care must be taken in using prior research to assess its technical quality, including the independence and objectivity of the researcher. See our discussion on recognizing threats to objectivity in our transfer paper entitled Case Study Evaluations, PEMD transfer paper 9 (Washington, D.C.: April 1987).

completed within 4 months, the selection of proposals to consider took on some importance. Clearly, full consideration of all 22 proposals was out of the question.

To aid in the selection of proposals to assess, GAO staff performed a content analysis of each program proposal, listing its program requirements, including such items as criteria for client eligibility, allowable and required services, and any required administrative arrangements.⁷ This information was presented in tabular form to facilitate identifying the elements that were similar and those that were different across proposals and how each bill resembled or differed from the others.

With a few exceptions, most of the 22 congressional bills proposed national programs of assistance services exclusively for pregnant and parenting young women. However, the bills differed on the scope of the services to be provided, the types of clients who would be served, and the administrative and financing arrangements that would be required. Therefore, rather than attempt to assess the feasibility and promise of all possible program options, the decision was made, in consultation with the customer, to focus the PES on those apparently key, congressionally relevant dimensions of difference between the proposals—that is, the choices presented to the Congress regarding scope of services, clients, and administrative arrangements. Picking alternatives that differed widely also would help in the evaluation of other proposals that differed along the same dimensions.

In order to further narrow the focus of the PES, GAO staff, again in consultation with the customer, selected two proposals that embodied these choices by differing substantially on each of these key dimensions.⁸ The first proposal was targeted to pregnant and parenting teenagers, flexible regarding the services that should be provided and administratively straightforward. Grants would be provided directly to local agencies that would design and deliver services. In contrast, the second proposal was more broadly targeted to include economically disadvantaged women up to age 25, was highly prescriptive about services to provide, and was administratively complex, requiring coordination with five other federal programs. This bill also included a proposed program for

⁷U.S. General Accounting Office, *Content Analysis: A Methodology for Structuring and Analyzing Written Material*, PEMD transfer paper 3 (Washington, D.C.: June 1982).

⁸It was understood that the first bill would be one of those evaluated, since the proposer had requested the report. The second bill was selected because of the contrasts it offered.

Chapter 4
The PES: Initial Steps

preventing teenage pregnancy, permitting the PES to address both of the problems for policymakers that had been identified at the start.

The PES: Middle and Final Steps

After narrowing the focus of the problem, we have the remaining tasks of analyzing the chosen bills in terms of conceptual and operational models of the proposed programs; identifying from those models the target populations and the program features of interest; selecting the appropriate evidence; arraying that evidence against the models to assess whether these proposed programs were likely to meet their stated objectives; and reporting the results.

3. The Conceptual Analysis

Underlying Logic

The key elements of this step are presented in table 5.1. At this point, the evaluation aims at revealing the underlying logic of the proposal: why—in theory—the proposer thinks it will work. For example, a proposal aimed at reducing urban congestion by subsidies for satellite location of offices and businesses probably is based on the assumptions that a dispersion of people is possible and desirable and that for a given community, the primary centralization comes from commercial or governmental requirements. A proposal aimed at reducing urban congestion by increasing mass transit and reducing individual parking facilities probably is based on the assumptions that dispersion of businesses attracting people centrally is not possible or desirable and that what will most motivate people to use mass transit is aversion to high parking-lot prices and having to walk long distances from parking lots to businesses, relative to cheaper, more readily accessible mass transit.

Table 5.1: Step 3: Conceptual Analysis

Aspect	Definition
What "conceptual analysis" means	Identification of the assumptions, beliefs, values, and theory underlying the proposal: why, in principle, it is likely to work or not work
Why this step is important	Two reasons. First, it helps set up criteria for figuring out what prior research or program evaluation is relevant: it is the research on the underlying theories or the program whose underlying assumptions were similar. Second, this step can identify gaps (or strengths) in logic that could lead to uncertainty (or certainty) about program success
The role of this step	In scoping, this step increasingly targets the research that will and will not have to be examined, and it increases the efficiency of the job
Activities that fulfill the requirements for this step	Content analysis of the proposed bill or idea. Graphic techniques are helpful in efficiently displaying the conceptual models and checking the accuracy and completeness of our interpretation. Can be supplemented by interviews with sponsors of the proposals or academicians who have worked on the ideas

Making the underlying assumptions or beliefs as explicit as possible helps identify gaps in the logic and helps focus the subsequent literature search on relevant prior research or program evaluations.¹ In the urban congestion example, the literature in the first instance might focus on evidence regarding the dispersion assumption and factors affecting business relocations. The second instance might focus our attention on research on individual incentives and disincentives involving money, convenience, safety, and so on in relation to using mass transit versus individual cars.

Illustration

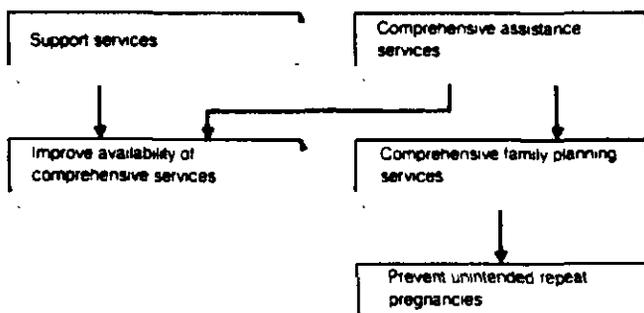
To assess both the promise and the feasibility of the two teenage pregnancy bills, it was necessary to break them down into components that could be addressed as subquestions. This required analyzing the texts of

¹ A conceptual analysis might usefully include examining the clarity and the simplicity or complexity of the outcomes anticipated. Some proposed innovations seem to be viewed as having clear outcomes, such as alleviating traffic congestion and reducing air pollution. However, it is possible that clarity may signify complexities that should not be ignored. Another aspect of the analysis of the assumptions might be the extent to which both immediate and longer-term outcomes are considered and the extent to which the links between them have been detailed. For example, a crime program may be aimed at putting more criminals in jail so that crime will be reduced. The PES could focus only on this immediate criterion, but it might be useful to consider more indirect consequences, such as increasing the size of the incarcerated populations, with the costs and complexities this will entail. Thus, in the conceptual analysis stage, a PES can inquire into these matters, finding out what potential problems have been recognized by proponents and opponents and, when the evidence is examined in a later stage, whether the arguments advanced to deal with the problems seem adequate.

the two bills to develop two types of model for each proposal: (1) a conceptual model and (2) an operational model. The strategy here was similar to that of developing an evaluation design, except that a PES reviews existing evidence instead of collecting new data.

The conceptual models would answer the following questions: What was the problem to be addressed? What was the treatment? (Or what actions would be brought about by the program?) And what was the intended outcome of those actions? Figures 5.1, 5.2, and 5.3, from GAO's report, contain the results of that disaggregation.² These models helped determine the previously studied programs that should be considered similar to those proposed and the outcomes that should be examined when judging their effectiveness. As can be seen from figure 5.1, the first bill had the objective of reducing the number of unintended repeat pregnancies, while the second bill, whose structure is shown in figures 5.2 and 5.3, articulated a fairly detailed theoretical model. It proposed to aid young mothers to avoid welfare dependence by allowing them to complete school and gain employment and, thus, the bill specified additional intermediate objectives.

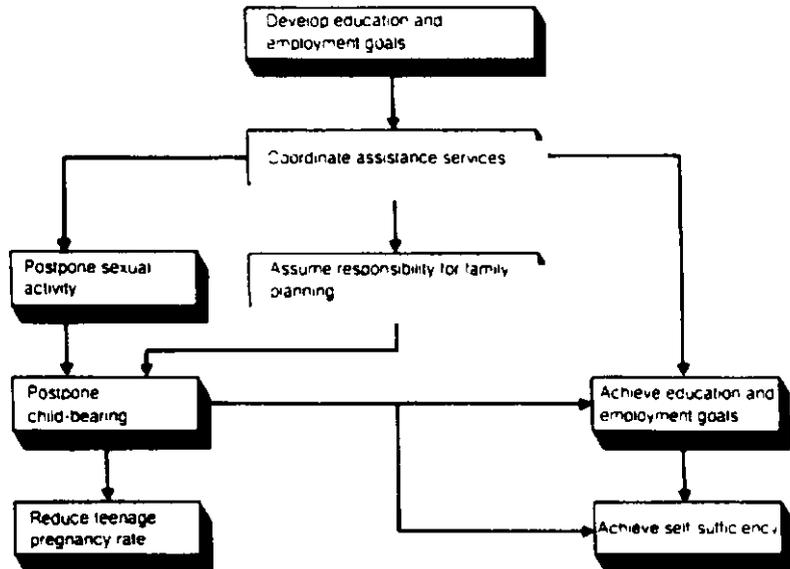
Figure 5.1: Underlying Conceptual Model of the First Bill



Source: U.S. General Accounting Office, *Teenage Pregnancy: 500,000 Births a Year but Few Tested Programs*, GAO/PEMD-86-16BR (Washington, D.C.: July 1986), p. 16

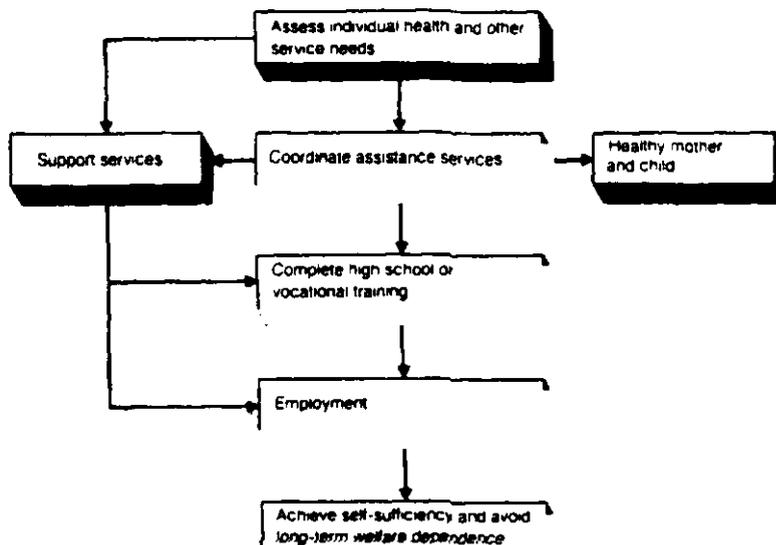
²U.S. General Accounting Office, *Teenage Pregnancy: 500,000 Births a Year but Few Tested Programs*, GAO/PEMD-86-16BR (Washington, D.C.: July 1986).

Figure 5.2: Underlying Conceptual Model of Program A in the Second Bill



Source: U.S. General Accounting Office, *Teenage Pregnancy: 500,000 Births a Year but Few Tested Programs*, GAO/PEMD-86-16BR (Washington, D.C.: July 1986), p. 17.

Figure 5.3: Underlying Conceptual Model of Program B in the Second Bill



Source: U.S. General Accounting Office, *Teenage Pregnancy: 500,000 Births a Year but Few Tested Programs*, GAO/PEMD-86-168R (Washington, D.C., July 1986), p. 16.

4. The Operational Analysis

Underlying Operations

The operational model of a proposed program shows how to accomplish the goals of the program. Like the conceptual model, it is constructed by a careful textual analysis of the legislation, but it answers the following question: Who is to be served, by whom, and under what financial and operational arrangements or constraints? An operational model defines the target populations, the intended service providers, the funding sources and amounts, and the administrative structures that should be the focus of the PES.

The details of the fourth step—operational analysis—are described in table 5.2. Here the emphasis is not on the “why” of the proposal. It is on the “how” of the proposal: how the proposed program would be carried out and how it would operate. The methods of operations research come into play in this step. The proposals are analyzed to determine who is doing what, when, and under what circumstances to whom in order for

the proposal to be carried out. This step can identify the operational complexities (or simplicities) in the proposal, the number of decisionmakers, and how contingent the final results will be on the agreement and coordination of many (or relatively few) actors.

Table 5.2: Step 4: Operational Analysis

Aspect	Definition
What "operational analysis" means	Identification of the mechanics of the proposal, how it is supposed to be carried out
Why this step is important	Two reasons. First, it sets up criteria for determining the relevant prior research or programs or the prior experience with operations similar to that of the proposal. Second, this step also can identify gaps (or strengths) in the proposed procedures that could lead to more or less certainty about program success
The role of this step	It sets limits within which the search for relevant prior research or program evaluations takes place, increasing job efficiency and completeness
Activities that fulfill the requirements for this step	Operations analysis of the proposal. The techniques of operations research—using the content of the proposal to identify the design elements—are appropriate. Graphic presentation of the operation helps check the accuracy and completeness of our interpretation. Interviews with proposal sponsors or developers provide final assurance of the operational model's quality

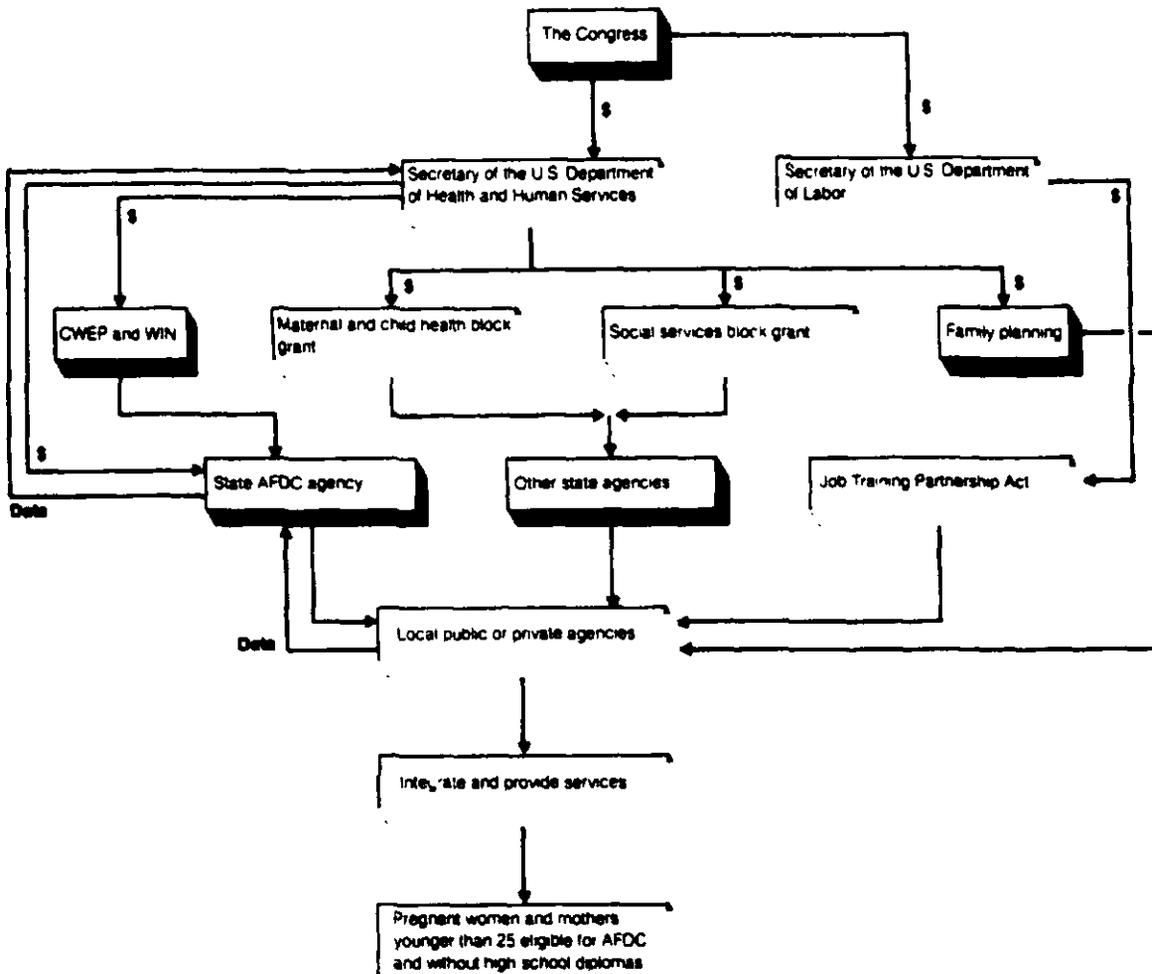
The analysis in itself can reveal likely sources of success or failure for the proposal: gaps, for example, in authority for making decisions or assumptions about the availability of resources other than those to be provided directly through the proposed program. The operational analysis also serves another function: it focuses the literature review on the relevant operational issues that could affect the success or failure of the new program. Finding, for example, that the operation of one proposal would require *establishing local stakeholder groups while that of the competing proposal would involve using elected officials* would turn attention to relevant prior experience of the efficiency and effectiveness of these contrasting modes of program management and control.¹

Illustration

Figure 5.4 shows the operational model constructed for the second teen-age pregnancy bill.

¹Other aspects of operations could be considered at this step, such as whether the process to be set into motion is fast-moving or slow-moving and whether it is easily reversible. For example, building an interstate highway system is inherently slow-moving, and the decisions could be fairly easily reversed. However, a decision to legalize the sale of assault rifles may be fast-moving and, at least in terms of consequences, may not be easily reversible.

Figure 5.4: Underlying Operational Model of Program B in the Second Bill



Source: U.S. General Accounting Office, *Teenage Pregnancy: 500,000 Births a Year but Few Tested Programs*, GAO/PEMD-86-16BR (Washington, D.C.: July 1986), p. 19.

5. Testing the Model

Two Substeps

Testing the model involves two substeps. The first substep—checking the centrality of the assumptions to be examined in depth—means

reviewing with the customer the assumptions selected as the focus of the review of prior evidence. The conceptual and operational models usually involve many steps, and it may not be valuable to delve into them all. The evaluator selects those that seem to be most pivotal or to offer the most useful contrasts between competing proposals. Discussion with the customer (or the developers of the idea or knowledgeable academic sources) is a final check that the best points of entry into tests of key assumptions have been selected.

The second substep—testing key assumptions against existing evidence—is summarized in table 5.3. This step uses the evaluation synthesis methodology but with two differences. The first difference is that what is relevant has been determined through the process of specifying the conceptual and operational models and through checking the importance of the assumptions to the customer. A second difference is that the evaluations are synthesized with respect only to the chosen assumptions.

Table 5.3: Step 5: Testing Key Assumptions Against Existing Evidence

Aspect	Definition
What "testing key assumptions against existing evidence" means	A complex body of evidence from prior research and program evaluation is collected, and the key conceptual and operational assumptions are compared with the findings from prior studies to determine the likelihood of new program success
Why this step is important	The conceptual and operational analyses can reveal gaps in logic that are likely to affect program success. This direct test against prior experience, however, is the major criterion for deciding whether the idea will work. If relevant prior research and experience indicate that the key assumptions have worked in the past, then, if conditions are similar, they are likely to work in the future (similarly, if they have not worked in the past and conditions are similar, they are not likely to work in the future)
The role of this step	It completes the triad of analyses (conceptual, operational, empirical) to give a conclusion on the proposal's success that is as solid as possible
Activities that fulfill the requirements for this step	(1) Complete identification of relevant prior research and program evaluation, (2) assessment of the quality of this evidence, (3) synthesis of credible findings. The evaluation synthesis method is applied. Systematic tabular or graphic comparison of the evidence against each key conceptual or operational assumption adds the efficiency and completeness of this analysis. Thus, techniques of meta-analysis and multiple case study comparisons are applicable

In table 5.3, step 5 is described as completing the triad of analyses. As noted earlier, a central methodological point in the PES is that the results of three different types of analyses—conceptual, operational, and

empirical—are all compared and otherwise taken into account in reaching conclusions, thus strengthening what can be said with some confidence about the future. When all three approaches give the same answer, we can be more confident about its soundness. When they differ, as seen from conceptual, operational, or empirical perspectives, we must qualify our results in terms of that lack of reinforcing agreement. Finally, we need to consider ways in which the future may differ from the past, identifying, for example, more or less optimistic scenarios for relevant factors. Where the future is likely to be similar to the past on key dimensions, we can have more confidence about the appropriateness of the PES to judge the likely success of proposals. As the scenarios differ from past or present experience, our certainty necessarily decreases, although we can still specify conditions under which a proposal is more or less likely to work.

Illustration

The review of evidence in the teenage pregnancy example we are following started with a basic question: How many people would be eligible for the programs in the proposed legislation? This was a relatively easy question to answer because of the excellent demographic data collected by the Bureau of the Census and the National Center for Health Statistics concerning the number of teenage women at present, in the past, and in the near future, as well as birth statistics. Less definitive data were available on births by socioeconomic level, although several surveys were the basis for our estimates. The next sections give further detail for the illustration.

Estimating Target Population Size

Good estimates of the size of the target population for a proposed program are important for projecting program costs. However, the target population is not identical to client population, since few programs are ever able to reach all the eligible members of a target population. In general, the more complex the eligibility requirements are, the less precise the estimates of client participation can be. An important data source can be experience with similar existing programs. If the clients of an existing program are identical (or nearly so) with the target population of some proposed program, a good basis for such estimates can be the existing program's current number of participants. For example, data from states with catastrophic illness insurance programs provided important insights for the PES on the proposed national system. More usually, it is necessary to synthesize population estimates, combining numbers from census and administrative data, for example, with information from population surveys and research data on the degree of association between eligibility characteristics.

In this illustration, no existing program served all pregnant and parenting teenagers. It was necessary to rely on published tables from the National Center for Health Statistics on the characteristics and numbers of women giving birth each year by age, marital status, years of school completed, and number of previous births. It was possible to add up the number of first births to women under age 18 over several years to calculate the number of young unmarried mothers who constituted the target population. However, this target population is too inclusive, since some of the young mothers are not poor and, hence, would be ineligible for program participation under the first proposal.

Unfortunately, the National Center for Health Statistics collects no information on the incomes of mothers. To estimate the number of poor young mothers required using sample survey data and applying survey findings to the vital statistics. Of course, the potential client populations of proposed programs are always problematic. Clients should not exceed in number the total target population, but participation rates can vary considerably, as suggested earlier.⁴

Some information on participation rates can be obtained by examining existing programs of a similar nature. The next task in the PES was to identify the existing federal programs with related objectives and target populations. This is important for several additional reasons. First, there is always an implied alternative to the proposals being considered, and that is the status quo, consisting of all the federal programs already in place. Second, in this instance, information on existing programs would also address the feasibility of both the proposed coordination of existing services and the proposed funding level.

For example, if a proposed program relies on coordinating services provided under another program or programs, whether those services are in fact available becomes crucial information. It is crucial because if the services funded by these other programs are not available, or if the providers are already operating at capacity and cannot take on new clients, then the new program has to find another way of providing those services, and it will need additional funds to provide them. Further, if existing services were apparently underutilized, a new program might not be needed.

⁴The number of clients should not exceed the total target population unless, of course, the existence of the program produces an increase in the target population.

The review of existing programs provided little information on what could be expected as participation rates in either of the two proposed programs. The main reason for this disappointing outcome was that the existing programs were, with one exception, not exclusively targeted at teenage pregnancies but included other target groups as well.

Finding the Studies

The next task was to conduct a search for all studies published in the recent past (5 years, in the illustration) that evaluated pregnancy prevention programs and comprehensive service programs for pregnant and parenting young women.⁵ The search included formal publications, such as professional journals and monographs, as well as computerized data bases, usually containing bibliographic citations, abstracts, and informal (or so-called fugitive) publications, including reports of limited circulation and monographs. It is especially important that every effort be made to (1) obtain coverage of the last category as wide as possible, since informal publications often contain the latest studies, and (2) collect and note negative findings, since studies showing positive results are more likely to be published than those that do not.

To obtain information on all relevant evaluation studies, it is usually necessary to rely on personal contacts with knowledgeable persons. This can normally be accomplished by sending out lists of publications already located and asking for the list to be supplemented by other publications known to the experts.

In this case, all the studies—whether containing outcome evaluations or not—were reviewed for analysis of program costs, their sources of funding, and implementation problems. The information on where individual projects gained their funds also augmented the information on existing programs and services collected earlier on the federal level. Articles about program failures can provide invaluable information that gives balance and perspective to the information gained from successes. For example, they may give clues as to the staff, public relations, client recruitment, or support services required for the proposed programs to operate as intended.

⁵Publication dates in journals can follow the time of data collection by several years. The studies covered up to a decade of research previous to the time of the PES. This time restriction recognizes that applied social research has only recently been used extensively in the evaluation of programs and that the credibility of remote data is slight for reason of age alone. For example, data on the effectiveness of the Great Depression programs, such as the Civilian Conservation Corps, are not likely to be viewed as relevant to similar contemporary programs. However, for some programs, time restrictions may be much looser. For example, in a PES on job training, studies that are a decade or two old may not be seen as irrelevant, especially if studies over time are quite consistent in their findings.

Special attention was paid to publications containing outcome evaluations. Each publication was read carefully to ascertain how closely the programs in question resembled the proposed programs, and a succinct summary of each program was prepared. The outcome variables used in the evaluation were noted separately, particular attention being paid to the quality of the "impact assessment" data. The end result of this careful examination was a profile for each evaluated program, recorded in tabular form, containing the crucial information on program description, outcomes, and ratings of data quality. Appendix III gives an example of one such profile.

As mentioned earlier, it is important to bring to bear on the literature the same conceptual framework used in examining the proposed programs. For each article describing a program and its evaluation, a profile form was filled out, characterizing that project's clients, services, and administrative arrangements. The categories were the same as those developed in the analysis of the two legislative proposals, in order to ensure that the derived information was directly relevant to the consideration of the proposed programs.

Quality Assessment

The most technically demanding aspect of the review of each evaluation was assessing the quality of the information. Since this task is essentially identical to that confronted in the evaluation synthesis, GAO staff borrowed from criteria employed in previous syntheses. Each evaluation outcome, as defined by the conceptual models of the programs being examined, was treated separately. The evidence on each objective was rated separately. An evaluation might provide evidence of adequate quality on one of the outcomes of interest but not on another, because, for example, of the use of different data collection methods. Other outcomes that were not of direct concern in the conceptual models of the programs under scrutiny in the PES were also noted, along with assessments of the quality of the evaluation evidence used.

Criteria

The quality-rating criteria used in the assessment of effectiveness evidence have to be tailored to some extent to the issues involved in the PES. Nevertheless, the criteria are largely the same from PES to PES. In this case, criteria centered primarily on the internal validity of the

research design used in arriving at effectiveness estimates.¹⁴ Most of the evaluation studies had used longitudinal comparison group designs, making the composition of the comparison group critical. (Appendix II presents more detail on criteria.)

The criteria included (1) appropriateness of the comparison (or control) group; (2) sample size adequacy, including attrition among clients and comparison group; (3) standardization of data collection, including measures of data reliability; (4) validity of measures used to represent outcome variables; and (5) appropriateness of statistical methods used, especially those used to enhance the internal validity of effectiveness estimates, by testing for competing explanations of estimates.

The assessment of data quality requires some training in evaluation design, measurement, and statistics as well as some understanding of the substantive area. Several readings are often required. For example, sometimes the fact that there is anything wrong with a particular measure of a variable is not obvious until another study has been examined that is more careful and accurate in its measurement strategy. It may often be necessary to read the set of evaluation studies several times before a final quality reading can be arrived at.

Reliability

As in other rating tasks, it is necessary to test the reliability of the ratings (that is, their replicability, or likelihood that other reviewers will reach the same rating conclusions) by ensuring that there will be at least two readers for at least a subset of evaluation studies. If a subset is used, the reliability check ratings should be done early, midway, and late in the coding process to avoid rater-drift and general fatigue. Discussion among raters concerning their disagreements on the subset often

¹⁴Internal validity refers to the attribution of cause and effect; external validity, to the ability to generalize. An "ideal" design would offer strong evidence that effects, if any, stemmed from the program (or event being studied) and would be obtained from groups and in situations as similar as possible to the whole range of circumstances in which the program was being applied. Further, this ideal design would be appropriately sensitive, able to detect effects of a size believed worth the costs of the program. Some experts believe the controls necessary for internal validity severely limit external validity, and they argue that for policy purposes, external validity, with its implications for extrapolation, is most important in judging quality. Other experts are more sanguine about optimizing both or place heavier emphasis on internal validity. We thought the question with top priority for this particular PES was evidence of any effects, and so we focused on that aspect of design. For some other PES, different criteria might be weighted more heavily, a point discussed in more detail in appendix II.

brings to light critical characteristics of studies that were not immediately discerned.

Aggregation

Although it is possible to arrive fairly easily at a reliable and credible rating for each criterion, arriving at an overall quality rating is usually more difficult. Much of the problem encountered in developing overall assessments for the teenage pregnancy study arose because many reports did not provide information with which to judge the adequacy of the evaluation on one or more of the criteria. In the absence of direct evidence, it is possible to judge the evidence only questionable, unless some other piece of information suggests that the absence of information stems from some serious flaw.

In addition, many evaluation studies provide data on several evaluation outcomes, each outcome varying in the quality of evidence presented. It would be a mistake to discount entirely a study that contains an acceptable evaluation of one outcome and a poor evaluation of another. For these reasons, rather than overall quality ratings for each evaluation study, each outcome was presented separately along with quality assessments of each outcome.

6. Presenting the Results

Product Type

Presenting the results of a PES differs from presenting the results of an evaluation synthesis. In a PES, the underlying conceptual and operational models have to be identified, the key assumptions have to be highlighted, and the evidence has to be summarized in relation to these assumptions. In contrast, an evaluation synthesis arrays the evidence in relation to the questions to be answered, and the underlying models need not be explicated. Table 5.4 summarizes the elements of step 6.

Table 5.4: Step 6: Presenting Results

Aspect	Definition
What "presenting results" means	Presentation of the conceptual and operational models (usually in graphic form) and of the results of the comparison of key assumptions and evidence concisely and clearly
Why this step is important	The PES involves an uncommonly detailed analysis of a proposal. The credibility of the results depends in part on the reader's being able to follow the PES procedures easily and to see in detail how the findings have developed
The role of this step	Promoting credibility and making our conclusions as simple, clear, and accessible as possible
Activities that fulfill the requirements of this step	Development of appropriate graphics and tables; preparation of necessary technical appendixes (for example, details on procedures used to rate the quality of prior evidence and to aggregate findings)

Table 5.4 emphasizes the value of tabular and graphic techniques. The result of a PES might look more like a briefing report than a chapter report. This would vary, of course, in terms of length, depth, whether or not recommendations are provided, and our other usual criteria for deciding on product type.

Illustration

The results of outcome evaluations are typically presented in tabular form, as shown in table 5.5 on page 46, where some of the findings from the teenage pregnancy PES assessment are presented. Table 5.5 was designed to draw the reader's attention to several different things. Across the top are the explicit objectives of the legislation plus some others that were found to be important in the field. Along the side are program types generated by clustering studies according to similarity with regard to the services they provided. In the body of the table are the descriptions of the studies' comparison groups and the results, expressed as whether the program group "did better" than the comparison group at a statistically significant level. The boxes represent findings we considered to be most methodologically credible. All this information was transcribed from the rating sheets.

A summary table such as table 5.5 provides information on how many studies addressed each particular outcome, how much of that data is credible, and the types of programs that had effects compared to other conditions or programs. The information is presented in narrative rather than numerical form. While this was an appropriate way to present the findings, an alternative would be to report effect sizes. Where there are quite a few studies with relevant results—and particularly where the programs' clients can be grouped by factors such as age, race, education,

and family income, which would be expected to influence the outcome variables—a quantitative presentation can be efficient and effective.

Note that in table 5.5, comparison groups are described in detail. This is also critical information, because some evaluations compared the program to nothing more than ordinary prenatal health care, while other studies compared their program with one that was only slightly different from it. The presence and absence of effects under these types of test condition are thus difficult to assess. That is, a high-quality test of a program includes the essential elements of a high treatment strength and a strong basis for causal attribution. This point became a conclusion of our illustrative PES: few programs had been adequately tested, a wide variety of programs appeared successful, and both comprehensive and less-comprehensive programs appeared to have been successful.

More specifically, the findings of the illustrative PES with regard to the requester's questions were summarized as follows.

1. The pattern of credible results showed no clear preference between the two proposed programs. A variety of past programs appeared successful, but there was little information on the components that were responsible for their apparent success. And there was no convincing evidence that the most comprehensive service packages were more effective than the least comprehensive.
2. Implementation analyses suggested that there were certain avoidable operational problems associated with the proposed administrative structures. For example, program administrators as well as evaluators frequently mentioned complex coordination arrangements as a significant obstacle to program success.

Table 5.5: Example of Presenting PES Findings*

<u>Table III.3</u>			
<u>Results of Service Programs</u>			
<u>Program type</u>	<u>Comparison^a</u>	<u>Health and delivery</u>	<u>Fertility</u>
Academic and vocational services, personal counseling, case management, health care, and parenting education	Similar teenagers delivering in same hospital receiving only prenatal care (Johns Hopkins Univ., C7)	Reduction in preeclampsia, premature births, and perinatal death; no change in % low birth weight	
	Perinatal patients not continuing (differences not tested)		Pregnancy rate lower after 1 year
	Teenagers in similar cities, no special program (Project Redirection, C12, C13)		Reduced pregnancy at 1st, not 2nd, year; no general change in birth control
Alternative school: personal counseling and health and parenting education	Pregnant students who remained in regular school (Continuing Education, N.C., C9)	No difference in prematurity	

Chapter 5
The PES: Middle and Final Steps

Education

Employment

Welfare

Higher graduation rates at 30 months; slightly higher attendance rates

Higher employment during 2 years after delivery

Lower participation at 30 months

No difference in graduation rates at 24 months; attendance rate higher after 1 year, not 2 years

No difference in employment rate; greater work experience in 1st year, marginally greater at 2nd year

Number of semesters of schooling completed higher at 1 and 2 year follow-up

Fewer graduated in program year; no control for age or grade level

*This is one segment of a longer table. It illustrates an intermediate summary of findings by offered services. The table included verbal and graphic material. The "C7" and the other code numbers in the second column refer to full bibliographic data for each comparison in the teenage pregnancy report from which we have taken the table

Source: U.S. General Accounting Office, Teenage Pregnancy: 500,000 Births a Year but Few Tested Programs, GAO/PEMD-88-168R (Washington, D.C.: July 1988), p. 47

3. Therefore, if the Congress wanted to initiate a nationwide program, then the administratively simpler model might have a greater chance of success. However, we concluded that the evidence was most consistent with initiating a large-scale demonstration program that would systematically test the feasibility, costs, and benefits of different approaches to reducing teenage pregnancy.⁷

In this particular instance, the conclusions did not clearly favor the legislative proposal that was prescriptive (given the lack of strong evaluative knowledge) and relied on existing services (given past experience with complex coordination processes.) In addition, the smaller, more flexible proposal had to take into account the need to develop information about which strategies work with which teenagers. Thus, no clear advantage adhered to the one compared to the other. This is not always the case for a PES and, in fact, did not occur in another example of the method dealing with catastrophic health insurance proposals.⁸ However, the importance of the teenage pregnancy example is real in that it saved taxpayer resources, since neither proposal had been introduced in a form that was likely to succeed.

⁷Two options were suggested as consistent with the analyses. (1) If expansion of available services is wanted, then it would make sense to target services to the teenagers who are at highest risk—young and unmarried teenagers—to allow flexibility in the type of services provided and to have a simple administrative structure. (2) In an alternative to a program of expanded services, the federal government could take the role of promoting innovation and ensuring both sound comprehensive evaluations of the innovations and dissemination of the programs (or their components) that have been shown to work.

⁸U.S. General Accounting Office, *Medicare: Catastrophic Illness Insurance*, GAO/PEMD-87-21BR (Washington, D.C.: July 1987). In this report, we looked at six legislative proposals for protecting Medicare enrollees from the financial hardships that often accompany catastrophic illness. Our review, and in-depth analysis of two of these six, determined that while protection would increase, some gaps would remain. We further identified issues requiring additional consideration, such as coverage of prescription drugs. Our conclusions played a significant role in both hearings and the subsequent configuration of the act.

Variants of the PES

Several variants of the PES are possible. They are of two types. The first variant derives from targeting the PES: customer interest in special aspects of a proposal. The second variant involves combining the PES with sources of information other than prior written evaluations. Using multiple methods would, of course, notably expand the range of the PES. Further, it is typical of designs for many of GAO's important or controversial jobs that we use several methods, so that the limits of one are offset by the strengths of another.

Targeted PES

The basic model of the PES we described in chapter 3 is appropriate when relatively well articulated proposals have been developed. However, the PES can be helpful in other, more limited situations, as when a problem is being defined or when costs are of particular interest. In essence, aspects of the full PES discussed earlier become the target of more limited work. Table 6.1 summarizes some of the variants of the PES.

Table 6.1: Targeted PES and Related Critical Issues

Target	Critical issue
Problem definition	Determining the fit between the perceived problem and legislative proposals
Problem characteristics	Assessing data quality and narrowing or resolving contradictory estimates
Relation of proposal to prevailing scientific models	Clarifying underlying assumptions
Assessing projected costs	Checking sensitivity of projections against varying assumptions

The PES and Problem Definition

For many issues that come before a legislative body, some critical problem has been identified by the proposers of legislation, along with suggested measures expected to resolve the problem. If the problem is a major one, it is rare that only one piece of legislation will be proposed. Even in such cases, as already noted, every proposal has an implicit alternative—namely, not to enact any legislation at all. In any case, before a judgment can be made about whether the proposed measure will resolve the problem, it is important to be clear about exactly what the problem is.

Proposed legislation designed to address a particular problem is necessarily based on some definition or understanding of the issue involved. For example, two contending legislative proposals may both be addressed to the issue of homeless persons, one identifying the homeless as needy persons who have no kin upon whom to depend and the other defining

homelessness as the lack of access to conventional shelter. The first definition centers attention primarily on the social isolation of potential clients, while the second focuses on housing arrangements. It is likely that the ameliorative actions that follow will be different, as well. The first might emphasize a program to reconcile estranged persons with their relatives, while the second might imply a subsidized housing program. Thus, the two definitions lead to different proposals.

Especially critical in problem definition is the fit between what is perceived to be the problem by those who have pressed for attention to the issue and the definition in the legislative proposals. In this connection, the PES evaluator would ordinarily refer to legislative proceedings, including committee hearings and floor debates, journals, newspaper and magazine editorials, and other sources in which discussions of the problem may appear. The purpose of this review of sources is to examine how the problem has been formulated and to state as clearly as possible the range of politically acceptable alternatives.

Problem Characteristics: Density and Distribution

To design a public program properly and to project its costs reasonably well, good information is needed on the density, distribution, and overall size of the problem. For example, in providing financial support for emergency shelters for homeless persons, it would make a significant difference if the total homeless population is 2.5 million or 250,000 (both estimates have been advanced). It would also make a difference whether the problem is located primarily in central cities or can be found in equal densities in smaller and larger places.

An identified problem is often a complex mixture of related conditions; for planning purposes, specific information is needed about that complexity. In the example of homelessness, the proportions of the homeless suffering from chronic mental illness, chronic alcoholism, or physical disabilities has to be known in order to appropriately design the relevant mixture of programs.

It is much easier to identify and define a problem than to develop valid estimates of its density and distribution. For example, only a small handful of battered children may be enough to establish that a problem of child abuse exists. However, to know how great a problem is and where it is located geographically and socially involves detailed knowledge about the population of abused children and its distribution throughout the political jurisdiction in question. Such exact knowledge

is ordinarily much more difficult to obtain with the kind of precision that may be needed.

To collate and assess whatever information exists on the issues in question, evaluators need to use what they have learned from the literature (consisting of government reports, published and unpublished studies, and limited-distribution reports) and their understanding of the designs and methods that lead to conclusive results. Equal emphasis is given in the last sentence to "collate" and "assess." Unevaluated information can often be as worthless as no information at all.

For some issues, existing data sources may be of sufficient quality to be used with confidence. For example, an issue on which measurements are routinely taken by either the Current Population Survey or the decennial census is typically an issue about which accurate and trustworthy knowledge ordinarily may be obtained from those sources. Data from some other statistical series, such as those published by the Bureau of Labor Statistics, also fall into the trustworthy category. But when we deal with data produced by other sources, it is necessary to examine with care how the data were collected.

A rule of thumb is that for any subject, existing data sources provide contradictory estimates. But even chaos can sometimes be reduced to some order. Seemingly contradictory data on the same topic collected by opposing stakeholders can be especially useful for assessment purposes. For example, both the Coalition Against Handguns and the National Rifle Association have sponsored sample surveys of the U.S. population concerning their approval or disapproval of gun-control legislation. Although the two reports issued by the coalition and the association differed widely in their conclusions, the one finding much popular support for more-stringent gun-control measures and the other the opposite, a close inspection of the data showed that many of the specific findings were nearly identical in the two surveys. The findings upon which both surveys substantially agreed can be regarded as having the greater credibility.

Relating Proposal Models and Prevailing Scientific Models

Whether explicitly intended or not, legislative and other proposals are based on some set of ideas or models of how the problem in question may have arisen and how it is currently sustained. For example, one welfare reform alternative suggests extending to all states the coverage of public welfare to intact families with unemployed parents in order to reduce the number of households headed by women. This proposal may

be based on a model that sees current welfare policies as penalizing marriage, since benefits to a woman and her children would stop upon marriage.

An alternative welfare reform proposal might suggest that benefits be continued upon marriage but reduced by some proportion to avoid subsidizing parasitic marriages. Both proposals involve extending benefits to intact families, one to support such families when both parents are unemployed and the other without regard to the employment status of a new parent. Each proposal is based on different models of how payments might affect marriages in households headed by women. In the first case, the proposal is based on the idea that women will avoid marriage to unemployed men because they would lose their benefits, and it ignores the effects that marriage to an employed man would have. The second proposal is concerned with the possibility that the continuation of benefits after the marriage of a woman head of household might render the woman susceptible to marrying a man who was primarily interested in sharing her benefits.

Both proposals are based on models that stress the role of economic incentives in marriage formation, a topic that has received considerable attention in microeconomic theory, econometric research, and social psychology and sociology. An appropriate tactic for the PES would be to review this literature, seeking to establish two things: (1) the extent to which experts agree and (2) the existence of empirical evidence concerning the intended effects of either proposal. A thorough review of the existing literature accompanied by consultation with subject-matter specialists and knowledgeable practitioners could determine that one of the proposals has more support than the other, that there is as much evidence for one as for the other, or, alternatively, that neither proposal has much positive backing in research and experience.

An important opportunity is presented when a PES finds that there are very few or no previous evaluations that are relevant because the proposed program is a notable departure from programs evaluated in the past. A clear message can be sent to decisionmakers that their proposals go far beyond firm knowledge and are, hence, subject to a more-than-ordinary risk of failure.¹ This advice need not be an admonition to stick to the programs of the past. For example, the advice may be to fund

¹We would need to take into account that not acting carries its own risks of failure. For example, while we may have little certainty about effective AIDG prevention measures, not making the best efforts we can also incurs risks.

demonstration projects incorporating the new proposals rather than to fund fully operational programs. Pointing out areas on which existing knowledge has nothing to say may be as important for the avoidance of public policy failures as gathering a rich harvest of firm knowledge.

Assessing Projected Costs

Legislative proposals are often accompanied by projected costs. In fact, all the bills that are reported out of committee include a Congressional Budget Office cost estimate. Although any projection can be easily upset by subsequent actual experience, it is usually possible to make a viable assessment concerning whether projected costs are based upon reasonable and likely assumptions. For example, the projected cost of a proposed measure that would subsidize flood insurance for structures built on flood plains can be profoundly affected by assumptions made about the number of structures that are to be covered and the participation rate among potentially covered households.

If the flood plains are defined as areas within a 100-year flood zone—where a major flood is expected at least once every century—coverage will be greater but flood incidence will be lower than if the limits of the flood plain were defined as a 20-year flood plain. If all the applicable property owners participated, anticipated costs might be more than if the participation rate were much lower. But there are also other complications that affect cost. If only the property owners who were close to the source of floods signed up, then the subsidy costs might be less than if participation rates were more uniform over the flood plain.

A PES can help assess cost projections by judging whether the appropriate assumptions have been made in their construction, as well as by proposing alternative assumptions. Here the statistical analysis tests how responsive the projections are to alterations in the assumptions. It raises questions like how much costs would be changed if participation rates were changed by a given amount or if unit prices of services were changed. Sensitivity analyses highlight the assumptions concerning the costs that are the most critical to the overall cost estimate. Further, as part of the PES analysis, estimates of the magnitude and direction of the problems of under- or overcosting that were identified could be applied to existing information and synthesized into a meaningful range.

Variants Using Other Sources of Information

The basic PES operational model uses prior evaluations or research as the source of information. If, in reviewing this literature, tradeoffs should be made between timeliness and comprehensiveness, strategies

such as sampling and time-limited searches could be adopted. There may be situations, however, when available information must be supplemented with some original data collection and when it may be more efficient to tap into existing knowledge through panels or expert judgments. Further, there may be situations where the PES is combined with original data collection and other audit work.

Combining the PES With Some Original Data Collection

The results of the PES may be supplemented with some original data collection, such as examination of agency records or surveys. That is, where existing data are insufficient and where time and resources permit, evaluators may want to use PES procedures up to the point of matching evidence and key assumptions. At this point, the PES could proceed on dual tracks with some highly targeted new data being collected while other, prior work is reviewed. Several of the reports already mentioned, such as one on the consequences of opening more combat support positions and units to women, involved multiple methods of data collection in answering a prospective question.²

For example, we were asked by the Congress to determine what might be learned from state and local experience in addressing mandate burdens. A law already in place since 1981 required the Congressional Budget Office to estimate such costs for proposed federal legislation. Similar requirements for reviewing the costs of proposed state legislation exist in 42 states. New legislation proposed by the congressional requesters would have required federal reimbursement for additional costs. This approach was already in use in 14 states that reimbursed local governments for burdens imposed by new state laws.³ The methods for answering the prospective question included a review of the literature, analysis of relevant bills, and visits to 8 states selected by searching prior studies, plus a telephone survey. Data from the 8 states were supplemented by questionnaires for state officials, state legislative leaders, and relevant interest groups. Using evidence from these 14 states, we found that estimating and reimbursing costs have had only a limited effect on the burden of mandates, except in some special circumstances.

²U.S. General Accounting Office, Women in the Military: Impact on Proposed Legislation to Open More Combat Support Positions and Units to Women, GAO/NSIAD-88-197BR (Washington, D.C.: July 1988).

³U.S. General Accounting Office, Legislative Mandates: State Experiences Offer Insights for Federal Action, GAO/HRD-88-75 (Washington, D.C.: September 1988).

When may such original data collection be particularly valuable? One might expect that in areas such as defense and tax policy, our unique access to data is likely to mean we would have better information than one could expect to find in the published literature. In other areas, however, such as certain aspects of health that require confidentiality in dealing with patients' records, physicians who are also evaluators and researchers might have the relative advantage and would find a richer data base in medical reports than we ourselves might be able to collect. That is, combining the PES with other forms of audit and evaluative work is consistent with the multimethod approach we typically use. However, evaluators planning a PES can also anticipate, to a certain extent, where we may find a relatively rich data base and where our unique authorization may suggest the need for new data collection to supplement the PES.

Combining the PES With Expert Judgment

The evaluator supplementing other evidence with the views of experts must be aware of the requirements of systematic methods such as Delphi techniques. Properly applied, these systematic methods yield information that differs in some key ways from the anecdotal evidence on which congressional testimony is often based. First, the effects of "charisma" in presenting testimony are ruled out. Second, since the same questions are usually asked of many key informants, it is possible to determine what opinion is generally held. Third, the bases for opinions are brought out and can be compared objectively with available evidence. Fourth, the experts or key informants can be selected primarily or solely by considerations such as knowledgeability and appropriate diversity.

We have used expert judgment and panels in a variety of ways to answer prospective (and also retrospective) questions. For example,

- to assess major welfare reform proposals dealing with case management, contracts between welfare recipients and agencies, coordination of services, and target populations, HRD contracted for two panels of experts. One panel consisted of experts at the national level and was convened by the National Academy of Public Administration; the other panel consisted of experts at the local level and was convened by the Federation for Community Planning. The findings of both panels were

synthesized by GAO and the numerous concerns, observations, and recommendations were presented to the Congress as the insights of expert panels.⁴

- to examine the probable effects of legislation that would change the conditions for legal immigration, we identified (in consultation with the customer) the issues and we brought together a panel of experts. The experts identified the highest-quality data relevant to these issues and presented their own conclusions. We then independently assessed the conclusions, relative to our own judgment of the quality of the evidence, in order to report the soundest available statement on probable effects.⁵

The use of expert judgment to supplement our prospective work requires (1) clarity in presentation when we are relying primarily on the opinions of others and (2) careful planning when the experts are a significant source but our own, independent judgment is needed. In the instance of proposed immigration legislation, the experts helped sharpen the issue, identified relevant empirical data, and examined points of consensus and dispute in the interpretation of the data. We then independently reviewed the available information and reached our own conclusions by the usual standards of audit and evaluation work.⁶

In another instance, GAO had a problem-definition assignment—examining the nature and extent of sweatshops in the United States and identifying the policy options that might help control the problem.⁷ In this study, which was clearly entitled opinions on the extent of the problem and possible enforcement options, we reviewed the relevant literature on sweatshops, particularly with regard to their origin and efforts at control; developed a working definition (since the term is not defined in federal statutes or regulations) in agreement with the customer; interviewed federal, state, and local officials, researchers, and union and management experts; surveyed state labor departments and agency officials; investigated possible sweatshops in New York and Los Angeles; and analyzed federal inspection reports. While this required more effort

⁴U.S. General Accounting Office, Welfare: Expert Panels' Insights on Major Reform Proposals, GAO/HRD-88-59 (Washington, D.C.: February 1988).

⁵U.S. General Accounting Office, "Immigration: S. 358 Would Change the Distribution of Immigrant Classes," GAO/T-PEMD-89-1, statement of Eleanor Chellmsky before the Subcommittee on Immigration and Refugee Affairs, Committee on the Judiciary, U.S. Senate, Washington, D.C., March 3, 1989.

⁶U.S. General Accounting Office, "Immigration," GAO/T-PEMD-89-1.

⁷U.S. General Accounting Office, "Sweatshops" in the U.S.: Opinions on Their Extent and Possible Enforcement Options, GAO/HRD-88-130BR (Washington, D.C.: August 1988). This was not formally a PES but illustrates a multimethod approach to analyzing a problem and possible action.

Chapter 6
Variants of the PES

than might usually be available for a PES, it illustrates that for certain prospective questions, GAO can negotiate with the congressional customer the time to undertake quite extensive involvement of experts, as well as site visits, to supplement the literature.

A Brief History of the PES and Some Other Prospective Methods

This appendix helps place the PES in relation to other methods. Traditionally, the basic concepts of evaluation have been used primarily in the assessment of policies and programs that are already in place. This *ex post* application has become so commonplace that it is the one most frequently associated with evaluation. Less frequently, evaluation methodology has been used to assess *ex ante* the potential success of policies that are under consideration.

The conventional approaches to prospective evaluations have ranged widely from relatively freewheeling "demonstrations" to highly controlled field experiments. However, proposed programs can be put into operation—often nationwide—with little evaluative evidence attesting to their potential for success. (Some of the unevaluated programs that have been put in place have to do with recent drug laws, various regulatory programs targeting improved health, "deinstitutionalization," "the strategic defense initiative," "pilot cities," "impact cities," "model cities," "operation push," and "operation breakthrough.")

But even when small-scale pilot efforts of an experimental sort are implemented—and most evaluators would agree that highly controlled field experiments yield the most credible results—the experiments have many practical drawbacks.¹ In particular, three serious limitations must be taken into account when they are considered for use as the only application of evaluation methodology to the assessment of prospective public policies. Consider, for example, three randomized public policy experiments: the five income-maintenance experiments, the housing allowance experiments, and the several experiments on demand pricing of electricity. First, they were costly. On this ground alone, it would not be likely that more than a small handful of experiments could be set under way during any decade. That is, only a minute proportion of the public policies and programs that are in any current policy space could possibly be assessed through field experiments.

Second, these field experiments were limited to the consideration of only a narrow band of alternative policies. Indeed, none of the income maintenance experiments came close to testing the actual public welfare policies that were considered by the Congress and the executive branch in

¹ Pilot and experimental studies can provide crucial intellectual capital on which synthesis draws. They are among the primary sources of information on which the PES relies. That is, a PES benefits from having available a good fund of knowledge based on evaluations of other programs, research knowledge, and so on. Thus, the PES does not replace the new data collection forms of program evaluation. Pointing out the limitations of pilot and experimental studies should not be misconstrued as arguing against this valuable prospective method.

the years since their completion. Policy space tends to be occupied by more contenders than can easily be accommodated in the design of the typical field experiment.² Furthermore, with every new administration or session of the Congress, the contending policies and programs, as embodied in various versions of proposed legislation, are never a static body and may in fact be constantly changing.

Third, public policy experiments take a long time to complete. Legislative proposals are often decided within the space of months and, at most, a few years. Clearly, field experiments that take 5 years to run and another 3 to analyze can rarely speak directly to any set of specific, proposed laws for the many years that typically pass before results appear. To some degree, these deficiencies are also characteristic of some other prospective efforts.

Pilot demonstrations that call for the collection of original observations in the field may take almost as long to carry through to completion as field experiments. Even cross-sectional surveys take significant periods of time. For example, a national household sample survey ordinarily takes from 6 months to up to 2 years to complete (depending on the complexity of sampling and analysis). In short, although "demonstrations" and quasi-experimental trials of prospective policies may take less time to conduct than the classical field experiments, they still may require more than several years to complete. In addition, they share the other drawbacks outlined above, being expensive and subject to increasing irrelevance with changes in the policy space.

In sum, the traditional ways in which evaluators have faced the problem of providing information to decisionmakers on the potential for success of policies and programs that may be under consideration at any time are not useful to a decisionmaking process that may take no longer than a year or two from proposal to definitive action. If evaluations are to contribute to decisions about proposed new programs, the contribution should be accomplished through procedures that are relatively inexpensive, speak to each of the variety of proposals under consideration, and provide timely results.

There is nothing especially new or startling about this idea, and many evaluators have given the problem some thought. A relevant example is

²This does not mean that the field experiments were irrelevant. Almost all the proposed welfare reform measures involved work-leisure tradeoff issues, a topic about which the five income maintenance experiments have much to contribute.

Appendix I
A Brief History of the PES and Some Other
Prospective Methods

an application of evaluative techniques to proposed legislation that advocated the use of national health screening for identifying abused children.³ The Evaluation Research Society has identified front-end analysis as a major focus of evaluative attention.⁴ Indeed, even the more extended forms of evaluation, such as randomized field experiments, could benefit from a PES conducted at the point of design. And there have been other efforts in recent years to come to grips with the problems of timeliness that are inherent in such front-end analysis. Many of the specific elements of PES have been advocated by others. In particular, evaluability assessment as developed by Joseph Wholey emphasizes the construction of underlying models of proposed programs in order to assess whether a program or policy can be evaluated for outcome effectiveness.⁵ In addition, many others stress the importance of the theoretical underpinnings of prospective programs.⁶

The main strength of the prospective evaluation synthesis is that because it draws upon existing knowledge and research to assess the potential success of a new proposal, it can be timely enough to be used within the policy development process. That is, the PES will not necessarily provide the best possible information that could be obtained under optimal conditions, but it can provide in a timely manner the best possible information that is currently available.

³Richard J. Light, "Abused and Neglected Children in America: A Study of Alternative Policies," *Harvard Educational Review*, 43:4 (November 1973), 209-13.

⁴"Evaluation Research Society Standards for Program Evaluation," in *Standards for Evaluation Practice*, no. 15, *New Directions for Program Evaluation* (San Francisco: Jossey-Bass, September 1982).

⁵Joseph Wholey, "Evaluability Assessment," in *Evaluation Research Methods*, L. Rutman (ed.) (Beverly Hills, Calif.: Sage Publications, 1977).

⁶For example, Huey-tyh Chen and Peter Rossi, "Evaluating with Sense: The Theory-Driven Approach," *Evaluation Review*, 7:3 (June 1983), 283-302; Margaret C. Wang and H. J. Walberg, "Evaluating Educational Programs: An Integrative, Causal-Modeling Approach," *Educational Evaluation and Policy Analysis*, 5:3 (1983), 347-66; Gary D. Gottfredson, "A Theory-Ridden Approach to Program Evaluation: A Method for Stimulating Researcher-Implementer Collaboration," *American Psychologist*, 39:10 (1984) 1101-12; John W. Finney and Rudolf H. Moos, "Environmental Assessment and Evaluation Research: Examples from Mental Health and Substance Abuse Programs," *Evaluation and Program Planning*, 7 (1984), 564-80; Karl E. Weick, *Social Psychology of Organizing*, 2nd ed. (New York: Random House, 1980).

Data Quality Judgment Models

The PES relies primarily on the results of past evaluations of previous or existing programs. That is, the results of a PES could be notably different if different rules were used for including a given study. Because the weighing of criteria used to judge the quality of prior studies is so critical to the results of a PES, this appendix discusses in some detail a point not elaborated upon in our paper on the evaluation synthesis: how criteria are aggregated in reaching a decision on whether to use (or how much emphasis to give) a specific study. There are at least four different ways to assess the quality of prior evaluation studies. Table II.1 summarizes the advantages and disadvantages of these four approaches.¹

¹We also note the special case of where quantitative estimates are required as part of a PES. In this instance, careful attention should be paid to the adequacy of our estimates of values that go into the PES analysis, including an examination of the quality of the data and methods for checking their validity. If data are not of truly high quality, provisions for boundary or sensitivity analyses should be made. Further, any time the functions we have to deal with are likely to be multiplicative rather than additive, the accuracy of values entered into the analysis is critical, particularly in going from local to national estimates. The PES could identify points at which data must be aggregated and could identify the vulnerability to multiplicative effects, where it is not possible as part of the PES to make these better estimates ourselves.

**Appendix II
Data Quality Judgment Models**

Table II.1: Advantages and Disadvantages of Four Data Quality Judgment Models

Model	Advantages	Disadvantages
One criterion	<p>Maximum number of prior reports brought to bear</p> <p>Large number of reports permits tests of interactions</p> <p>Analysis may be quicker since time for multiple quality screens is not taken</p>	<p>One strong report may be better than 20 weak ones</p> <p>One criterion is unlikely to be adequate, and interactions of data of mixed quality may be misleading</p>
Equally weighted	<p>If all criteria are in fact equally important, this model may best represent the quality of the prior evaluations</p> <p>Permits direct test of whether taking quality into account would make a difference in the findings</p>	<p>When several criteria are relevant, one may have little to analyze if a threshold for all is set, but not setting a threshold may permit a modest strength to offset a serious flaw in a study</p> <p>Rare to find all criteria equally important</p>
Unequally weighted	<p>Better represents relative importance of different criteria</p> <p>Permits direct test of whether taking quality into account would make a difference in the findings</p>	<p>A modest strength in one significant criterion can still offset a serious flaw in another criterion if there are two or more heavily weighted criteria</p> <p>Can be cumbersome to assign and compute weights for each criterion for each study, as well as to make ratings on each criterion on each study</p>
Threshold or fatal flaw	<p>Efficient in focusing on most crucial criteria</p> <p>Ensures that a study with high scores on several relatively minor criteria but a fatal weakness in one or more crucial criteria is not included</p>	<p>Must be sure the fatal flaw is sufficiently serious to be a screen ruling out studies that otherwise are potentially useful</p>

One Criterion Only

In this method, the set of prior research and evaluation studies on the general topic is developed—say, on food-stamp participation, military base closings, the effectiveness of federal programs aimed at disseminating knowledge, or the quality of executive and managerial personnel. The set is examined against a single criterion.

For example, a decision might be made that only one criterion such as measure validity should be really important for the job. This might be true if we are asked to assess the probable cost of a certain type of child care. Prior evaluations of child care that did not have information on costs that we considered complete and properly measured would be rejected. Those with valid cost information would be retained.

Except for the one selected criterion, other aspects of the quality of the relevant reports are not assessed in this method of synthesis. Rather, “strength through numbers” is the intention, with the notion that the

largest possible set of prior studies that meet the selected criterion will offer the soundest guide to answering the question. In a variant of this method, the information in the entire set of reports can be judged on the single criterion. The extent to which the answer to the evaluation question would differ when higher-quality and lower-quality studies (as judged by the single criterion) are used can be determined.

Among advantages of this approach are that it draws on the largest body of data. A prime disadvantage is that it is quite rare that only one criterion of study quality would be important. The evaluative question, as noted, would have to be quite limited in scope.

Equally Weighted Criteria

In this approach, a set of criteria for selecting the prior research to be synthesized is developed. Typically, the set includes relevance, recency, context similarity, and a variety of indicators of technical adequacy including those appropriate to measurement, design, analysis, and reporting.

Each of these criteria is given equal weight in deciding whether or not to include the report, article, or book in the set of material to be synthesized. That is, a high score on relevance might offset a lower score on technical adequacy when a "total" quality score is derived and the cut-off established for whether a study is included. Or, alternatively, a threshold score in all criteria may be required for the report to be used.

GAO has examples of this approach, including the criteria described in the reviews of the effect of illegal aliens on legal workers and the effect of the drinking-age laws on highway safety.² An advantage to this approach is that the effects of various aspects of quality can be tested empirically. A disadvantage is that particularly if a high threshold is set for all criteria, almost no studies may pass the quality screen.

Unequally Weighted Criteria

In this approach, the criteria receive different weights. For example, technical quality may be seen as more important than recency in deciding whether or not to include the study. Among the technical-quality criteria, for some questions the extent to which the design permits strong inference about causality may be seen as much more important

²U.S. General Accounting Office, *Illegal Aliens: Influence of Illegal Workers on Wages and Working Conditions of Legal Workers*, GAO/PEMD-88-13BR (Washington, D.C.: March 10, 1988), and *Drinking Age Laws: An Evaluation Synthesis of Their Impact on Highway Workers*, GAO/PEMD-87-10 (Washington, D.C.: March 16, 1987).

than, say, the extent to which documentation of measurement reliability exists. The weights are not arbitrary but are guided by the theory underlying the methods. Again, there are examples of this approach.³

This approach has the advantage of better representing the importance of different criteria. It is still possible, however, that modest strength on several relatively less important criteria can offset a serious flaw on a significant criterion, if scores on each criterion are aggregated.

Threshold or Fatal Flaw

In some situations, a report that does not pass muster on a specific criterion is not considered at all, and other criteria come into play only after the "fatal flaw" test has been passed. For example, in a synthesis of studies on the homeless mentally ill, reports that did not attempt to estimate the size of the local population of the homeless were excluded from consideration. Further, within the useful studies, a fatal flaws criterion (sampling the range of settings) set a cap on rated quality. That is, among the studies that estimate population size, the quality of the report was judged against seven other criteria and the direction and extent of bias were judged. The technical-quality rating was the profile of whether the errors were likely to lead to an overestimate bias or an underestimate and the size of the bias.⁴

This model is the most efficient way to ensure quality. The fatal flaw must be carefully examined, however, to be sure that no offsetting features are possible, since potentially informative studies that fail on only one criterion may be excluded from the review set.

Table II.2 provides a detailed example of the criteria used and how they were applied with regard to the number of homeless mentally ill persons.

³U.S. General Accounting Office, *WIC Evaluations Provide Some Favorable But No Conclusive Evidence on the Effects Expected for the Special Supplemental Program for Women, Infants, and Children*, GAO/PEMD-84-4 (Washington, D.C.: January 30, 1984).

⁴U.S. General Accounting Office, *Homeless Mentally Ill: Problems and Options in Estimating Numbers and Trends*, GAO/PEMD-88-24 (Washington, D.C.: August 3, 1988). Another example is U.S. General Accounting Office, *Influences of Illegal Workers on Wages and Working Conditions of Legal Workers*, GAO/PEMD-88-13-BR (Washington, D.C.: March 10, 1988).

**Appendix II
Data Quality Judgment Models**

Table II.2: Example of a Fatal Flaws Analysis

What we did	How we did it
Screening the studies	<p>In defining our universe of studies for the evaluation synthesis, we purposefully kept our inclusion criteria broad. We included any study, regardless of methodological quality, that attempted to estimate the size of the homeless or homeless mentally ill population. We did, however, have some minimum inclusion criteria. Of our universe of 83 studies, 27 were selected as useful. Specifically, we included a study in our universe if it met each of the following three criteria.</p> <ol style="list-style-type: none"> 1. The study was in written form. Telephone conversations, speeches, or conference proceedings without a written product were not included. 2. The study provided a count or estimate (by whatever method) of the homeless or homeless mentally ill persons or assessed trends in a designated geographic area. This would exclude case studies of individuals or studies describing service needs without a count or estimate. 3. The method used to make the estimate of the number of homeless or homeless mentally ill was sufficiently described to permit us to evaluate its merits (or shortcomings). By "sufficiently described," we mean the study provided some information on <ul style="list-style-type: none"> • the data used to make the estimate (for example, expert judgments or actual counts of persons in shelters); • how those data were collected (for example, shelter-providers were interviewed over the telephone, streets were canvassed by car, and so on); • how the estimate of the size of the homeless or homeless mentally ill population was actually computed (for example, how shelter and street counts were aggregated). That is, there was some kind of link between the data collected and the final population estimate.
Assessing the studies	<p>Next we rated the 27 relevant studies on two dimensions: technical quality and soundness (that is, the extent to which the chosen method would produce an underestimate or overestimate of the size of the homeless population). We discovered that many of the studies involved multiple methods for counting the homeless, reflecting the various settings (shelters, streets, institutions) in which the homeless and chronically mentally ill can be found. We considered each of these "nested studies" for how well it met survey methodology standards for soundness. Criteria for methodological soundness encompassed such issues as adequacy of universe definition, coverage of sampling frame, implementation procedures, and soundness of data analysis. We developed and applied a coding form to extract data relevant to these criteria. Finally, two staff members rated the full studies on criteria related to their overall sampling, measurement, implementation, and population estimation procedures.</p>
Sampling design	<p>Did the design cover the range of settings where homeless persons were likely to be found (shelters, streets, other public places, institutions)?</p> <p>Was the sample of shelters and institutions representative in terms of the area's shelter size (that is, number of beds) and type (public or private)?</p> <p>Did the sample of streets and other public places (such as census blocks) adequately cover the locations where the homeless are known to congregate?</p> <p>Did the sampling design account for seasonal variation in homelessness?</p> <p>Was the unit of analysis (such as municipality) clearly defined?</p>
Measurement	<p>Was the estimate of the number of homeless based on an actual count rather than expert judgment?</p> <p>Was a respondent's homeless status determined on the basis of screening questions?</p>

(continued)

**Appendix II
Data Quality Judgment Models**

What we did	How we did it
Implementation	<p>Were survey procedures explicitly stated in the report?</p> <p>Were interviewers trained to engage with and administer interviews to homeless persons?</p> <p>Were instruments pretested?</p> <p>If a street survey was conducted, were canvassing procedures consistently applied in areas searched? Were areas enumerated before the actual street survey was conducted?</p> <p>If a shelter-and-institutions survey was conducted, was the count based upon administrative records rather than subjective estimates? Were procedures developed to ensure an unduplicated count of the homeless within shelters and institutions?</p>
Deriving the population estimate	<p>Was the estimate of the number of homeless based upon a probability sample of areas (such as a national estimate based upon a probability sample of cities)?</p> <p>Were adjustments from the sample made to estimate the population (for example, was the application of a shelter-to-street ratio obtained from previous studies) appropriate and justified?</p>
Fatal flaws analysis	<p>In applying these criteria, we gave a higher priority to the sampling dimension. That is, if a study did not adequately sample the range of settings where homeless persons stay, there was a limit on how high the study could be rated, no matter how strong the measurement, implementation, and estimation procedures. To illustrate, a study that had a strong sampling design (for example, surveyed many settings) but used simple estimation procedures was rated higher than a study that had a weak sampling design (for example, surveyed only shelters) and used sophisticated statistical adjustments to account for the fact that streets or institutions were not surveyed. Accounting for sampling bias by using statistical adjustments—in some cases the only option available—is based on assumptions about the size of the homeless population in the settings not included in the survey, not an actual count. Applying the criteria in this manner, we rated each study's technical quality very high, high, moderate, low, or very low.</p> <p>Our second rating helped us distinguish where on the technical-quality scale (very high to very low) studies could be considered sound enough to provide reliable estimates. The soundness of studies was determined by rating each study on the extent to which its methodology would produce, in our judgment, an underestimate or overestimate of the number of homeless persons. For example, a study that employed a design that relied solely on the estimates of service providers would be rated as having the potential for overestimating the size of the homeless population. Each study was assigned a rating on a 7-point scale that ranged from -3 (serious underestimate) to +3 (serious overestimate). A written justification was given for each bias rating.</p> <p>To determine a cutoff point for the methodological soundness, we selected studies that received a bias rating of -1, 0, or +1. In addition to providing a cutoff point, this second rating indicates the direction and likely magnitude of the bias in each study.</p> <p>We used the information from these ratings to get an overview of the current approaches and research designs that are being used to count homeless and homeless chronically mentally ill persons. This information formed the basis for a closer examination of the patterns of strengths and weaknesses that were evident in the various studies and was applied in developing our alternative approaches.</p>

A Project Evaluation Profile

Study Code: _____ Reviewer: _____ Date: _____

A. Information Relevant to Conceptual and Operational Models

1. What services are provided, and how are they provided?

<u>Service</u>	<u>Available?</u>	<u>Directly?</u>	<u>Frequency?</u>	<u>Comments</u>
Perinatal health	no			
Well-child care	no			
Child care	no			
Transportation	no			
Counseling				
Educational	no			
Vocational	no			
Job skills and search	no			
Family planning	no			
Parenting education	yes	yes	2 hrs weekly	Both home and groups
Support groups	no			
Personal counseling				
Individual	no			
Group	yes	yes weekly	2-1/2 hr section	See results below

2. Arrangements for providing services

- How many services are provided in a single setting? 1
- Is there an explicit case management system? no
- Are referrals, if any, followed up? NA
- Is an individual service plan created and maintained? no
- What is the expected length of program participation per client? 4 months
- What percent of clients complete this expected stay? unknown

Comments: For each individual client, there is a plan with regard to location of parenting education; there is not, however, anything like a comprehensive case management plan.

3. Client characteristics

- What age groups are served? most are 16-29
- Is there an age limit? no
- What income eligibility requirements are there? none
- Percent below poverty line? unknown
- Percent recipient of or eligible for AFDC or General Assistance? unknown

Comments: Participants must be at high risk of having problems and be without adequate income resources, but no specifics given.

Appendix III
A Project Evaluation Profile

f. What proportion of program clients are

Low income	unknown but assume all
Students	unknown
Nonwhite	unknown
Dropouts	unknown
Fathers	none
Pregnant	unknown but all are mothers
Grandparents	none
Parents	100%

Comments: Program focuses on women with infants.

4. Provider characteristics

a. What is the primary setting of the provider or sponsor?

School	no
Alternative school	no
Community health clinic	no
Family planning clinic	no
Public welfare agency	yes
Private welfare agency	no
Other	yes

Comments: Service provided by the Children's Aid Society in cooperation with Utah State University's early childhood research program.

b. Sources of funding

SSBG	no
M&CHEG	no
AFL	no
Title XX	no
Medicaid	no
State or local government	yes
Private	no
Other	yes

Comments: local governments plus grant from federal agency, demonstration or research program.

c. Program cost per client? no information given

d. If costs not available, can they be calculated from number of clients and funds per year? no

Appendix III
A Project Evaluation Profile

5. Primary results claimed

Increased independence and development of viable support systems are claimed and attributed to the group counseling sessions.

Other claims are increased birth control use and use of family planning clinics (self-reports).

**Appendix III
A Project Evaluation Profile**

B. Evaluation quality

1. Outcome variables addressed: (Circle all that apply and record rating below from attached specific comments sheets)

- | | |
|---|-------------------------------|
| A. (Repeat) pregnancy | F. Pregnancy and birth rates |
| B. Birth outcomes--prematurity, birthweight | G. Sexuality information |
| C. School completion or continuation | H. Interpersonal skills |
| D. Employment, apprenticeship or training | I. Birth control use |
| E. Income, public assistance receipt | J. Family planning clinic use |

<u>Specific rating criteria/outcome</u>	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>	<u>E</u>
Comparison group				Q	Q
Sample size				Q	Q
Data collection				U/Q	U/Q
Measures				Q	Q
Threats to validity				U	U
Quantitative measure of difference				Q	Q

General remarks: Comparison group undefined. When and how data collected on whether employed or receiving AFDC not described. Proportions are simply tabulated. There is also a simple statement that 62% of program graduates and 28% of current participants are employed, but unclear whether this is intended as a description or an effect.

Ratings: A - Acceptable
Q - Questionable
U - Unacceptable

**Appendix III
A Project Evaluation Profile**

2. Details

<u>Criteria</u>	<u>Rating</u>	<u>Comments</u>
Comparison group compatibility (same age groups, demographic, denominators)	Q	Unidentified nonprogram parents
Sample size	Q	No information on whether 29 represents <u>all</u> participants and the rest are other agency clients or what
Data collection (surveys, administrative data, legally required records, self-report)	U/Q	No information on how collected
Measures (standardized, same period, adequately reflect programs objectives)	Q	Do not know if employed fulltime or parttime
Threats to validity (attempts to correct for recognized limitations, limit conclusions)	U	No control for age of client, no description of comparison gap
Quantitative measures of differences (netting out other causes, test significance)	Q	Figures are not given. Use term "significantly," but do not know if tested statistically
Ratings:	A - Acceptable Q - Questionable U - Unacceptable	

References

Cronbach, Lee. Designing Evaluations of Educational and Social Programs. San Francisco: Jossey-Bass, 1982.

"Evaluation Research Society Standards for Program Evaluation." Standards for Evaluation Practice. No. 15. New Directions for Program Evaluation. San Francisco: Jossey-Bass, September 1982.

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

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

U.S. General Accounting Office. Content Analysis: A Methodology for Structuring and Analyzing Written Material, PEMD transfer paper 3. Washington, D.C.: June 1982.

U.S. General Accounting Office. Drinking Age Laws: An Evaluation Synthesis of Their Impact on Highway Safety, GAO/PEMD 87-10. Washington, D.C.: March 16, 1987.

U.S. General Accounting Office. Estimated Employment Effects of Federal Economic Development Programs, GAO/OCE-84-4. Washington, D.C.: August 1984.

U.S. General Accounting Office. The Evaluation Synthesis, Methods Paper 1. Washington, D.C.: April 1983.

U.S. General Accounting Office. Farm Payments: Analysis of Proposals to Amend the \$50,000 Payment Limit, GAO/RCED-88-42BR. Washington, D.C.: October 1987.

U.S. General Accounting Office. Federal Land Management: Consideration of Proposed Alaska Land Exchange Should Be Discontinued, GAO/RCED-88-179. Washington, D.C.: September 1988.

U.S. General Accounting Office. Financing Higher Education: Examples Comparing Existing and Proposed Student Aid Programs, GAO/HRD-87-88FS. Washington, D.C.: April 1987.

U.S. General Accounting Office. Hazardous Waste: Uncertainties of Existing Data, GAO/PEMD-87-11BR. Washington, D.C.: February 18, 1987.

References

U.S. General Accounting Office. Homeless Mentally Ill: Problems and Options in Estimating Numbers and Trends, GAO/PEMD-88-24, Washington, D.C.: August 3, 1988.

U.S. General Accounting Office. Illegal Aliens: Influence of Illegal Workers on Wages and Working Conditions of Legal Workers, GAO/PEMD-88-13BR. Washington, D.C.: March 10, 1988.

U.S. General Accounting Office. Illegal Aliens: Limited Research Suggests Illegal Aliens May Displace Native Workers, GAO/PEMD-86-9BR. Washington, D.C.: April 21, 1986.

U.S. General Accounting Office. Immigration: The Future Flow of Legal Immigration of the United States, GAO/PEMD-88-6. Washington, D.C.: January 1988.

U.S. General Accounting Office. Legislative Mandates: State Experiences Offer Insights for Federal Action, GAO/HRD-88-75. Washington, D.C.: September 1988.

U.S. General Accounting Office. Medical Devices: FDA's Forecast of Problem Reports and FTEs Under H.R. 4640, GAO/PEMD-88-30. Washington, D.C.: July 1988.

U.S. General Accounting Office. Medicare: Catastrophic Illness Insurance, GAO/PEMD-87-21BR. Washington, D.C.: July 1987.

U.S. General Accounting Office. Milk Marketing Orders: Options for Change, GAO/RCED-88-9. Washington, D.C.: March 1988.

U.S. General Accounting Office. Minerals Critical to Developing Future Energy Technologies, Their Availability, and Projected Demand, GAO/EMD-81-104. Washington, D.C.: June 1981.

U.S. General Accounting Office. Models, Data, and War: A Critique of the Foundation for Defense Analyses, GAO/PAD-80-21. Washington, D.C.: May 1980.

U.S. General Accounting Office. Patent Policy: Recent Changes in Federal Law Considered Beneficial, GAO/RCED-87-44. Washington, D.C.: April 1987.

References

U.S. General Accounting Office. Simulations of a Medicare Prospective Payment System for Home Health Care, GAO/HRD-85-110. Washington, D.C.: September 1985.

U.S. General Accounting Office. "Sweatshops" in the U.S.: Opinions on Their Extent and Possible Enforcement Options, GAO/HRD-88-130BR. Washington, D.C.: August 1988.

U.S. General Accounting Office. Tax Administration: Difficulties in Accurately Estimating Tax Examination Yield, GAO/GGD-88-119. Washington, D.C.: August 1988.

U.S. General Accounting Office. Teenage Pregnancy: 500,000 Births a Year but Few Tested Programs, GAO/PEMD-86-16BR. Washington, D.C.: July 1986.

U.S. General Accounting Office. USDA's Commodity Program: The Accuracy of Budget Forecasts, GAO/PEMD-88-8. Washington, D.C.: April 1988.

U.S. General Accounting Office. Welfare Reform: Projected Effects of Requiring AFDC for Unemployed Parents Nationwide, GAO/HRD-88-88BR. Washington, D.C.: May 1988.

U.S. General Accounting Office. What the Department of Agriculture Has Done and Needs to Do to Improve Agricultural Commodity Forecasting and Reports, GAO/RED-76-6. Washington, D.C.: August 1975.

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

U.S. General Accounting Office. Women in the Military: Impact on Proposed Legislation to Open More Combat Support Positions and Units to Women, GAO/NSIAD-88-197BR. Washington, D.C.: July 1988.

Wholey, Joseph. "Evaluability Assessment." Evaluation Research Methods, ed. L. Rutman. Beverly Hills, Calif.: Sage Publications, 1977.

Requests for copies of GAO reports should be sent to:

**U.S. General Accounting Office
Post Office Box 6015
Gaithersburg, Maryland 20877**

Telephone 202-275-6241

The first five copies of each report are free. Additional copies are \$2.00 each.

There is a 25% discount on orders for 100 or more copies mailed to a single address.

Orders must be prepaid by cash or by check or money order made out to the Superintendent of Documents.

**United States
General Accounting Office
Washington, D.C. 20548**

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

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