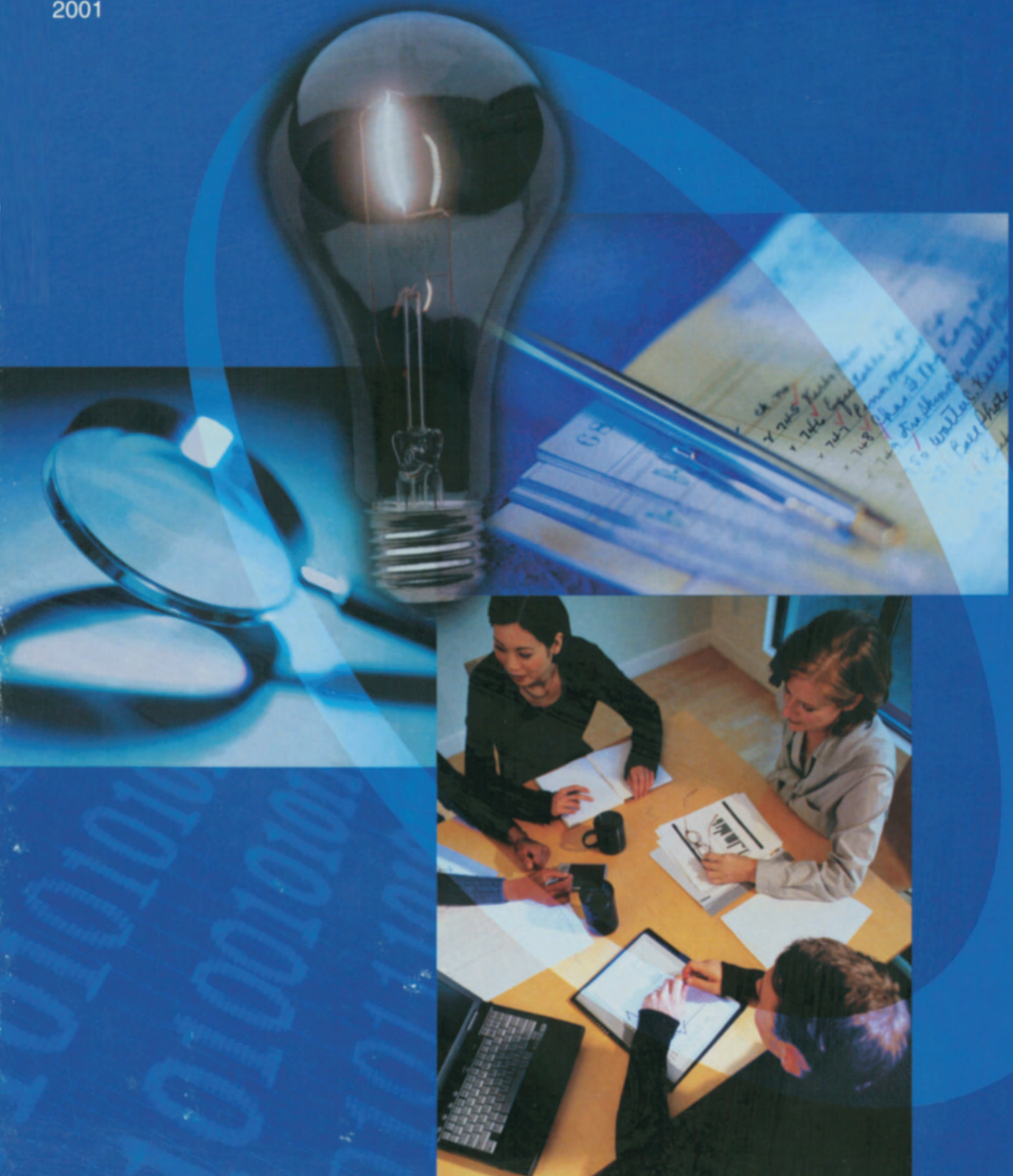
Improving the Evaluation of DOL/ETA Pilot and Demonstration Projects: A Guide for Practitioners

Research and Evaluation Report Series 01-A

U.S. Department Of Labor
Employment and Training Administration

2001



Improving the Evaluation of DOL/ETA Pilot and Demonstration Projects: A Guide for Practitioners

U.S. Department of Labor
Elaine Chao, Secretary

Employment and Training Administration
Raymond Uhalde, Deputy Assistant Secretary

Office of Policy and Research
Gerard F. Fiala, Administrator

2001

Abstract

The U.S. Department of Labor, Employment and Training Administration (DOL/ETA), funds numerous pilot and demonstration projects to test innovative employment and training approaches. Evaluating these projects is critical to the Department's store of knowledge about successful and unsuccessful program designs. This paper explores ways to refine and improve DOL/ETA demonstration and pilot evaluations to maximize what is learned, while holding the costs and intrusiveness of the research to a minimum. It provides local program operators, DOL/ETA staff, and evaluation contractors with a set of decision rules for designing and executing policy evaluations to produce more successful and valuable studies in the future.

Acknowledgments

In preparing this document, the author was assisted by several able colleagues. Janet Javar of the U. S. Department of Labor, Employment and Training Administration, guided planning for the paper and provided essential background materials. She and others at the Department, including Sande Schifferes and Jon Messenger, gave essential feedback and suggestions for revising an earlier draft. Demetra Nightingale of the Urban Institute directed the task order contract under which the paper was developed and provided helpful comments on early versions. Jonathan Fischbach conducted the literature review. A final debt is owed to Larry Orr of Abt Associates, whose collaboration with the author over many years in designing and executing employment and training evaluations sparked many of the ideas in this *Guide*.

This report is published as part of the Research and Evaluation Report Series in the Office of Policy and Research (OPR) of the Department of Labor's Employment and Training Administration. This series presents information about and results of projects funded by OPR. The Series is published and disseminated under the direction of OPR's Division of Policy, Legislation, and Dissemination: Terence Finegan, Division Director; and Andre Robinson, team member. This publication, as well as other research and evaluation reports, can be ordered online or downloaded through the website: www.ttrc.doleta.gov/opr/reports.html. Publications can also be ordered on the OPR publications' phone line at 202-693-3666 (please note that this is not a toll-free number), or requested by mail at: The Office of Policy and Research/The Division of Policy, Legislation, and Dissemination, U.S. Department of Labor, 200 Constitution Avenue, N.W., N-5637, Washington, D.C. 20210.

About the Author

Stephen H. Bell is an expert on employment and training evaluation design who, for the last 15 years, has helped design and execute numerous evaluations of employment and training programs for DOL/ETA and other agencies. His research includes the National JTPA Study, the Evaluation of the Lifelong Learning Demonstration, the Evaluation of Project NetWork (an SSA return-to-work demonstration for people with disabilities), and—most recently—the National Evaluation of the Welfare-to-Work Grants Program. Dr. Bell specializes in quantitative evaluation methods, particularly techniques for measuring training program impacts, benefits, and costs. Currently at the Urban Institute as a Principal Research Associate, Dr. Bell previously worked on evaluation projects at Abt Associates for more than a decade. His publications on evaluation methodology include *Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test*, "Are Nonexperimental Estimates Close Enough for Policy Purposes?", and "New Federalism and Research: Rearranging Old Methods to Study New Social Policies in the States."

A Guide for Practitioners... A Guide for YOU

As a project manager who must test new pilot and demonstration programs or policies, a researcher within or outside DOL/ETA who must design and execute evaluations, or a State or local practitioner who organizes and operates test programs, have you ever thought about how useful it would be to have some guidance in choosing the right type of evaluation and in ensuring that the evaluation is well-designed? Perhaps you already have an idea of what you want to address in your evaluation, but need additional information on fleshing out the design. Or, perhaps you are not aware that different types of evaluations exist, and now you must choose which type is best for your project. These types of scenarios make this *Guide* a useful reference tool for you to keep throughout the various stages of designing, operating, and completing evaluations.

Unsure whether this *Guide* is for you?

This *Guide* will help you consider important design questions such as:

What activity is to be tested in the pilot demonstration initiative?

What result(s) does DOL/ETA and/or your organization intend to achieve by the intervention?

Does an evaluation of these activities make sense?

What policy questions would the evaluation address?

How soon are answers needed for the key questions? Are results feasible within that time frame? If not, are there preliminary results that could substitute for final findings?

For the various types of findings needed, what analytic strategies can be adopted to ensure that findings are both credible and reliable?

Are there ways to scale back the evaluation agenda and still obtain valuable findings? What cut(s) in evaluation activities should come first?

This *Guide* was prepared just for *you, the practitioner.* Do not be discouraged by the thickness of this *Guide.* Like many reference tools, you will probably not need to read this *Guide* from cover to cover. Although reading each page will help you gain a better understanding of evaluations, Part I will guide you to the type of evaluation you need (including a quick "road map" on pages 14-17), and then you can simply reference to specific sections of Part II to learn more about the evaluation type that makes the most sense for your project.

Is all this effort put into evaluation worthwhile? Yes, and for several reasons, as the author states: (1) A new intervention tested once may never be tested again– it may well be judged a success or a failure based on this one set of demonstration or pilot results; (2) The very process of doing high quality research provides a unique opportunity for agency staff to learn-as-they-go about new policy ideas and how the test intervention plays out in the field; and (3) When evaluation research covers the right questions and provides staff with close-quarters familiarity with results, project managers will be strongly equipped to respond to any challenges to the study's substantive conclusions.

In short, by doing pilot and demonstration evaluation well, workforce development agencies will find the result of the added spending and effort will not be frustration and vulnerability, but solid and demonstrable accomplishment in one of its key mission areas: learning from test experience.

Using This *Guide*

This *Guide* provides advice on designing and conducting reliable research concerning new policy initiatives "field tested" at the U. S. Department of Labor/Employment and Training Administration (DOL/ETA). Such research, known as pilot project evaluation or demonstration evaluation, examines the accomplishments and worth of new policy ideas while still in "test mode" and reports its conclusions to policy makers responsible for deciding whether the intervention should be adopted as national policy, discontinued, or tested further.

The paper is both a treatise on evaluation methods and a reference guide for obtaining information on specific aspects of demonstration evaluation design and management. It is intended for use by:

- Project Officers and other federal government planners involved in testing new pilot and demonstration programs or policies at DOL/ETA,
- Research staff within or outside DOL/ETA who are asked to design and execute the evaluation, and
- Local demonstration staff who organize and operate test programs and who are frequently called upon to provide evaluation data.

Evaluations need to be crafted to maximize what is learned about the test programs/policies under study. DOL/ETA would like to see this learning process improved.

The *Guide* is organized into two parts. *Part I* provides an overview of the role of evaluation research in demonstration projects, summarizes the four basic types of evaluation (formative, descriptive, operational, and outcome), and offers practical advice to sponsors on how best to manage demonstration and pilot evaluations to achieve their goals. *Part II* covers the four evaluation types in detail, introducing readers to the design, data collection, and analysis components of each type and identifying "best practice" approaches to the many individual activities that characterize successful evaluations of employment and training pilots and demonstrations.

Users of this *Guide* will benefit most by

- Reading the document from front to back to familiarize themselves with the information presented and the range of topics covered,
- Looking up and considering more deeply topics of special interest (using the page references in the table of contents), and
- Returning to the *Guide* as a reference tool when dealing with specific issues in implementing a pilot or demonstration evaluation.

Those who wish to streamline the process should instead

- Read the "Evaluation Issues" and "Terminology" portions of *Section 1*,
- Examine *Section 2* to ensure that the appropriate evaluation types are implemented to answer the policy questions of vital interest,
- Read about the relevant evaluation types in *Sections 4 - 7*, looking up the topics of immediate interest within any section (using page references in the table of contents),
- Check *Section 3* for tips on key management decisions,
- During evaluation planning, return to the *Guide* to read the complete document in order to alert themselves to any research and management issues that may have been overlooked or mishandled up to that point, and
- As the evaluation proceeds, return to the *Guide* as a reference tool for dealing with specific oversight and design issues.

The importance of the next-to-last step can hardly be overstated. Those who do not read the entire Guide at the beginning should read the full text as soon as possible thereafter to identify evaluation issues that may have gone unrecognized or unresolved.

A third option—most appropriate for *employment and training practitioners* implementing a demonstration or pilot intervention—is to read all of Part I and then browse through Part II to identify any evaluation issues or components that may have gone unrecognized but carry implications for pilot implementation.

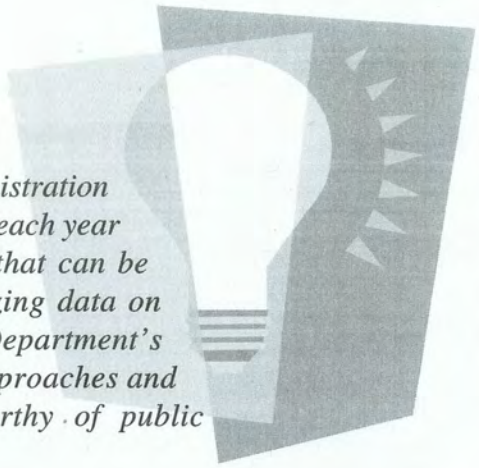
Used in this fashion, it is hoped that the *Guide* will help improve evaluation practice for DOL/ETA pilot and demonstration projects, increase the value of that research, and— as a result—contribute to the development of the Department's future policy agenda.

Table of Contents

PART II RESEARCH METHODS

List of Exhibits:

Part I: Overview of Evaluation Goals, Types, and Management Issues



The U.S. Department of Labor, Employment and Training Administration (DOL/ ETA), funds numerous pilot and demonstration projects each year which try out innovative employment and training interventions that can be disseminated if found effective and efficient. Having and analyzing data on these projects--i.e., evaluating them--is critical to increasing the Department's store of knowledge about successful and unsuccessful program approaches and fundamental to making demonstration and pilot projects worthy of public investment.

This paper explores how the evaluation component of DOL/ETA pilots and demonstrations might be refined to maximize what is learned from "test" runs of new ideas, while holding the costs and intrusiveness of the research to a minimum. Different approaches are recommended in different situations, based on state-of-the-art research techniques and recognition of the practical constraints facing "real world" field studies. The goal is to provide local program operators, DOL/ ETA staff, and evaluation contractors the background and decision rules needed to adopt and execute evaluation designs sensitive to local conditions yet capable of capturing all the "lessons" to be taught by trial runs of new policy ideas.

1. The Role of Evaluation and the Issues It Raises

Demonstration and pilot approaches to testing new policy initiatives combine several key elements, one of which is the evaluation--or research--component. To begin an examination of evaluation methods, it is important first to understand the larger framework in which demonstration research takes place and the specific roles played by the evaluation.

1.1 The Overall Structure of a Demonstration Project

Exhibit 1.1 summarizes the structure of the typical employment and training demonstration project, highlighting the role of evaluation as one of several operating components within the whole. As shown, demonstration and pilot projects begin with a set of inputs (detailed below) and--once running--consist of four or five operational components:

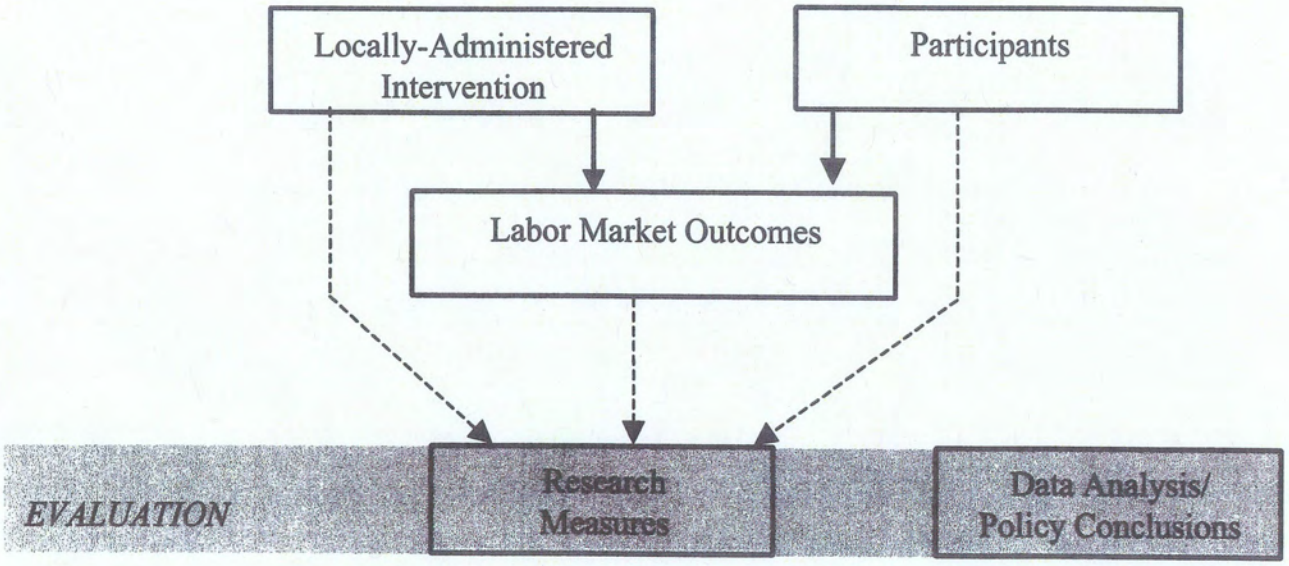
- The locally-administered intervention--the new program or policy implemented on a limited scale for test purposes;
- Participants--individuals who participate in the intervention;

Exhibit 1.1

Structure of Employment and Training Demonstration and Pilot Projects

INPUTS
[See Exhibit 1.2]

OPERATIONAL COMPONENTS



OUTPUTS
[See Exhibit 1.3]

- Labor market outcomes of participants, during and after demonstration participation; and

- An evaluation of the success of the intervention, combining research measures that describe other demonstration components with data analysis and policy conclusions.

A number of outputs, or "end products," result from these activities.

Exhibits 1.2 and 1.3 diagram the input and output components of the process--the start and end points of the typical employment and training demonstration. As illustrated in the two sides of Exhibit 1.2, some demonstration inputs are within the control of the sponsoring agency and others not. For example, federal officials at DOL/ETA specify the program or policy to be tested and its participant targeting requirements. DOL/ETA also provides funding for the test intervention and staff to oversee and assist local program operators as they implement the intervention. Local implementation agencies--the organizations that receive federal grants to conduct the demonstration--constitute the final controlled input to the demonstration.

Uncontrolled inputs can also play a major role in employment and training demonstrations. These include:

- The existing program environment in the demonstration sites;

- Characteristics of the local population targeted for inclusion in the demonstration; and

- Local economic conditions.

The existing program environment--as a supplement or substitute for the demonstration intervention--influences both how the new intervention is implemented and how participants fare in the labor market. If the target population has other effective employment assistance programs in place for it in the community, all members of the target population--including demonstration participants--may do better in the labor market independently of the effects of the demonstration. Alternatively, the presence of other similar interventions may make local project administrators more able to administer the demonstration effectively based on experience gained in administering other successful programs.

Exhibit 1.2

Controlled and Uncontrolled Inputs to Employment and Training Demonstration and Pilot Projects

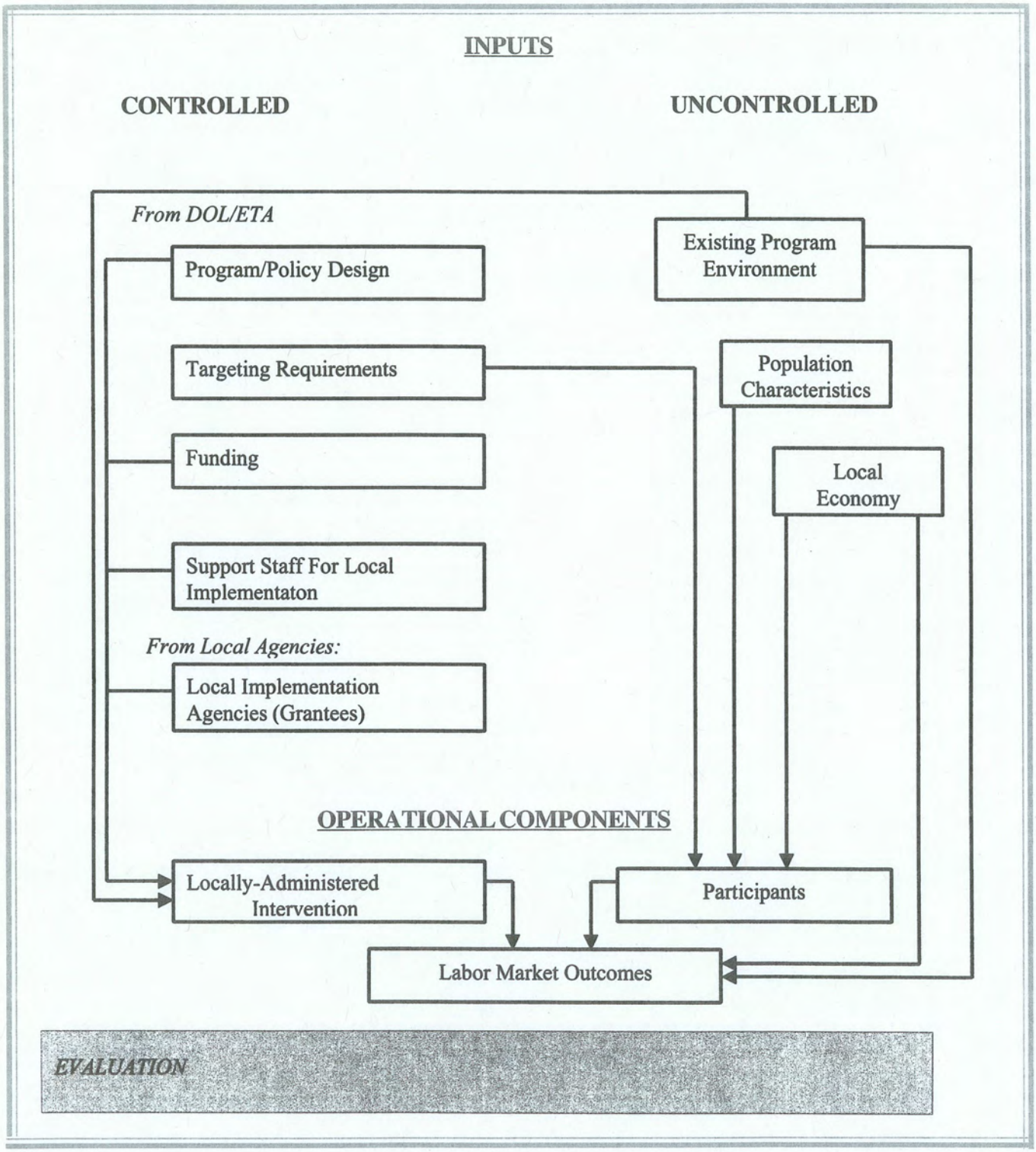
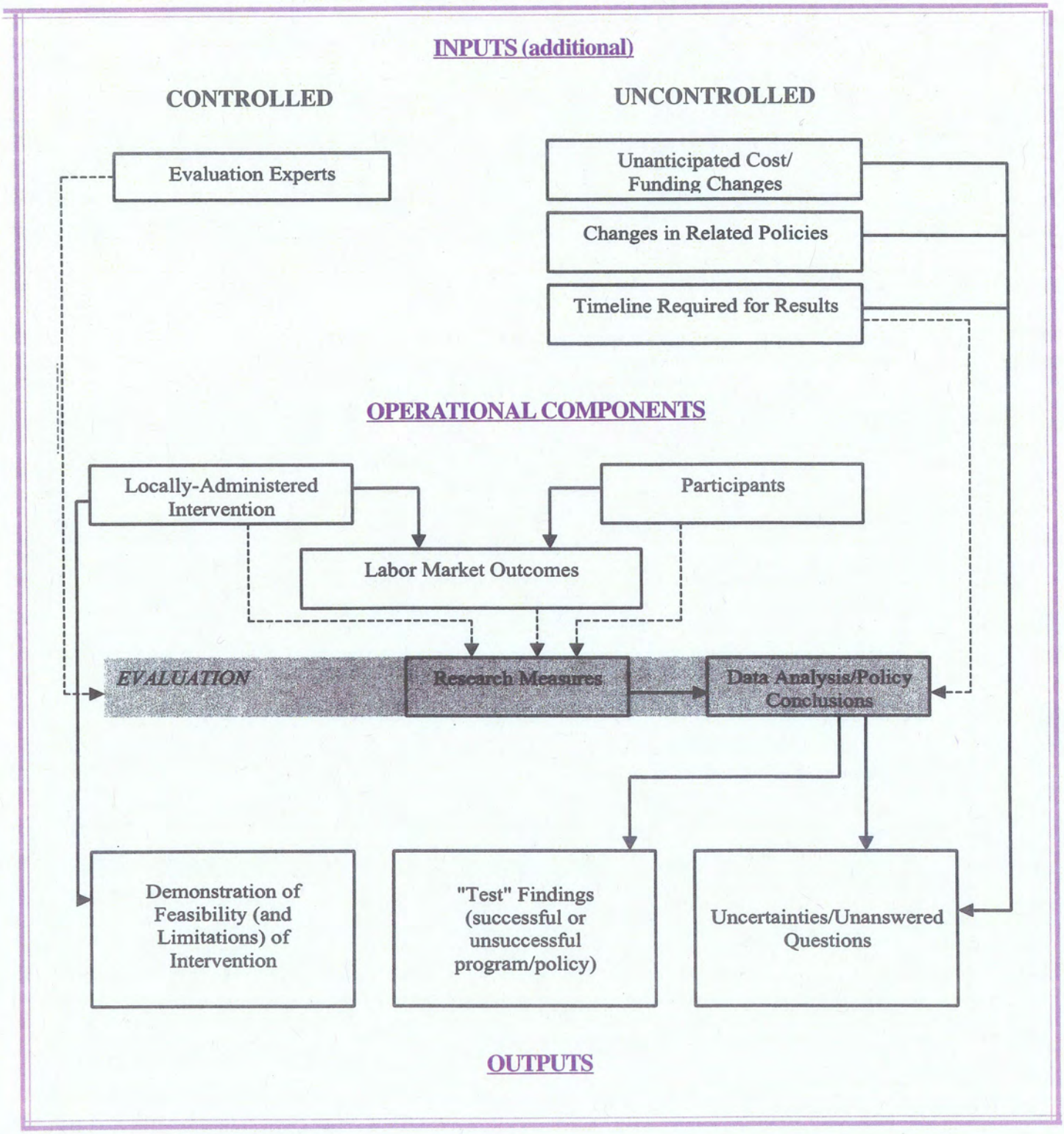


Exhibit 1.3

Outputs of Employment and Training Demonstration and Pilot Projects



Finally, the types of workers who seek help from a demonstration will be affected by local economic conditions. Localities with unusually high unemployment or unusually rapid economic change are likely to attract more experienced and educated workers into employment-assistance projects. In a strong economy, however, very few experienced workers need the help of a demonstration project to find or hold a good job, so demonstration participants will tend to be less educated and less seasoned as workers. The health of the local economy will also influence how demonstration participants do in the labor market irrespective of the type or effectiveness of the intervention tested by the demonstration.

Additional inputs to the demonstration influence only the evaluation component of the project. These are shown in Exhibit 1.3, along with typical outputs sponsors can expect of a demonstration. The exhibit links inputs to operational components, and operational components to outputs. Most employment and training demonstrations provide three major outputs:

- A demonstration of the feasibility of implementing the test intervention in a real-world setting. This step can also be used to identify the strengths and limitations of the implementation approach taken by the various sites.

- Findings on the success of the test intervention relative to its goals. Essential to this assessment is a comparison of demonstration accomplishments to the policy goals of the intervention as identified by DOL/ETA.

- Uncertainties and questions remaining at the end of the evaluation regarding the new intervention tested.

This last unwanted, but inevitable, output of pilot and demonstration projects results from the inherent limitations of current research techniques (as discussed in the remainder of the document) and the "uncontrolled" input factors shown in the upper right of the exhibit. Factors that add uncertainty to a demonstration's results include:

- Unanticipated changes in demonstration or evaluation costs and funding levels;

- Changes in non-demonstration programs or policies that confound or obscure the intervention of interest; and

- Tight timelines for providing demonstration results to policymakers.

A major function of the *Guide* is to explore "best-practice" responses to these types of externally-imposed constraints on the evaluation to minimize the number of unanswered questions remaining at the end of a demonstration. As noted in the exhibit, one additional controlled input is needed to accomplish this goal--evaluation experts from within or outside DOL/ETA.

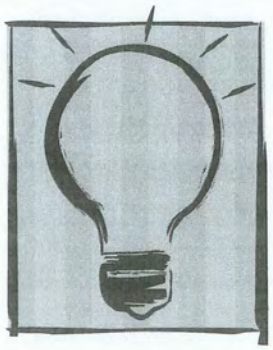
1.2 How Evaluations Contribute

Within this framework for demonstration research, evaluations of new employment and training policies and programs serve several purposes. Specifically, they:

- ***Identify*** the ***goals*** of a new initiative and stipulate the ***outcome measures*** to be used to judge its success.
- ***Anticipate potential weaknesses*** in the intervention--along with possible difficulties in its implementation--prior to start-up, to improve the "treatment" before the demonstration begins.
- ***Monitor*** the initiative's ***start-up process*** to give program operators "real time" feedback used in refining the intervention after the demonstration begins.
- ***Document*** the specific ***intervention tested*** in each of the test sites, including the program's organization and management and the delivery process for individual services.
- Give timely reports on the ***progress*** of project operations in each site--customers served, dollars spent, types of participant activities, etc.
- Gauge the ***achievements*** of the initiative, relative to the project's goals and administrative hurdles.
- Identify the ways that the intervention might be ***improved*** procedurally and consider its potential for successful ***replication*** in other settings.

Of course, not every evaluation of a demonstration or pilot project encompasses all of these goals. But most pursue several of the objectives.

All of these purposes will be advanced if DOL/ETA carefully considers--and, if appropriate, implements --evaluations that follow the strongest possible research approach for any given pilot or demonstration project. Substantial progress toward this goal has already been achieved: DOL/ETA has conducted evaluations of many of its pilot and demonstration projects over the years.[1] Some of DOL/ETA's most successful and well-known evaluations examined such demonstration and pilot initiatives as:



- National Supported Work Demonstration;
- WIN Labs Experiments;
- Unemployment Insurance Reemployment Demonstrations in Illinois, New Jersey, Pennsylvania, Washington, Florida, and the District of Columbia;
- STEP: Summer Training and Education Program;
- JOBSTART; and
- Unemployment Insurance Self-Employment Demonstration in Washington and Massachusetts.

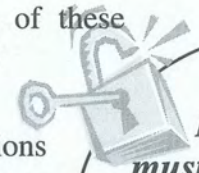
[1]DOL/ETA has also been one of the few federal agencies to rigorously evaluate its ***ongoing*** programs, most visibly Title IIA of the Job Training Partnership Act (JTPA) and the residential Job Corps.

Much has been learned through these and other DOL/ETA demonstration evaluations, which likely have paid for themselves--perhaps many times over--through better policy and program decisions. However, the use of evaluation has not always been universal or grounded in the best possible research practices. Even greater (and sharper) lessons can be learned in the future, if careful planning and up-to-date research techniques are adopted throughout the agency when running demonstration programs.

1.3 Evaluation Issues

In pursuit of better evaluations, this paper provides a "users guide" to the effective use of social science research techniques to study DOL/ETA pilot and demonstration projects. It focuses on the key questions that must be answered in order to design and execute a successful demonstration evaluation:

- What activity will be tested in the pilot demonstration initiative?
- What result(s) does DOL/ETA intend to achieve by the intervention?
- Does an evaluation of these activities make sense?
- What policy questions would the evaluation address?
- How soon are answers needed for the key questions? Are results feasible within that time frame? If not, are there preliminary results that could substitute for final findings?
- Will a process analysis of demonstration operations--planning, implementation, ongoing service delivery--be needed? What about an analysis of participant outcomes? An assessment of the *improvement* in outcomes spurred by the new policy? A comparison of benefits and costs?



Key questions must be answered in order to design and execute a successful demonstration evaluation.

- For the various types of findings needed, what analytic strategies can be adopted to ensure that findings are both credible and reliable?
- On what scale should program assessment take place to assure statistical reliability and broad meaning in the results?
- What are the data demands of such a study, and how can they be filled without major data collection initiatives?
- Are there ways to scale back the evaluation agenda and still obtain valuable findings? What cut(s) should come first?
- What do demonstration sponsors need to do on a continuing basis to assure successful execution of the chosen evaluation plan?

1.4 Outline of the Paper

To address these questions, the remainder of this section defines some critical terms used throughout the *Guide,* such as "evaluation" and "pilot." Section 2 provides a guide to choosing the right type or types of evaluation research for a given pilot or demonstration study, working from a list of four major evaluation types --"formative," "descriptive," "operational," and "outcome"--defined below. The right choice among these options depends on the questions the sponsor (in this case, DOL/ETA) wants the evaluation to answer.

This material, along with section 3, constitutes Part I of the *Guide*, Overview of Evaluation Goals, Types, and Management Issues. Section 3 outlines the supports (financial and non-financial) sponsors need to commit to evaluations if they are to succeed, considers how funding priorities might be set within a given study, and notes some of the pitfalls that can arise in managing the evaluation process. It also explains why all this bother is worthwhile, in terms of pay-offs to the sponsoring agency, local program operators, the DOL/ETA consumers who benefit from demonstration-induced policy refinements, and the taxpayers who finance the resulting government programs.

In Part II of the paper, "Research Methods," sections 4 through 7 discuss the four main types of evaluation analyses in detail, including their roles and methods. This second portion of the document is both essential reading and a general reference for learning more about specific evaluation issues during the course of a study's design and execution.

1.5 Terminology

To discuss evaluation methods for DOL/ETA's pilot and demonstration projects, the first step is to define terms. In this guide, "evaluation methods" are defined as follows:

- ***Evaluation methods*** = research tools for compiling information and drawing conclusions about specific government programs or policies.

The DOL/ETA pilots and demonstrations to which those tools might be applied also have a specific definition here:

- ***Pilots and demonstrations*** = special program or policy initiatives undertaken on a small scale to test the feasibility and effectiveness of a new (or at least previously untried) or improved policy idea.

- ***DOL/ETA pilots and demonstrations*** = pilots and demonstrations funded in whole or in part by the U.S. Department of Labor, Employment and Training Administration.

Historically, the two terms "pilot" and "demonstration" differed in the degree of formalism applied when specifying the intervention and conducting the research. Pilots tended to be less formal and ambitious, concerned almost exclusively with the feasibility of a new idea--i.e., whether a new policy or program could be created and run smoothly in the field, from the government perspective. The exact definition of the intervention, and its consistency across pilot sites, did not matter much, nor did its broader social accomplishments or participant benefits. The primary goal was simply to determine whether the general concept "would fly" when tried in the real world. Pilot results were often reported on an *ad hoc* basis by practitioners and oversight officials without any formal research structure, and generally stated project leadership's impressions of what worked or did not work along with (in some instances) tracking data on the number and characteristics of participants and their involvement in the project. Project management reports of this type proved useful, despite their limitations, and often led to a recommendation for more structured implementation (and study) or adoption of the approach as national policy.

Demonstrations have generally set loftier goals. In addition to piloting an intervention to see if those running it view it as a success, agencies that initiate a demonstration hope to be able to ***demonstrate*** to others that the intervention worked, using not just testimonials and management summaries but a body of factual evidence compiled from a formal research protocol that meets accepted scientific standards. Ideally, demonstration sponsors end up with a dossier of "proof" that the test policy worked (or did not work)--and why--to aid them in choosing the right course and making their case with others. This approach has proven so much more reliable and persuasive to policy makers that most projects today conceived as "pilots" include a formal evaluation component and easily could be called "demonstrations."[2] Given this blurring of distinctions, the two terms are used interchangeably in this *Guide*.

It may be helpful to note the DOL/ETA projects and research studies that are ***not*** demonstrations (or, equivalently, not pilots), and hence are not examined here. These include:

- Research on labor market issues that provides vital background information without focusing on the consequences of specific program or policy approaches (e.g., research on the extent and nature of employer-based skill training);
- Studies of ongoing national programs (e.g., the National JTPA Study); and
- Implementation studies of newly enacted national policies (e.g., DOL/ETA's survey of grantees' progress implementing programs funded by the new Welfare-to-Work Grants Program).

[2]The Workforce Investment Act of 1998 requires that all DOL pilots and demonstrations have a formal evaluation component.

Not one of these three types of research is undertaken specifically to ***test*** new policy ideas; rather, each provides knowledge useful in ***developing*** new policy ideas or measuring the consequences of ***existing*** national policies.[3] This *Guide* focuses exclusively on tests of pilot and demonstration interventions undertaken as models for future national policy.

Finally, the four main types of evaluation research are defined as follows:

- ***Formative evaluation*** = an analysis of pre-existing data to determine the feasibility of evaluating a proposed demonstration or pilot intervention, and to help planners refine the intervention before it is implemented.

- ***Descriptive evaluation*** = an analysis of the steps taken by local agencies to create and run a demonstration or pilot intervention, including a summary of the characteristics and in-program activities of demonstration participants.

- ***Operational evaluation*** = an analysis of the operational strengths and weaknesses of a demonstration or pilot intervention once implemented, with suggestions for how it might be strengthened in future applications.

- ***Outcome evaluation*** = an analysis of demonstration or pilot participants after they leave the demonstration, tracking their labor market outcomes (levels and changes) and assessing the short-run and/or long-run impacts of the intervention on those outcomes--and on overall social benefits and costs.

2. Choosing the Right Evaluation Type or Types

There is much to choose from when designing pilot and demonstration project evaluations and many, many factors to take into account. Given the complexity, no evaluator or evaluation sponsor can expect to ever reach the optimal plan when conducting a demonstration or pilot study. Still, all evaluation teams are likely to move substantially closer to the ideal by applying more thought and evaluation expertise to the panoply of research design and analysis issues identified in this document. It is not possible to compile a complete "look-up" table of evaluation options and recommendations for each of the combinations of circumstances that can arise in doing evaluations of employment and training pilot and demonstration projects. That is why careful thought and input from evaluation experts will always be essential--bordering on indispensable--when undertaking a demonstration or pilot study.

Still, one can outline at least one step of the research design process in a comprehensive way--the first and most fundamental step of choosing the right evaluation type or types for any particular demonstration. This first in-depth section of the paper provides such a "road map."

[3]Of course, it is a good to idea to periodically (or even continuously) check on the efficacy of existing national policies and programs, as has been done once in both the National JTPA Study and the National Job Corps Evaluation. But such studies are not examined here.

2.1 Prioritizing Research Goals

From a sponsor's point of view, deciding what broad type, or types, of evaluation research to undertake amounts to nothing more than choosing the research goals for the study. With a complete list of the questions about the pilot or demonstration initiative that one would like to answer, and the resolve to shorten the list if necessary to fit within available evaluation resources, the tools needed to select the right research type or types for a given pilot or demonstration evaluation are here in this section.

Unfortunately, establishing the exact set of research goals or questions to be addressed by an evaluation is not as simple as it sounds, particularly when circumstances require balancing policy needs and evaluation costs. Often placed between a rock and a hard place, the sponsoring agency must make tough decisions. Evaluation experts can help by identifying the tradeoffs presented and, if asked, providing recommendations on the research goals it sees as the most worthwhile and attainable. Ultimately, though, the demonstration program office must decide what is to be learned from a given pilot or demonstration initiative, as the organization closest to the policy issues that inspired the test in the first place. It is also the office most "on the line" to provide valuable policy information from the demonstration investment.

Evaluation experts could also assist in goal-setting by supplying a list of potential, or candidate, topics on which a demonstration study could make an important contribution. Sponsors would then use this list as an "options sheet" while discussing research goals and the priorities among them. Exhibit 2.1 provides such a list by summarizing the research goals associated with the various evaluation types of interest (see first column of the exhibit). The goals are stated here in generic terms, since each could apply to any type of labor market intervention that DOL/ETA might test.

To make it easier for sponsors and other planners to find a particular research goal on the list, the exhibit divides goals into seven topical categories. For example, research goals concerning the feasibility of the demonstration's program approach appear in the second panel of the exhibit and include:

"Check ability to create intervention,"
"Decide whether intervention is interacting with participants as planned," and
"Identify problem spots, resolution."

Individual goals are numbered consecutively through the entire list.

The full range of topic areas covered by the list includes, in sequential order:

- ✔ Ways to Strengthen the Demonstration Intervention,
- ✔ Feasibility of Approach,
- ✔ Operations Management and Accountability,
- ✔ Post-Intervention Trends,
- ✔ Effectiveness of Intervention,
- ✔ Major Beneficiaries, and
- ✔ Future Applications.

2.2 Translating Research Goals into Evaluation Types

Subsequent columns of Exhibit 2.1 indicate the particular evaluation type that addresses each of the research goals stated in the first column. As can be seen, ***each research goal can be addressed by one and only one evaluation type***. It follows that the research goals set by sponsor agencies dictate the particular type or types of evaluation research required. Shifting to an alternative type for convenience or cost reduction simply is not possible.[4]

This means that the research requirements of a demonstration often cannot be met doing research of just one type. For example, a management-focused study that seeks to determine how demonstration funds were used ***and*** whether the intervention implemented matched the planned intervention would have to include both descriptive and operational components in its evaluation (see rows 8 and 12). This particular blend would not be particularly expensive, since (1) addressing only these two topics requires but a fraction of what it would take to conduct full-scale versions of descriptive and operational evaluations, and (2) the two topics involve complementary research activities. Other combinations do not dovetail so well; for example, an alternative agenda again with just two goals--comparing actual and planned interventions and determining how much taxpayers paid per dollar of social benefit (rows 12 and 14)--would require that two distinct evaluation components, operational and outcome, be executed on a large scale with little overlap between them.

When situations of this sort arise, ***the only way to avoid the complexities and costs of multiple evaluation types[5] is to selectively drop one or more of the study goals***. For example, an evaluation sponsor that wants to achieve all three of the goals just mentioned (goals 8, 12, and 14 in Exhibit 2.1) and, in addition, strengthen the demonstration as a setting for good evaluation (goal 7) faces the prospect of undertaking all four evaluation types at once--or dropping one or more of the goals. There is no other way out: the evaluation setting cannot be strengthened without a formative evaluation, funds cannot be tracked absent a descriptive evaluation, fealty of the intervention to plan cannot be confirmed without an operational evaluation, and taxpayer payments per dollar of social benefit can only be calculated using an outcome evaluation.

Specialization among types has one advantage, however: once the list of evaluation goals is in hand for a given demonstration, it translates immediately into the one and only combination of research types that can meet that agenda, making the first step in a multi-stage research design fairly deterministic. Exhibit 2.1 provides the "road map" for this initial step, setting the foundation for the second phase of more detailed decision-making that can only be accomplished by applying the judgments and suggestions provided elsewhere in this *Guide*.

One further point of discretion does arise at this point, however: ***how much*** of a particular evaluation type to implement when only a portion is needed. If other research goals within that type can be accomplished using a less comprehensive approach than that described in section 4 (or 5 or 6 or 7) below, a less-is-better approach makes sense from a budgetary and management standpoint. Exceptions will arise, however, when some required research element depends critically on completion of another piece of analysis within that same type. Several such situations can arise in conducting outcome evaluations, and other evaluations have similar interdependencies. For example, "testing hypotheses about who benefits most" from the demonstration intervention (goal 27) is impossible without first having accomplished goal 22, "estimating impacts on participant outcomes." Most professional evaluators will recognize these interdependencies, while lay people generally will not, highlighting the importance of getting expert input at the very beginning of evaluation planning.

---

[4]This lack of overlap among types is not an accident, since new evaluation types evolved historically to meet informational requirements that previous types could not. As a result, each type has come to be defined in terms of what other types ***cannot*** do, leaving little or no overlap between the missions and capabilities of the four different approaches.

[5]Evaluation costs are driven substantially by the number and nature of the research types undertaken. Issues of cost, and technical responses to limited evaluation funding, are discussed in section 3.

Exhibit 2.1

What Research Goals Imply about Evaluation Types for Pilots and Demonstrations, by Topic Area

TOPIC AREA/ Research Goal	Evaluation Type Needed to Address Goal			
	Formative	Descriptive	Operational	Outcome
WAYS TO STRENGTHEN THE DEMONSTRATION INTERVENTION				
1. Refine/improve intervention design	✔			
2. Smooth initial implementation	✔			
3. Guide mid-course corrections			✔	
FEASIBILITY OF APPROACH				
4. Check ability to create intervention			✔	
5. Decide whether intervention is interacting with participants as planned			✔	
6. Identify problems spots, resolution			✔	
OPERATIONS MANAGEMENT AND ACCOUNTABILITY				
7. Make demonstration a better setting for evaluation	✔			
8. Describe use of demonstration funds		✔		
9. Document how treatment was produced		✔		
10. Distinguish test intervention from other similar policies		✔		
11. Provide information for later normative evaluations		✔		

(continues)

Exhibit 2.1 *(continuation)*

What Research Goals Imply about Evaluation Types for Pilots and Demonstrations, by Topic Area

TOPIC AREA/ Research Goal	Evaluation Type Needed to Address Goal			
	Formative	Descriptive	Operational	Outcome
OPERATIONS MANAGEMENT AND ACCOUNTABILITY (continuation)				
12. Determine if actual intervention matched plan			✔	
13. Gauge achievements vs. goals regarding outcome levels, changes, and impacts				✔
14. Determine how much taxpayers paid per dollar of social benefit				✔
15. Reject interventions that are bad investments of taxpayer's money				✔
POST-INTERVENTION TRENDS				
16. Describe what happens to participants after exit for key outcome measures				✔
17. Examine outcome trends over time				✔
18. Quantify improvements in outcomes over time				✔
EFFECTIVENESS OF INTERVENTION				
19. Ascertain strengths and weaknesses of treatment			✔	
20. Note goals that may be unattainable given operational limitations			✔	

(continues)

Exhibit 2.1 *(continuation)*

What Research Goals Imply about Evaluation Types for Pilots and Demonstrations, by Topic Area

TOPIC AREA/ Research Goal	Evaluation Type Needed to Address Goal			
	Formative	Descriptive	Operational	Outcome
EFFECTIVENESS OF INTERVENTION (continuation)				
21. Develop hypotheses of intervention's effects on participants			✔	
22. Estimate impacts on participant outcomes				✔
23. Test hypotheses regarding demonstration impacts				✔
MAJOR BENEFICIARIES				
24. Develop hypotheses about who will benefit most			✔	
25. Note subgroups with best outcomes				✔
26. Note subgroups with largest outcome gains over time				✔
27. Test hypotheses about who benefits most (who has largest impacts)				✔
FUTURE APPLICATIONS				
28. Know exact nature of treatment		✔		
29. Craft intervention "blueprint" for future		✔		
30. Consider potential for replicating successful intervention(s)			✔	

(continues)

Exhibit 2.1 *(continuation)*

What Research Goals Imply about Evaluation Types for Pilots and Demonstrations, by Topic Area

TOPIC AREA/ Research Goal	Evaluation Type Needed to Address Goal			
	Formative	Descriptive	Operational	Outcome
FUTURE APPLICATIONS (continuation)				
31. Identify ways to improve intervention			✔	
32. Decide if intervention has enough social benefits to offset full social costs and, thus, warrant expansion				✔
33. Provide factual basis for advocating/opposing pilot or demonstration intervention as national policy based on full social costs				✔

2.3 Beyond Agenda Setting

After thinking carefully about evaluation objectives and identifying the required evaluation type(s)--or subparts of types--a demonstration sponsor could be tempted to engage outside experts to do the rest of the work and retreat to an oversight role.[6] This would be a mistake. Sponsor agencies are better advised to remain integrally involved in the research design process for some time, as the list of selected types, expands into a full-fledged "blueprint" for the analysis. As evidenced by the earlier discussion of individual types, there are still a myriad of design questions to be answered, many with important substantive implications for the meaning and utility of the findings (e.g., issues surrounding the exact population studied, the indicator measures used, and the timing of data collection and reporting). Continued direct involvement of the sponsor through this second level of planning can be essential to making the research "work" in policy terms. More specifically, sponsors will gain from this investment in a number of areas:

- Sponsor staff need to be able to ***explain the study's design and methodologies***--including virtues and limitations--***from their own knowledge***, not simply that of the research team.

- A broader knowledge base--and early, in-depth thinking about research issues--makes the sponsor a ***better overseer of researchers as the evaluation is carried out***. The importance of effective oversight in assuring top-quality, policy-relevant research can hardly be overemphasized. (See section 3 for specific recommendations in this area.)

- The credibility and salience of an evaluation may depend on the sponsor's ***ability to gauge the meaning of intermediate findings***, and to make prudent decisions regarding their release and interpretation.

- Ongoing engagement with research issues makes the sponsor ***a more able "voice" in conveying the demonstration's*** findings and ***policy implications*** to government decision makers--and in demonstrating to budget officials the value of the intervention.

In concrete terms, what does a sponsor need to do to stay involved in the design process, keep the research on track, and realize these benefits? Three things:

Devote adequate time and concentrated attention to major evaluation design issues on a continuous basis before the intervention goes into the field. An evaluation structure only loosely or partially defined on D-Day (demonstration "launch" day) invites trouble, as does having too little definition of the research to engage an evaluation team on a timely basis (3 months prior to D-Day--6 months if conducting a formative evaluation).

[6]For example, the sponsor could hire evaluation specialists to develop the technical "blueprint" for the study, based on principles in sections 4, 5, 6, and 7 below. Next, an evaluation contractor could be brought on board to refine and implement the plan and report findings.

- ***Use this Guide and other resources to understand the technical issues in each research area before discussions begin*** on that component of the evaluation. This does not mean mastering the technical details of the theories and solutions surrounding an issue, only recognizing the basic question, understanding what is at stake, and grasping a rough outline of possible methodological responses.

- ***Participate actively in all design discussions concerning the evaluation or*** the structure and implementation of ***the demonstration*** project and intervention. The latter can have major implications for the former, and always needs to be considered from a research perspective before anything is set in stone.

For those not expert in demonstration research methods, summary volumes on evaluation methods such as this *Guide* and other works can lead sponsor staff through this process, serving as primers, prompts, and reference documents at different stages of the process.

3. Special Challenges for Sponsors

Beyond this section, the *Guide* will focus on methodological aspects of evaluating employment and training demonstrations and pilots. But first, the current section addresses three topics of a more practical bent -- topics of great interest to demonstration sponsors when guiding and monitoring evaluation activities:

- What cuts in evaluation scope are appropriate when budget constraints force a "leaner" study than the sponsor might like?

- Where should a sponsor agency focus its attention once a demonstration evaluation is underway?

- Is planning and overseeing evaluations of pilots and demonstrations with the level of commitment and sophistication described here worth all the trouble?

Answers to these questions should incorporate, among other things, the principles of top-quality evaluation research practice. This section offers some thoughts on what those principles imply.

3.1 Choosing the Right Cuts

As noted already, most evaluations begin with ambitions that exceed their funding base. Section 2 provides a framework for correcting this imbalance through tough decision-making on study scope and the prioritization of research goals. The main theme there bears repeating:

The first and best response to overly-tight budgets is to reduce a study's ambitions by dropping some of its research goals (preferably those with the lowest priority) and--at the same time--continuing to fund the rest of the research agenda at a level that ensures its success.

Strong, thorough research on most of the key questions surrounding a new policy idea will make a much larger contribution to public discourse on it than spotty, suspect research on all questions.

But what if scope has been cut and the budget still looks too limited for the remaining scope. Such situations can occur under at least two scenarios: ambitions and research goals have been cut too little or an adequately-funded evaluation hits financial difficulties part way through its execution. What steps can evaluators and sponsors take in these instances to hold down costs while preserving the topical coverage of the study? Besides management economies (i.e., doing everything with less through increased efficiently), the research team should strive for creative technical solutions to the funding shortfall, based on the intersecting roles and complementary nature of many evaluation tasks. Prime options for savings of this sort include:

- ***Elimination of analyses that do the same thing two different ways***, such as the use of both focus groups and "customer satisfaction surveys" to document participants' views of demonstration services for the operational evaluation described in section 6 below. As noted there, with only one of these methods, findings will not be as rich or nuanced as before, but findings will not be entirely absent either. The original two-pronged attack is not absolutely essential unless testing an intervention that makes client outlook its primary measure of success.

- ***Confining analysis to the highest level of study possible for each aspect of the demonstration studied***. Evaluation of participant outcomes provides perhaps the best example of this strategy: rather than examine all three of the outcome dimensions discussed in section 7 below--levels, changes, and impacts--determine how far the evaluation can take this sequence technically and do just that level. In other words, if a credible ***impact*** analysis is possible, do it alone since no one will care terribly about outcome levels and changes if the demonstration is shown to have no impact on participants. Similarly, where circumstances preclude impact analysis, but not other options, do just the ***change*** analysis for outcomes; again, from the perspective of the intervention, few will care about ***levels*** of outcome if they do not change over the course of participation. This form of cost cutting makes most sense when--as in the outcomes example--completion of the higher-level analysis requires collection of the principal data needed at ***all*** levels, leaving open the possibility of someone else eventually looking at lower-level measures should they remain of interest.

- ***Conducting formative evaluation on a highly selective basis.*** All of the formative research activities described in section 4 below have value in almost any setting. But with a tight budget, each can be considered as a stand-alone option. When strapped for funds, it makes sense to do just the one or two that will contribute most to refining the particular intervention under study. This does not mean the idea of formative evaluation should be dropped entirely: if there is at least one formative check with a large potential payoff, do it and cut elsewhere, remembering that good ex-post research on a poorly-conceived and raggedly-implemented policy is worth little.

- ***Compression of task schedules***. When targeted on the right tasks, shortening the implementation of certain evaluation tasks can eliminate an appreciable share of their total cost. Specifically, costs associated with task leadership and supervision are roughly proportionate to the number of months the task runs, regardless of the intensity of activities in any given month. Also, the inefficiencies of research support staff moving back and forth between tasks can be reduced when a single task--at a faster pace--occupies their full-time attention. Planners should look particularly hard at participant surveys for savings of this sort; as noted in section 7 below, when interviews are confined to fewer months (at higher volume) total overhead costs go down. In addition, some highly complementary evaluation activities can be overlapped, such as description and normative assessment of pilot or demonstration project operations, both of which involve substantial amounts of on-site time and other shared resources (e.g., site MIS data). However, there are also hazards to this approach to cost-saving if not properly targeted (see below).

- Though often shunned (at least during an evaluation's design phase), ***reductions in target sample sizes*** --for demonstration participants and, particularly, for follow-up survey respondents--should be put on the table when facing an under-funded situation. For example, sizeable cuts in the number of observations impose relatively little damage to statistical reliability if sample sizes are toward the high end of the usual range to begin with (see section 7 below).[7] Unlike most cost-based sacrifices regarding research technique, here the loss is not only modest but ***quantifiable and unambiguous***,[8] traits that make possible careful, well-informed--and much safer--decisions on economizing. Moreover, there is no formal basis for deciding which levels of statistical reliability are acceptable and which are not, nor any conventional standards. Large samples that look adequate initially to a given researcher tend to look acceptable in terms of uncertainty when shrunk by one quarter or one third.[9] In evaluations with two or more waves of survey data collection, a cut of this magnitude can save a good deal of money, while sample size reductions for just a single survey can make an important contribution.[10]

There are also two "don'ts" to bear in mind when looking to economize. With the temptation to sacrifice as little as possible, these "don'ts" warn against viewing ***illusory*** measures as real solutions to budget woes:

Don't drop a line of research or an entire evaluation type and count it as savings expecting that "a minor addition here, a trivial extension there" and so on across ***the remaining portions of the plan will put back what was sacrificed "on the cheap."*** Synergisms among analyses are never as real as they seem to be when scouring a work plan for budget savings: without lowering goals, the work that has to get done still has to get done even if it is changed from a separately designated task to an appendage on another task. And experience suggests that, in the end, it ***will*** get done, but using pretty much the same means--and the same level of resources--as in the original plan.

---

[7]For example, a 50-percent reduction in sample size does not double the minimum detectable effect (MDE), as one might expect. Rather, MDEs go up in proportion to the ***square-root*** of the sample size decline. This square-root translation has a substantial mitigating effect on percentage changes in MDEs: the MDE in this example, rather than doubling (i.e., increasing by 100 percent), goes up only 41 percent.

[8]The quantifiable consequences of sample size reductions are usually expressed by evaluators as a percentage increase in minimum detectable effects, or MDEs, for the impact analysis (see section 7).

[9]A one-fourth reduction in sample size increases MDEs by just 15 percent, while a one-third reduction increases MDEs by 22 percent.

[10]Analyses of the entire participant pool based on administrative data suffer even less from sample cuts of a given percentage, since they start out with larger samples (often much larger samples). Little money will be saved by sample size cuts here, given the trivial marginal costs of adding another case to administrative data file extracts (see section 7 below). However, a demonstration's operational costs should fall substantially if the overall participant group shrinks as well.

In a similar vein, ***don't combine two full tasks expecting economies of scale to lower total costs***, then fund at that lower level. This is wishful thinking: economies of scale in evaluation research-- to the extent they occur at all--occur ***across*** projects, when carrying out the same function over and over. Changes such as putting MIS data collection and process analysis site visits together under one task sound great but often do not integrate in practice, often because of non-overlapping personnel. Thus, ***economies of scale are mostly, if not entirely, illusory (or even counter-productive) when evaluators are forced to work on two conceptually different goals within a single task structure***. Pretending otherwise will simply strain relationships between the sponsor and the research team and yield disappointing results for everyone.

Collectively, these "dos and don'ts" may lead to real and substantial cost reductions while diminishing the strength of the research and the value of its results by a modest and acceptable amount. Sometimes, however, the problem is larger, and sponsors face a tougher question:

"What is the minimum amount of evaluation research needed to make a demonstration worthwhile at all?"

Obviously, there is no "one-size-fits-all" answer to this question--it depends on the intervention and the policy questions to be addressed. Still, there are some research minimums common to almost all demonstrations. These, in turn, translate into ***a set of guidelines for assuring that even the most tightly constrained pilot or demonstration evaluation makes the most of its money but--in the extreme case--avoids investing in a hopeless cause.***

Minimum Funding Guideline 1: *If an agency cannot afford at least some normative analysis--a piece of an operational evaluation, or a part of an outcome evaluation--it should not do the evaluation or the demonstration at all.*

Without normative analysis to provide clues as to the ***desirability*** of what a pilot or demonstration intervention does, a simple description of its activities in a descriptive evaluation provides no "actionable" information. It is only in judging activities helpful or hurtful in some way--or capable of improvement--that the government, and therefore society, gains from its investment in research. Without normative follow-up to gauge, judge, and improve project operations or participant outcomes, no part of an evaluation should go forward. The sequel to this conclusion is that--lacking an evaluation to impart useable lessons--the demonstration itself has little purpose.

Minimum Funding Guideline 2: *Any normative analysis, whether operationally or outcome-focused, must be accompanied by a descriptive evaluation.*

An operational evaluation simply cannot be done without the programmatic information provided by descriptive study. In contrast, an outcome evaluation would be possible, but it would have little meaning. This is the famous "black box" problem: simply knowing that an intervention worked for participants gives a sponsor no advantage in formulating future policies if the intervention itself is in a "black box," invisible and unknown.

Minimum Funding Guideline 3: *If forced to choose between the two types of normative evaluation, operational or outcome, choose operational as the more economical.*

This oft-debated question has a clear answer and a simple two-step rationale. As just noted, descriptive evaluation is essential for either type of normative research—operational or outcome—to make a useful contribution. Once a descriptive evaluation is done, the added effort to draw out normative conclusions about demonstration ***operations*** pales in comparison to what it takes to draw normative conclusions about participant ***outcomes***--essentially because the later involves major amounts of follow-up data collection, including one or more high cost participant surveys.

Minimum Funding Guideline 4: *Stay flexible and don't commit funds until you have to.*

Things will change over the course of a medium to long-term demonstration evaluation (those two or more years in length), making the flexibility to reallocate research dollars across tasks a particularly valuable commodity. Thus, when facing a choice of what to cut at or near the beginning of a study, sponsors should favor the course that holds open the most options for the future. This principally means scaling back on baseline data collection as much as possible and preserving resources for later decisions. Viewed in hindsight, long-term evaluations almost always end up "cutting-corners" of necessity in later years, having failed to make serious sacrifices in the first two years--often to the regret of sponsor and evaluator alike.

3.2 Monitoring the Evaluation: What to Worry about When

As a further aid to sponsors, Exhibit 3.1 lists the areas of evaluation implementation that are most likely to encounter technical (as opposed to budgetary) difficulties.[11] If looking for a way to invest their time, oversight staff at DOL/ETA would do well to focus first on these areas (if the referenced evaluation component is part of the study) during each of the time intervals indicated. As can be seen, technical threats to evaluation research concentrate in the first phase of a study, as many types of data collection need to begin at once, on schedule, and in sync with demonstration implementation in the field. This reinforces a message emphasized earlier: sponsors **must** have their evaluation team in place and operational at least 3 months prior to the beginning of demonstration operations. This accomplishment is critical to seamless, reliable data collection in the early going. Sponsor agencies should also consider assigning strong, highly experienced oversight staff to the project during this period, with ample time allocated to that role.

[11]The list begins at the start of demonstration operations. Section 2 above offers advice on sponsor priorities during an evaluation's planning phase.

Exhibit 3.1

Potential Hazards at Various Stages of the Evaluation Process

Potential Hazards at the

START-UP STAGE

of the Evaluation Process

- Delays in obtaining evaluator feedback on early pilot or demonstration events for formative evaluation
- Research staff shortages for initial descriptive site visits
- Late or incomplete collection of baseline data on participants

  --unexpected omissions from site MIS data
  --late or erratic use of evaluator-provided data collection forms during demonstration intake
  --delayed start-up of baseline personal interviews
  - late questionnaire development
  - late OMB submission
  - interviewer recruiting and training problems
  - difficulty establishing sample frame
- (for experimental impact studies) Random selection procedures for control group not ready when demonstration enrollment begins
- Loss of pre-demonstration earnings, welfare, and/or Unemployment Insurance benefit information removed from State data systems before extraction process begins running smoothly
- Legal issues surrounding confidentiality and release of administrative /MIS data

Exhibit 3.1 *(continued)*

Potential Hazards at Various Stages of the Evaluation Process

Potential Hazards at the DATA COLLECTION STAGE of the Evaluation Process

During Pilot/ Demonstration Operations

- Difficulties compiling and interpreting participant-level data from many different data sources.
- Gaps in participant-level data when building the research database, including problems with personal identifiers needed to link files.

Post-Demonstration

- Insufficient time to prepare for participant follow-up survey (less than 12 months advance).
- Tracking and response rate problems in conducting follow-up interviews.
- Premature cut-off of the follow-up period for late demonstration entrants as time is squeezed between final data collection and preparation activities and the end date for the study.

Potential Hazards at the REPORTING STAGE of the Evaluation Process

- Analysis staff turnover.
- Inadequate agency-level review of interim findings.
- Inattention to the public release of interim findings and efforts to ensure appropriate technical interpretation of the results.
- Incomplete coverage of intended research topics due to time, budget, and data constraints.
- Delayed or incomplete preparation of public use data files as research staff members with the greatest knowledge of the data's origins, structure, and content move on to other assignments.

A second theme runs through the list as a whole: data collection and data processing pose by far the greatest challenges and risks to demonstration and pilot project evaluations.[12] Though these may seem like dry and semi-rote tasks, inattention to data work on the part of the sponsor can be very costly. The best counter to these threats is to attack each new data issue aggressively as soon as it arises. So many data challenges can arise in a large, complex study that the research team finds itself with an almost overwhelming backlog of "data salvage" activities on its hands. Quick and sustained support from the sponsor agency greatly reduces this risk. In the extreme, multiple data problems threaten the forward progress and long-term success of all evaluation activities. While this is rare, in a complex evaluation it remains a possibility during much of the study period.

3.3 Why It's All Worthwhile

Good evaluation practice, and successful results, require sustained commitment, vigilance, and high-level expertise in a mix that only the demonstration sponsor can put in place. This places high-level demands on the agency involved, as DOL/ETA knows from its support of several top-quality evaluations in the past. But as will be emphasized throughout this *Guide*, even stronger, more thorough going pursuit of the quality goal is both possible and desirable. To inspire such an effort, this concluding portion of Part I of the *Guide* deals with an important question: Is all this effort worth it? The answer, simply put, is "yes" for a number of powerful reasons.

First, ***a new intervention tested once may never be tested again--it will be judged a success or a failure based on this one set of demonstration or pilot results.*** If deemed a success and adopted nationwide, years or even decades may pass before any re-test takes place. If deemed a failure, the policy approach will quickly be forgotten. So mistakes on the initial verdict are always costly. If, for example, a sub-par evaluation concludes that an ineffective intervention actually works, the process may run exactly counter to core of the demonstration philosophy: rather than making sure an intervention works before undertaking nationwide implementation, the study will have hoisted the policy into the ranks of "proven good ideas" and advanced, rather than retarded, its progress toward national implementation. If instead a poorly-done evaluation reaches a negative conclusion about what is, in fact, an effective intervention, a rare "winning" idea may be abandoned forever. Strong evaluation, on the other hand, settles the question of worth correctly and decisively--or at least points in a helpful, not a misleading direction. When taken to their logical conclusion, these observation imply that ***policy tests that cannot be done well should not be done at all,*** lest they legitimize failed interventions or foreclose successful policy ideas. If evidence is to be a tool, decision makers need the best evidence available.

The very process of doing high quality research provides a unique opportunity for agency staff to learn-as-they-go about new policy ideas and how the test intervention plays out in the field. In an ambitious, well-crafted study, agency learning of this sort is inevitable from beginning to end, starting with concentrated thinking about the objectives and expected paths of influence of the intervention and continuing as research activities and findings unfold. Along the way, involvement in better studies will enable Departmental

[12]This is not necessarily true for studies that cover only a few evaluation components, which will not necessarily face even a majority of the challenges on the list.

staff to develop stronger information-based theories regarding a test intervention's strengths and weaknesses and to use that input to develop clearer ideas regarding future policies worth testing. In contrast, a muddled or cursory evaluation will do more to mystify than illuminate.

When its research covers the right questions and provides staff with close-quarters familiarity with results, ***project managers will be strongly equipped to respond to any challenges to a study's substantive conclusions***. The Department will also acquire a heightened ability to counter technical challenges to the research using the expanded in-house knowledge that results from strong evaluation involvement. Finally, sponsor staff in a state-of-the-art evaluation can rest assured that the research stands on solid methodological ground when calling upon the research team to answer technical questions for an outside audience in greater detail.

In short, by doing pilot and demonstration evaluation well, the Department will find the result of its added spending and effort to be not frustration and vulnerability, but solid, demonstrable accomplishment in one of its key mission areas: learning from test experience.

Part II: Research Methods

4. Formative Evaluation: Fixing Problems Before They Start

As noted in section 1, evaluations of DOL/ETA pilot and demonstration projects can take several forms:

* ***Formative evaluations*** use pre-existing data to check the "evaluability" of a proposed demonstration innovation and to help planners refine the intervention before it is implemented.

* ***Descriptive evaluations*** document the steps taken by local agencies to create and run the intervention, and summarize the characteristics and activities of demonstration participants.

* ***Operational evaluations*** (also called process, or implementation, studies) assess the operational strengths and weaknesses of the intervention once implemented, and consider ways the approach might be strengthened in future replications.

* ***Outcome evaluations*** track participants following demonstration exit to monitor their labor market outcomes (levels and changes) and assess the short-run and/or long-run impacts of the intervention on participant outcomes and overall social benefits and costs.

Most pilot and demonstration studies include at least two or three of these evaluation types. Section 2 above provides guidelines for determining which type or types are most appropriate to particular situations, based on the research goals of particular studies. The current section, and the three that follow, define and illustrate each of the four research types in turn, describing research approaches and analytic methods that maximize what can be learned from pilot and demonstration evaluations. Discussion focuses first on formative evaluation. For this evaluation type--and others covered in later sections--researchers do not have to undertake all of the component analyses described for the individual components to be useful.

Formative evaluation (also called pre-project analysis) occurs at the front-end of a demonstration initiative, before the new innovation is actually implemented in the test sites. The goal is simple: to give the new intervention--and its evaluation--the greatest possible chance of success. Despite its potential to improve pilot and demonstration tests, formative evaluations are probably the rarest--and certainly the least visible--type of demonstration research conducted today.[13]

[13]A broad sweep of the literature on evaluations of DOL/ETA pilot and demonstration projects over the last 20 years found no examples in the public domain of written formative evaluations.

Formative evaluation draws on several information sources:

-- the ***opinions and intentions of the innovation's creators***, to identify the intervention's intended activities, target population, and policy objectives;

-- ***existing information from outside the demonstration***, to consider the potential and possible pitfalls of the intended intervention; and

-- ***the very earliest steps of demonstration implementation***, to uncover real-world problems and opportunities that could not be anticipated without actually moving to create the intervention.

These sources may point out desirable traits of the new test policy before it fully takes shape, or--if the intervention has already been specified in operational detail--flag weaknesses in the plan. Planners then can devise and test a better version of the new policy concept, making the demonstration evaluation more valuable to policy makers and other interested parties. Formative evaluations can also strengthen the technical aspects of other types of evaluation research. Demonstration planners need not include ***all*** of the types of formative analysis described below, though a case can be made in each instance. More realistically, the discussion here provides a "menu" of options for improving pilot demonstrations through pre-demonstration research.

4.1 Evaluability Assessment

Anyone who sets out to understand and report on a pilot or demonstration intervention needs to know--and lay down in writing--what the intervention is all about--who it will serve, what services it will render, and what it hopes to accomplish. Agencies sponsoring pilot and demonstration initiatives almost always cover these basics themselves while obtaining authorization and funding for an initiative. However, even when a written record of these features exists and can be incorporated into the evaluation, it may not be as clear and detailed as is needed to guide implementation and evaluation planning.

This raises basic questions regarding the "evaluability" of a proposed policy or program that must be settled before it is tested:

Has the proposed intervention been specified in enough detail, conceptually, to allow for unambiguous statements about its nature and relationship to other distinctive policies, once research results are available?

Does the intervention as planned have a reasonable chance of being "implementable"--i.e., is it capable of execution in a real-world setting in rough conformity to its intent?

Will enough observational information about the intervention be available to allow for descriptions of its nature and judgments as to its success or failure, strengths, and weaknesses?

The best way to begin any demonstration project is with an "evaluability assessment" that addresses these questions, based on a hard-nosed projection of whether the proposed intervention can be understood conceptually, implemented faithfully, measured reliably, and interpreted unambiguously in relation to other policy options. If not, actors in the policy arena will have trouble knowing what to make of the policy tested even with

substantial investment in its implementation and study. If the answer to any of the above questions is "no," the next step is to identify changes that would right the situation, or to determine that the demonstration should not go forward. Even when the conditions of "evaluability" are met, the thinking needed to reach that conclusion can alter the intentions and opening steps of a demonstration in important and beneficial ways.

To further ensure a clear, clean policy study, the demonstration's sponsor should provide a written definition of the intended policy intervention to one or more employment and training evaluation experts to see if it is sufficient to support clear-cut research. Employment and training practitioners also should be consulted as to the feasibility of constructing a real-world intervention that reflects the design. The "nuts and bolts" of this process are described succinctly by Wholey (1994).

Sponsors will also benefit from working closely with both groups to identify the indicator measures on which demonstration progress can be judged and the evaluation focused. Three traits are essential in any such measure:

- It must cover a factor of clear importance to the demonstration's target population or society in general (e.g., quantity of services received, speed of service delivery, labor market outcomes for participants);
- It must be capable of being influenced by the demonstration intervention; and
- It must be measurable on a reliable, consistent basis across multiple years, individuals, and localities.

To maximize the influence of study results, sponsors should also seek indicators that are:

- Widely used and accepted in similar evaluations--evaluations concerning similar programs or policies and similar target groups.

The last two of these criteria--"measurability" and broad acceptability--draw directly on the knowledge base of professional evaluators, while the first two can best be informed by policy makers and program operators. Whether found within or outside the sponsor agency, expert knowledge of this sort can be particularly valuable at the very beginning of a demonstration initiative.

4.2 Examining the Target Population

Once definitional and feasibility problems have been resolved, a number of other dimensions of the demonstration or pilot intervention should be examined prior to implementation. For example, to check that the intervention makes sense, one needs to understand the ***target population*** to which the new policy will be applied, both in terms of that group's background characteristics and its behaviors under existing, non-demonstration policies. Sponsors can then use this information to identify problems in the planned intervention and take early steps to correct them.

Often, the projected accomplishments of a demonstration intervention hinge on assumptions about the target population that, on closer inspection, turn out to be unfounded. These include assumptions about:

- the number of potential program participants,
- the pre-demonstration circumstances of potential participants, and
- the likelihood that the participant behaviors needed for the intervention to attain its goals will actually take place.

There are many examples of how analysis of pre-existing data has helped—or might have helped—to refine the test "treatment" before a demonstration intervention began. Boxes A, B, and C at the end of the section describe three such instances, one for each of the three types of testable assumptions mentioned above.[14]

Most demonstration projects present one or more opportunities to use outside data to check underlying assumptions. Often, the ***assumptions*** are as hard to identify as the data needed to test them. When the assumptions that make one policy approach more appealing than another are not obvious, sponsors and evaluators need to spend some time "digging" them out.[15] The best way to do this is to:

- Detail the anticipated actions and accomplishments of the pilot or demonstration, and
- Ask the demonstration's creators for their vision of the mechanisms that connect the two--i.e., the linkages, or paths, between policy actions and participant outcomes.

By definition, these interconnected, causal linkages capture the assumptions that led the policy's initiators to believe the intervention would work, and that therefore need to be examined prior to the demonstration itself. When a demonstration's creators cannot specify exactly what they have in mind, economic theory and other behavioral science models (e.g., from sociology or psychology) can help fill the gaps. These paradigms are ideal for developing possible explanations of how policies (inputs) produce results (outcomes) through human behavior. Any predictions they produce that have participants moving in the intended direction may be a crucial underpinning of the intervention and, thus, worth examining in advance.

Actual analysis of these assumptions can take many forms, too many to summarize here. Therefore, it is not possible to specify the best methods for this type of advance research. However, one can specify the generic skills needed carry out this type of research successfully:

---

[14]Example A describes successful use of external data in a DOL/ETA demonstration. Examples B and C point out how extra target group analyses could have helped refine two other employment-related demonstration interventions outside the Department.

[15]Even if skipped at the beginning, the effort to clarify implicit assumptions about participants often becomes a necessity later in the evaluation, when researchers attempt to explain ***why*** the results turned out as they did. Possible reasons for certain outcomes begin with hypotheses about how the intervention was ***expected*** to interact with the target population. "Surprise" findings at this point usually stem from unrecognized, or heretofore misinterpreted, assumptions about target group behavior—something that evaluators must get straightened out before publishing their findings.

- Good knowledge of available individual-level data bases, and
- Experience modeling inputs and behaviors of the target population, whether that be drop-out youth, dislocated workers, or some other DOL/ETA constituency.

Assuming a formative evaluation takes place, demonstration sponsors should engage an evaluation team--either internally or externally--that meets these requirements.

4.3 Learning from Earlier Studies

Decision makers can also benefit from formative evaluations that review findings on previous, similar interventions and translate them to the current setting. Of course, no other existing study can substitute fully for a new demonstration and evaluation of the exact policy in question. But program evaluation is expensive and should be "recycled" from past studies when germane, both to redeem the cost of those evaluations and--of greater importance--to save the expense of a ***new*** demonstration study when its findings can be anticipated with a good deal of confidence. In other words, there are times when a strong *a priori* case can be made that a proposed demonstration intervention is likely to achieve its goals (or fail to achieve them) based simply on experience with similar interventions. Example D at the end of the section provides one such example.

Occasionally, a new pilot or demonstration test may be found unnecessary based on past research. By acting on this information, a sponsor agency can turn its attention to where uncertainty is greater. Of course, a new replication of a previously successful intervention is not guaranteed to produce the same result, even when it seems identical. As a result, it is often worthwhile to run a pilot or demonstration "replication study" just to check whether the initial findings hold up in other settings.[16] But in other instances--where a general approach has been tried and studied two or more times--prior research can provide powerful hints as to what to expect from a currently proposed intervention. Such hints can be especially valuable to an agency that has two (or more) competing policy proposals on the table--both untested in their current form--but not enough resources to demonstrate and evaluate both. Clearly, such an agency is better off letting prior research guide its recommendation of which of the proposals should be tested in a new demonstration initiative--the one least informed by prior research.

Sponsors can rely on past evaluation findings in one more way: to identify the policy models most worth testing next. Then, if the currently proposed demonstration looks considerably different from the most promising test option identified in the literature, a shift in approach might be in order. Sometimes, the existing wisdom on promising policies for a specific target population (e.g., mid-career women, tribal groups) has already been assembled in a recently published literature review or in a "think piece" by one or more scholars.[17] When grounded in past research in this way, these summaries provide an excellent basis for making and defending one's choice of the policy innovation most worth testing. If no such document exists, commissioning one as part of the formative evaluation can prove a very wise investment.

---

[16]This is being done now, for instance, in DOL/ETA's Center for Employment Training (CET) replication study (Walsh, 1996) and the Quantum Opportunities Program replication (Marlani and Maxwell, 1999).

[17]For example, Gueron and Pauly (1991) produced a comprehensive and influential document of this sort regarding employment and training programs for welfare recipients.

4.4 Reanalyzing Existing Evaluation Data

A third area of demonstration refinement supported by formative evaluation reanalyzes existing evaluation data to predict the success of a proposed new initiative. Large-scale demonstrations and program evaluations often include within them narrower evaluations of potential relevance to subsequent demonstration undertakings. For example, a demonstration proposing to provide employment and training services to non-resident fathers as a means of increasing child support payments could benefit from a closer inspection of the National JTPA Study data. That study included over 1,100 low-income men living apart from their children, some of whom received JTPA services (the "treatment" group) and some of whom did not (the "control" group).[18] If JTPA's standard services did not raise earnings for that group, a demonstration that extends these services to a larger share of absent fathers is probably not worth testing. Instead, a program that both reaches out to the absent father population and, at the same time, provides ***new*** service types would seem the stronger candidate for testing.

Current demonstration practice may routinely overlook opportunities for guidance of this sort, with the increasing number of large evaluation databases available. Unfortunately, this oversight is not easily remedied. To take advantage of all such options, one first must identify prior evaluations that could include (as a subgroup) the target group and intervention of interest. Then questions arise as to whether separate analysis of those individuals is feasible and statistically reliable, and whether the data can be obtained.

In one important instance, the goal of identifying studies of potential relevance is easily accomplished: when the main goal of an evaluation is to measure the impact of a demonstration intervention on participants. In this case, one can rely on the *Digest of Social Experiments*,[19] which provides a comprehensive review of all social program impact studies ever conducted using random assignment research methods.[20] Descriptions of interventions and target groups in this volume provide a good understanding of the scope of each study, though the presence or absence of specific subgroups usually cannot be determined. For other types of analyses--descriptive, operational, and non-impact outcome evaluation--no such summary exists (to the author's knowledge). For those evaluation types, one must turn to someone with up-to-date knowledge of the literature, either within DOL/ETA or outside the Department.

---

[18]This example has actually been actuated, when the author developed estimates of JTPA's effects on absent fathers (and other fathers) using the National JTPA Study data, for a recent Ford Foundation conference on "Improving Labor Market Outcomes for Poor Fathers." Neither set of fathers experienced statistically significant increases in their earnings over 30 months compared with control group members excluded from JTPA. Sample sizes equal 1,177 for absent fathers and 1,559 for resident fathers.

[19]Greenberg and Shroder (1997).

[20]See section 7, below, for details on random assignment impact analysis. Randomized trials are widely viewed as producing the most reliable impact estimates possible for employment and training programs; in the view of many evaluators and government officials, they are the ***only*** credible estimates available concerning demonstration effects. In this sense, the *Digest of Social Experiments* can be viewed as comprehensive of all the impact studies of interest—those reliable enough to be used in place of a new demonstration evaluation.

Regarding feasibility and sample size, the only way to be sure about a re-analysis is to contact someone from the original study team.[21] This may present the first opportunity to check on the inclusion of the pertinent subgroup, and it is often the only way to determine whether that subgroup can be isolated within the overall study and whether the data can be shared. For descriptive and outcome analysis, the ***number*** of subgroup members in the sample carries critical importance (see section 7.5 below) in determining whether ***reliable*** analysis can be done.[22]

In almost all cases, the actual re-analysis of prior evaluation data can be accomplished most efficiently by the original evaluator (again, either inside or outside the agency). The effort may become prohibitively expensive when someone with no prior experience with the data set is assigned the task; also, the risk of errors and misinterpretation of findings goes way up when engaging someone from outside the original study team. With a sufficient investment--and for re-analyses of sufficient importance--it is possible, of course, to complete a reliable re-analysis using personnel other than the original study team, as long as those chosen have strong general capabilities in conducting the type(s) of research sought.

4.5 Feedback on Early Demonstration Events

Important lessons may be gleaned from a demonstration's early implementation experience, after it starts but before it reaches scale--information that can further improve the project's design and procedures. Gains can be substantial in two realms: in the implementation and management of the intervention itself, and in the evaluation research. At this early stage, the sponsor agency or evaluator looks for signs that things have gone off track, or picks up suggestions from the field on how operational procedures could be improved over the original plan. These are good things to know early, whether they are actionable or not at that stage. Early "course corrections" that offset off-track or unanticipated developments can have tremendous value over the remainder of the demonstration, and newly-recognized ways to strengthen the research--or the intervention--can be nearly as valuable. In each instance, the sooner the correction is implemented, the more the demonstration benefits.

Unwanted developments early in implementation, even when not correctable, can help identify new analysis needs on the fly, either different types of analysis or adjustments in data collection. Even where research needs don't change, evaluation ***procedures*** might have to.[23] Either way, early vigilance and awareness are essential to keep a demonstration running smoothly. In the extreme, an operational deviation could be discovered

[21]For extramural evaluations, the "study team" includes both lead personnel at the contract evaluation organization and the cognizant agency personnel (e.g., the Project Officer). Greenberg and Shroder include contact information on these groups in their *Digest.*

[22]Even the final report of an impact evaluation will not necessarily contain all (or any) of this information, and most such reports include very little.

[23]As an example, consider a demonstration run by JTPA in a large urban site where a major source of client referrals for the project (say, the state Transitional Assistance to Needy Families, or TANF, program) quickly departs from plan by sending potential participants to JTPA intake points scattered throughout the city, rather than routing them all to the main JTPA office for that service delivery area (SDA) as had been planned. This change would necessitate the addition of an item to the demonstration's intake form to indicate the neighborhood office that received the referral. More substantially, it might imply that the study team visit many local offices to document intake procedures, rather than just one as expected.

that is so large and intractable--and carries such sharp consequences for demonstration procedures or research--that the policy test should simply be abandoned altogether. Fortunately, few precedents exist for this "train wreck" scenario (perhaps none at DOL/ETA). And, as should be apparent from the above discussion, an early end to a demonstration project is particularly unlikely when a formative evaluation has been set up to anticipate and avert problems of this sort before implementation begins.

So how does one monitor early implementation to detect untoward developments at the very beginning of a pilot or demonstration project? Often, the sponsor agency will be in the best position to keep abreast of developments in demonstration operations, especially if--as is common--it has the local organizations implementing the intervention under contract. By staying in close touch with the DOL/ETA staff responsible for overseeing local demonstration contracts, the evaluation team can receive rapid feedback on any start-up events with implications for the research. Some of these events may not be ***recognized*** as research problems unless the evaluator is put "in the loop" in this fashion--on an almost daily basis in the early phases.

There are also situations where the evaluator--if extramural--has good or excellent access to demonstration implementation information itself. These "natural openings" should be exploited whenever possible to complement the oversight-based feedback just discussed. Opportunities of this sort arise when:

- the evaluator ***administers*** the intervention, as occurred in the National Supported Work Demonstration and the Lifelong Learning Demonstration[24];
- the evaluator has worked with one or more of the local demonstration agencies on other previous studies, earning trust and access to knowledgeable sources; or
- the evaluator is viewed as a neutral party in instances in which relationships become strained between a local implementation agency and the federal sponsor, for reasons unrelated to the demonstration.

Particularly in this last situation, information passed on to an outside evaluator may be more accurate and complete than that reaching the federal oversight organization.

[24]In the former case (Hollister et al., 1984), the evaluator created field offices in the demonstration sites that conducted outreach, intake, and service delivery; in the latter (Bell et al., 1996), the evaluator designed and implemented the informational campaign that constituted a major part of the treatment. While offering communications and access advantages—and often leading to much more "evaluable" interventions—situations like this create the potential for a conflict of interest as researchers are asked to determine the success of their own creations. The lack of objectivity this implies need only be perceived—and not necessarily real—for this arrangement to undercut a demonstration's credibility, although there are instances in which such arrangements are appropriate and especially advantageous.

Whoever does the monitoring, there is no magic formula for covering all the areas in which early implementation issues could arise, short of conducting in-depth information-gathering visits to the demonstration sites on a regular basis. While the site-visit approach would provide maximum protection against unwanted operational events going undetected throughout the demonstration period, it could also be quite costly--especially in multi-site demonstrations. A more affordable strategy combines the off-site monitoring steps discussed above with the capability (and will) to respond rapidly and decisively to any serious issues that do arise. In-between options might also be considered, including:

- Brief on-site checks of just those procedural steps in which breakdowns seem most likely or would have the most costly implications; and

- One or two full-scale site visits in locations where procedural breakdowns seem most likely, given the history of the local demonstration agency and indications of dissonance or reluctance during the demonstration planning phase.

Beyond these steps, DOL/ETA has gone one step further on some of its test initiatives by running a "pilot within pilot." This "pre-pilot" approach means starting the intervention ahead of schedule on a very small scale for the exclusive purpose of catching operational "glitches" in a test-run of implementation. The events that occur--and the participants served--during this initial test phase ***are not included*** in any other evaluation components. They are essentially "throw-away" data points for the main demonstration study and will not be included in the project's main reports or used in developing its policy recommendations. Thus, nothing that goes wrong in the "pre-test" phase can skew the demonstration's main mission of bringing forth information on the **best** version of a new policy idea based on field trials. As a result, data gathering during the pre-test can focus exclusively on detecting problems or surprises, and any responses can be more considered and complete--and fully in place before implementation of the main demonstration begins.

Pre-tests are clearly the most comprehensive and effective way to benefit from the lessons of early implementation. They are not necessarily that expensive, depending on the nature of the intervention. The actual ***number*** of pre-test participants can be quite small; far more important is that all ***steps and components*** of the intervention be carried out for at least ***some*** people. This of itself may be expensive for interventions with high up-front costs for recruiting and intake or low final participation rates among those recruited, and for interventions implemented by newly-created units or divisions within a parent agency. This latter circumstance can sharply escalate the costs of a pre-test simply by extending the total amount of time that special units must be maintained and operated--an increase of from 2 weeks to 3 months, typically.[25] If affordable, however, special test-runs can do more than any other mechanism to check whether "all systems are go" and to find ways to improve those systems that are not fully functioning before any of the "real" demonstration takes place.

---

[25]The Project NetWork Demonstration conducted by the Social Security Administration ran a 2-week pre-test of intake and random assignment procedures within specially created demonstration units (Rupp et al., 1994). Main demonstration intake began within a week after that since no serious problems were encountered. In contrast, DOL/ETA's Lifelong Learning Demonstration (Bell et al., 1996) went 6 months between its pre-test and the main intervention, for two reasons: the treatment (mailings encouraging incumbent workers to upgrade their skills through education) only made sense every 6 months, in conjunction with the academic calendar, and the pre-test looked not just at treatment administration but at the first-level response of the target population (replying to the outreach mailings) which continued over several months. In the meantime, no special demonstration unit had to be set up. It is also important to note that the added round of outreach required by pre-tests cost very little in both these instances (all outreach was done by mail), another factor to consider in deciding whether to do a "dry run" of a demonstration intervention.

Example A

Using Outside Data to Check the Size of the Participant Population

The Lifelong Learning Demonstration, co-sponsored by DOL/ETA and the U.S. Department of Education, tested the impact of a public information campaign encouraging mature, incumbent workers to go back to school (Bell et al., 1996). Early in the demonstration's planning phase, a decision was made to conduct the campaign through promotional mailings to the homes of thousands of mature, incumbent workers. Which workers to target in the mailing was not so easily resolved. To make this decision, demonstration planners used external data from State Unemployment Insurance wage files and commercial marketing firms to simulate various population sizes under alternative targeting criteria. This allowed the project to adopt appropriate job tenure and age restrictions for participants while still ensuring an adequate number of participants.

Example B

Using Outside Data to Check the Pre-Demonstration Circumstances of Participants

The AFDC Homemaker-Home Health Aide Demonstrations of the mid-1980s provided a striking example of the value of formative analyses that examine the characteristics of potential participants. In addition to their AFDC job training component, these demonstrations tested whether Medicare could save money by paying for home care for frail elderly people rather than allowing their daily functioning to decline to the point where they needed much-more-expensive nursing home care. A rigorous random assignment evaluation of the demonstrations showed (among other things) that the answer is "no"--no savings occurred (Bell et al., 1987). When viewed after the fact, it became clear that the same conclusion might have been reached much sooner, without initiating demonstrations or an evaluation. Follow-up data on the control group showed that nursing home stays are so rare in the target population ***absent*** *the home care intervention that there was no room for the treatment to substantially reduce nursing home costs. Indeed, even elimination of* ***all*** *nursing home stays in the treatment group--the most savings the intervention could possibly have produced, would not have been enough to offset the added costs of providing home care to the much larger pool of all treatment group members. It is possible that this result could been established using data on nursing home admission rates for the population of interest and a rough estimate of nursing home and home care services costs from an external source.*

Example C

Using Outside Data to Check the Likelihood of Desired Participant Behavior

The Project NetWork demonstration, operated by the U.S. Social Security Administration (SSA), included changes in how earned income affects SSA's Disability Insurance benefits (Rupp et al., 1994). Prior to the demonstration, SSA wondered if beneficiaries had enough understanding of the usual earnings rules to make this change in the rules meaningful. The demonstration went forward, however, primarily to test other return-to-work strategies (case management and rehabilitation and employment services); the changes in the earnings rule stayed in the design as a separate evaluation component.

Providing information never before collected, the project's baseline survey showed beneficiary knowledge of pre-demonstration earnings rules to be too sketchy for rule changes to have much chance of influencing beneficiary behavior. Such information, available ahead of time, might have influenced the intervention's design in important ways.

Example D

Prior Evaluations as Substitutes for New Demonstrations

In the mid-1990s, one of the nation's many state-based welfare-to-work initiatives provided wage subsidies to AFDC recipients who obtain jobs, lowering the cost to employers of hiring these workers. An evaluation of the initiative sought to determine whether subsidies increase participant employment and earnings and reduce welfare dependency. Based on several prior evaluations, the answer to this question was pretty clear: "yes." Specifically, between 1981 and 1996 rigorous multi-state evaluations looked at the impact of three similar labor market interventions for welfare recipients: the National Supported Work Demonstration (Hollister et al., 1984), the AFDC Homemaker-Home Health Aide Demonstrations (Bell and Orr, 1994), and the National JTPA Study (Orr et al., 1996). Each of those interventions provided a wage subsidy to AFDC women seeking employment, and each found that this intervention produced modest but sustained earnings gains and short-run reductions in welfare receipt. It is not clear whether the same results should simply have been assumed for the later state initiative, but a government agency facing critical evaluation needs elsewhere could reasonably have taken that tack.

5. Descriptive Evaluation: The Intervention Unfolds

Descriptive evaluations summarize what happens once a pilot or demonstration project gets underway at full scale--which offices or agencies play a role, how the project recruits participants, what services participants receive once in the project, and how long they stay in the project. Descriptive evaluations also may report on participants' status at exit, the costs of demonstration operations, and the characteristics of the communities in which the demonstration operates. Information on several of these subjects is indispensable to project oversight and other forms of evaluation. However, unlike the operational evaluations discussed in section 6 below, a descriptive evaluation does not attempt to identify the demonstration activities that are good or bad from the standpoint of the sponsor's policy goals; rather, it adopts a "just the facts, please" approach when documenting events.

5.1 Goals, Limitations, and Topical Coverage

As just noted, a descriptive evaluation documents what a demonstration project looks like--and what it does--in a straightforward, factual account. It does not judge or comment on the extent of the demonstration's success; success indicators (e.g., effective collaboration among participating agencies, achievement of recruitment targets, delivery of services on a timely basis, cost-effective use of taxpayers' money) are part of the next two types of evaluation, operational evaluations and outcome evaluations. As a precursor to these components, a descriptive evaluation attempts to remain neutral on the ***desirability*** of the events it chronicles. It does, however, present a great deal of information that others may find helpful in reaching normative conclusions--information that will be essential to the main demonstration study should the research encompass an operational evaluation component as well.

Descriptive evaluation also differs from formative evaluation in important ways. It does not seek out facts that can be known ***prior to*** or at the very beginning of a demonstration effort. Rather, it looks back at what has already happened over a sustained period of demonstration operations. This may still have operational importance; for long-running projects, descriptive information for the first 6 to 9 months of demonstration operations can lead to mid-course corrections in how the demonstration is run, similar to the use of formative information at an earlier point. But the main purpose is to provide an account of events ***after the fact***, information that can be critical in:

- Demonstrating to others that it is possible to create the type of intervention envisioned by the demonstration;
- Defining the intervention in detail as a "blueprint" for future replications;
- Establishing the size and breadth of the population served;
- Characterizing the context of the demonstration--local labor market characteristics, existence and size of other programs, urbanicity, cultural norms, etc.;
- Providing sponsors with an account of how demonstration funds were spent, relating dollars expended to services delivered; and

- Making clear how the pilot or demonstration program differs from other similar interventions that may have been in place for some time or previously tested.

This last contribution of descriptive evaluation poses a particular challenge, requiring in-depth knowledge of how conventional--as well as special demonstration--policies and programs operate in the field. While the conceptual distinction between these models may dominate discussion during the design phase, documenting true differences between pre-existing government programs and the test intervention ***as implemented*** can receive short shrift during the research phase. Yet, when done well, *ex post* policy differentiation provides an extremely worthwhile sense of how the test intervention changed the programmatic landscape for the target population of a demonstration. This point is made clear by DOL/ETA's Evaluation of the Defense Conversion Adjustment Demonstration, a study that extended its descriptive field research to 12 localities where the demonstration intervention was ***not*** implemented (Berkeley Planning Associates, 1997). Though this type of investment in comparative analysis is not common, it is in principal essential for ensuring that a demonstration in fact tested what it set out to test: something ***new***.

One should also note that not all descriptive evaluations cover all classes of demonstration events. Some focus exclusively on:

- agency planning
- field implementation
- problem solving
- community characteristics
- participant profiles
- types of services
- participant progress/attrition,
- total and per-participant expenditures,
- or a combination.

Written reports from descriptive evaluations take many forms, combining both qualitative and quantitative information. A diverse range of topics can be covered, including:

- The process of demonstration start-up;
- Expansion of program scale and service options over time;
- Longer-run service delivery procedures and partnerships;
- Number of clients from different populations or program subgroups (e.g., Hispanics, Employment Service registrants);
- Status of participants compared with milestones in the treatment process;
- Characteristics of participants reaching--or failing to reach--certain milestones; and
- Comparisons of one demonstration site to another.

Operational, or process, evaluations (see section 6 below) look at these same topics, but with a critical eye to the desirability of the demonstration's actions and accomplishments in each area given DOL's policy goals in undertaking the project.

5.2 Progress Indicators Monitored by Descriptive Studies

As just noted, descriptive evaluations fill a valuable support role in later assessments of what is good or bad about a demonstration's operations and/or outcomes. But descriptive studies also have an audience of their own, and in almost all cases attempt to release their findings to interested parties as the demonstration evolves. Sponsors--and other interested parties--often want to know at a particular point in time how far demonstration operations have progressed. Descriptive analysis meets this need by charting the intervention's progress over time and across the topics listed above. The story can change substantially from month to month, making ***repeated*** checks of progress particularly valuable.[26] To meet the need for regular summaries, researchers and/or program operators must use consistent measures of demonstration progress and gain rapid access to key data sources on a repeated basis. Strategies for achieving these goals are discussed here; analysis and reporting options for descriptive evaluations follow later in the section.

Several types of information contribute to descriptive analyses and provide input to a range of progress indicators for the intervention. Qualitative reports from the field describe a demonstration's operational planning, implementation actions, and ongoing issues and events. Complementing this, information from demonstration agencies' records can give a quantitative account of the initiative's outreach and intake volume, number of participants (and their characteristics), services provided, and dollars expended. Finally, secondary data from external sources describe the non-demonstration characteristics of local sites--local unemployment rate, percent of employment in high-growth industries, etc. Ideally, evaluators would compile--and then report on--all of these types of information for each demonstration site at multiple points during demonstration operations; not a small matter. Even less ambitious versions of this approach require careful planning to achieve their purposes.

The first step is to select the appropriate indicators of demonstration progress. Though these vary from one demonstration to another, the list of possibilities can be stated in general terms. Exhibit 5.1 does this, and indicates the most widely available data sources for each type of indicator. All of the measures on this list (1) have meaning in almost any context, (2) presage the success or failure of the test intervention as a whole, and (3) are sensitive barometers of changes of direction over fairly brief time spans (from a few weeks to several months). Each set of measures is examined in detail below. Any additional indicators that meet the same criteria and hold interest for a specific demonstration intervention should be added during the planning phase.

---

[26]This is especially true for demonstrations with long intake and program participation intervals. The full interval of demonstration operations can stretch to two years or more if multiple sites are involved and different sites begin operations at different times. For example, the demonstration period ran from 1985 to 1988 in DOL/ETA's Summer Training and Education Program (Grossman and Sipe, 1992) and from 1989 to 1992 in the New Chance Demonstration (Quint et al., 1997).

The larger challenge is accessing good data on a rapid, consistent basis. The best approach depends on the type of data one seeks among the descriptive information sources noted earlier:

- qualitative information on demonstration operations,
- quantitative data on client progress,
- fiscal information on demonstration expenditures, and/or
- secondary data on the demonstration's setting.

The discussion below considers each of these data types in turn, focusing on the best methods for accessing these data on a repeated basis over the course of a demonstration. Further details on data sources for individual indicator variables appear in Exhibit 5.1 (right-hand column).

5.3 Qualitative Information on Project Operations

For qualitative measures of demonstration operations, most pertinent data will come from those who know the program best, local demonstration operators. Some portion of this information will be committed to paper at the local or state level--in printed project summaries, demonstration grant applications/ contracts, and ongoing progress reports/correspondence with federal sponsors.[27] The rest of what is needed must be gathered in interviews with key operational personnel at the local, federal, and possibly state levels or through on-site observation. Interviews with representatives of community groups other than the demonstration agency are also essential to understand the project's integration and coordination with related organizations. Principal among these are the demonstration subcontractors and referral agencies that contribute to the intervention itself.

One's access to this type of information--particularly information gathered through discussions with State and local personnel--depends in part on who one is. For example, senior members of the federal team that launched the demonstration may have special access to well-informed individuals in the local demonstration agencies, using connections developed during the exploratory phase of demonstration planning and/or prior interactions. Strong bonds of communication may also run between local agencies and DOL regional office staff. Alternatively (and perhaps most commonly), the best channels of information flow between local demonstration staff and the federal project officers who oversee demonstration grants. These two groups often form strong working partnerships; as a result, federal liaisons are often consulted--or at least apprised--by local staff when new issues arise in the field, or when major milestones are attained. In the bargain, federal project officers become highly efficient and well-informed sources of information for descriptive analysis.

[27]Useful printed information can also come from the "case files" of individual participants compiled by project staff in the course of enrolling and serving clients. For technical and cost reasons, it is rarely possible to examine a representative sample of these files, limiting the use of case records to qualitative descriptions of the service delivery process and (anonymous) accounts of the case history of individual clients. (Leiter et al., 1997, provides a nice example of the latter.)

Exhibit 5.1
Demonstration Status Indicators and Potential Data Sources

QUALITATIVE INFORMATION ON OPERATIONAL ISSUES / EVENTS	
Indicator	**Potential Data Source(s)**
Planning	
■ Number and types of local organizations involved, key omissions.	■ Printed project summaries ■ Field interviews -- local
Formal structure of project	
■ Federal contract terms. ■ Legal and *de facto* relationships among local partners. ■ Who does what operationally.	■ Contract(s) ■ Printed project summaries ■ Field interviews -- local -- federal
Operational coordination	
■ Degree to which partners fulfill their formal roles ■ Means and extent of communication/collaboration among partners ■ Role of federal sponsor and state agencies ■ Key benefits/issues as seen by each party.	■ Field interviews -- local -- state -- federal ■ On-site observation
Start-up progress	
■ Leadership. ■ Timing. ■ Staff learning. ■ Bottlenecks (if any). ■ Federal support. ■ Reasons for lags.	■ Field interviews -- local -- federal ■ Local progress reports
Planned services	
■ Types. ■ Sources. ■ Availability (point when first offered, number of "slots"). ■ Duration. ■ Intensity. ■ Expected Outcomes. ■ Expected Cost.	■ Grant applications/contracts ■ Printed project summaries ■ Field operations manuals ■ Field interviews -- local

(continues on page 46)

Exhibit 5.1 *(continuation)*
Demonstration Status Indicators and Potential Data Sources

QUALITATIVE INFORMATION ON OPERATIONAL ISSUES / EVENTS *(continuation)*	
Indicator	**Potential Data Source(s)**
Details of operations ▪ Staffing ▪ Support facilities (office space, PCs, etc.) ▪ Service provision arrangements ▪ Client outreach ▪ Client intake ▪ Service delivery ▪ Schedule ▪ Financing	▪ Local progress reports ▪ Local/federal correspondence ▪ Field interviews -local -federal ▪ On-site observation
Demonstration "close-down" ▪ Advance Planning ▪ Individual client "hand-off" ▪ Project unit shut-down	▪ Field interviews -local ▪ Local progress reports

Exhibit 5.1 *(continuation)*
Demonstration Status Indicators and Potential Data Sources

QUANTITATIVE DATA ON PARTICIPANTS	
Indicator	**Potential Data Source(s)**
Background information ▪ Demographics ▪ Labor market history ▪ Current employment and family status	▪ Demonstration agency MISs ▪ Evaluator-designed intake forms
Outreach and intake ▪ Type of initial contact ▪ Date of application ▪ Screening date ▪ Reason ineligible/not selected ▪ Enrollment date	▪ Referral agency MISs ▪ Demonstration agency MISs ▪ Evaluator-designed intake forms
Service receipt ▪ Type of service ▪ Date started ▪ Date ended ▪ Completion status ▪ Rating/grades	▪ Demonstration agency MISs ▪ Provider records
Demonstration exit and follow-up ▪ Date of exit ▪ Status at exit/reason for termination ▪ Employment status X weeks after exit ▪ Earnings and UI benefits after exit	▪ Demonstration agency MISs ▪ Post-exit tracking surveys ▪ State UI wage & benefit records

(continues on page 48)

Exhibit 5.1 *(continuation)*
Demonstration Status Indicators and Potential Data Sources

FISCAL INDICATORS OF SPENDING	
Indicator	**Potential Data Source(s)**
Expenditures on intervention ▪ Outreach/recruiting costs ▪ Intake/enrollment costs ▪ Service costs by type of service ▪ Overhead (administrative/supervisory costs) ▪ DOL/ETA oversight costs	▪ Accounting systems of -- demonstration agencies -- major service providers -- DOL/ETA ▪ Grant invoices submitted to DOL ▪ Time use surveys
Special demonstration costs ▪ Planning and start-up costs ▪ Evaluation costs	▪ Accounting systems of -- demonstration agencies -- DOL/ETA -- outside evaluator ▪ Grant invoices submitted to DOL ▪ Invoices from outside evaluator

SECONDARY DATA ON LOCAL CIRCUMSTANCES	
Indicator	**Potential Data Source(s)**
Population characteristics ▪ Income ▪ Race ▪ Education ▪ Country of origin ▪ Health status ▪ Housing status ▪ Age ▪ Urban/rural	▪ U.S. Bureau of the Census
Labor market characteristics ▪ Unemployment rate ▪ Prevailing wage rates ▪ Industry/occupational mix ▪ Growth rate of employment by sector	▪ Bureau of Labor Statistics
Program participation ▪ Unemployment Insurance ▪ JTPA ▪ TANF ▪ Food Stamps ▪ SSI	▪ Published statistics/special data runs from program oversight agencies

On some matters, local demonstration operators may feel most comfortable sharing pertinent information with someone other than a federal employee. This is when an outside evaluator can be particularly useful--when local staff are hesitant to share all they know with federal oversight officials but would open up to outside researchers. The information gain from working this channel only increases should--as is not uncommon--a researcher already has a personal connection with someone in the local agency based on prior collaboration.

In light of these considerations, the best way to ensure that a descriptive evaluation obtains comprehensive, high-quality qualitative information involves four steps:

- Hire an outside (i.e., non-governmental) evaluator;
- Provide the evaluator with access to all printed materials pertinent to site operations;
- Put the evaluator in touch with federal staff who oversee demonstration grantees or have other channels of communication with local project managers; and
- Fund the evaluation to conduct additional field interviews where needed to complete a demonstration's qualitative research agenda.

5.4 Quantitative Data on Project Participants

The process of obtaining good quantitative information on program participants tends to be more straightforward. Many demonstration agencies have or will set up their own management information systems (MISs) for tracking participant enrollment, in-project activities, service receipt, and exit--dating and identifying each transaction (with participant identifiers). The same point applies to financial records on dollar expenditures in the demonstration (see section 5.5 below), except that sites are less likely to adopt new accounting practices or set up special financial software just for the sake of one demonstration.

Evaluators most often access data on demonstration participants through file transfers from local agency MISs, although in some instances local data have already been compiled by a state agency and can be obtained from it most easily.[28] Having tapped these data, it is sometimes possible to trace participants outside of the interval of demonstration participation to enrich project summaries. For example, some site MISs track individuals prior to enrollment, as they pass through outreach, application, and screening.[29] Others continue tracking beyond demonstration exit using special follow-

[28]The Washington State Unemployment Insurance Alternative Work Search Experiment is one example of this latter circumstance (Johnson and Klepinger, 1994). As in other Unemployment Insurance (UI) demonstration projects, the participants in this study—new UI benefit claimants—were tracked by local staff using a state-provided and compiled MIS system.

[29]Demonstrations can only track outreach at the individual level if the initial outreach steps rely on lists of potential participants (e.g., employee rosters at plants about to close) or one-by-one referrals from other agencies and non-governmental organizations.

up surveys and administrative data on other programs (e.g., Unemployment Insurance wage and benefit records). The best MIS systems also contain a rich array of background descriptors and include flexible, user-friendly reporting systems that can generate made-to-spec tables for evaluation use. The transfer of the individual-level data sets is not difficult with these systems and can occur on a routine basis (with appropriate privacy protections[30]) between local demonstration operators and the central management/evaluation team. This gives evaluators access to summary counts and breakdowns of almost any type without requiring local operators to create special (often complex and irksome) reports on their behalf.

Where such systems exist and contain all the required information[31] (or can be made to do so), national sponsors need to worry about only one thing: their right of access to all individual-level information in the system. This should always be put in the grant agreement between the federal sponsor and the local or state demonstration agency; a verbal or non-contractual written commitment by the local agency to release MIS data carries considerably more risk for the federal partner than taking the extra step of formalizing the arrangement in the agreement. If possible, the grant agreement should also stipulate the agreed-upon format, content, and schedule for information transfer. While it may not matter for certain purposes whether data are passed along as printed tables, electronic cross-tabs, or individual-level records, a sponsor should never have to settle for printed tables alone in today's computerized environment.

Federal sponsors and their evaluation teams can go even further to ensure access to individual-level program participation data by ***creating*** the data collection, compilation, and transfer tools sites need to fulfill their data-provision obligations. Centrally-provided MIS tools are recommended, where affordable, for several reasons:

- The measures collected by centrally-developed tools will fully reflect the priorities of the sponsor agency, not just those of the local program operator.

- Adequate attention will be given to not just in-program measures but also individuals' background characteristics where important to the evaluation. Exhibit 5.2 provides a list of individual background characteristics that tend to be important in descriptive and outcome analyses of employment and training initiatives and that are usually best collected by local site MISs.

---

[30]Privacy requirements depend on the nature of the data and the terms under which it was originally acquired by the demonstration operator. Data on more sensitive topics (e.g., arrests, immigration status) and data obtained from individuals under a pledge of confidentiality require the strongest protections. At most, this means stripping any MIS file of individual names and personal identifiers prior to transfer and replacing them with new demonstration identifiers that facilitate data checks between organizations and across successive releases.

[31]This was true, for example, for the state-maintained MIS systems used to track participants in the Washington State Unemployment Insurance Alternative Work Search Experiment (Johnson and Klepinger, 1994). Based on local data entry, these systems provided individual-level data on participants' entry into the demonstration project, demographic characteristics, and work search services received during the demonstration period.

- The sponsor can fully define the individual data elements that become part of the system.
- The sponsor can assure a high level of consistency in measurement across sites by providing user training on the system and building in automatic internal data checks.
- The technical quality of the system need not be left to outside parties, but instead can be fully controlled by the sponsor agency through oversight of software development and extensive home-office testing.
- Evaluators can know that all data sets will reach DOL/ETA in identical format, thus avoiding the burden of reading and processing files in different formats for each site.
- Changes in data capture and packaging cannot take place without central office knowledge, since such changes require modifications to the data-entry software and file transfer programs.
- Output file structures and identifiers can be made as convenient and powerful as possible for linking demonstration data with other administrative data sources, such as State wage records and welfare participation data.

Exhibit 5.2

Participant Background VariablesTypically Included in Sites' MIS Systems

Age

Race

Gender

County of residence

Education

Employment status

Work history

Current earnings

Unearned income, by type

Current program participation

- Unemployment Insurance
- TANF
- SSI
- food stamps
- employment & training

Given these advantages, it is a good idea to pursue central development and control of MIS software even if all sites have adequate data systems already in place--if those systems are highly divergent in content and structure. Several DOL/ETA demonstrations in the Unemployment Insurance (UI) area have taken this view, seeking to centrally develop and supply MIS tools to all demonstration sites. Special "participant tracking systems" were planned for this purpose in the entire series of UI Reemployment Demonstration Projects and reached fruition in the Lifelong Learning Demonstration.[32]

5.5 Fiscal Information on Project Expenditures

Expenditure data provide a means of measuring the resources invested in a new demonstration initiative. They come from the financial records of the federal sponsor agency and its State and/or local affiliates. Local demonstration operators ultimately expend most of the money consumed by demonstration projects, either directly (e.g., in rent and staff time) or indirectly through subcontracted service providers. But the federal sponsor--DOL/ETA--incurs some direct costs as well, mostly staff time and travel in initiating and overseeing the project.

In general, these expenditures divide into three categories relevant to descriptive evaluations:

- Planning and start-up costs that occur only once for ongoing national programs;
- Costs accrued in conducting the evaluation rather than implementing the intervention; and
- Everything else--expenditures essential to the intervention that would arise continually in an ongoing program.

When considering whether a demonstration or pilot intervention should be adopted and run continuously as a national program, one would like to separate and ignore both evaluation costs and planning/start-up costs.[33] At the same time, these costs need to be part of the equation for DOL/ETA's project management purposes, since all costs count equally from a budgetary perspective. It follows that a descriptive evaluation should track each of the three categories of expenditure data separately. Careful conceptual thinking is needed to define the first two categories (evaluation-only costs, up-front costs) and extract them from each agency's financial accounting records.

---

[32]The Lifelong Learning Demonstration was something of a special case, however. DOL/ETA-produced software was used in just one site (Baltimore) and by the evaluation grant contractor (Abt Associates), which created and administered the demonstration's treatment—promotional mailings to mature incumbent workers.

[33]Evaluation costs are clearly irrelevant to a program run year after year without a research component. In contrast, planning and start-up costs are program-generated; with or without an evaluation, they will occur in each locality if the intervention goes national—but as one-time expenditures only. As "once in a lifetime" costs, these expenditures should not play a very large a role in assessing the merits of an ongoing program. Yet they can dominate total costs in a limited-time demonstration. Ideally, one would amortize these up-front costs over the "lifetime" of an ongoing program when considering national implementation. It is far simpler—and a good first approximation—to simply exclude planning and start-up costs from the cost analysis.

Certain breakdowns may be important even within the third cost category. For example, to calculate the cost-effectiveness of a demonstration's recruiting strategy, one might want to divide the money spent on outreach and intake by the number of people who apply for admission to get a cost-per-applicant measure. Or, the efficiency of a subsidized job program might be assessed by dividing the wages paid by the demonstration by the number of persons who remain employed after the subsidy ends. To achieve these goals, an evaluator must disaggregate ongoing intervention costs into separate figures for specific functions like outreach and subsidies. Unfortunately, local demonstration agencies do not always include such distinctions in their cost accounting systems, and it is not reasonable to expect changes in long-term accounting practices and software. To compound the challenges, different local agencies almost always measure conceptually-similar costs in somewhat different ways.

In the face of these difficulties, DOL/ETA should ensure that a detailed dialog occurs between agency staff who know the expenditure data best and evaluator/oversight team members with expertise in cost analysis. This exchange with evaluation staff needs to occur at both the federal and State/local level, since both of the latter expend funds and track expenditures. In some instances, discussions of cost accounting practices may even need to include local organizations that provide services to the demonstration agency under contract--if they do an important share of local spending and it is not clear from records at the demonstration agency how their spending breaks down into the required accounting categories.

In trying to assure the completeness and comparability of expenditure data across sites and over time, it helps to start early. At the outset, the federal sponsor should make clear that (1) local agencies will have to submit regular expenditure reports and (2) the sponsor will assist agencies in this endeavor. Several steps need to follow from there:

- Development, with site input, of a common reporting format or worksheet;
- Detailed definition of each of the accounting categories shown on the worksheet;
- Careful investigation by the sponsor of what each site can do to obtain reliable measures of the desired concepts (where such measures do not already exist);
- Detailed training and monitoring of the personnel who transcribe information from agencies' financial records to the worksheet; and
- Review and discussion of completed worksheets as they come in to resolve any remaining uncertainties or difficulties in data transcription.

Each of these steps can be carried out by federal staff or by an independent evaluator. Typically, the federal sponsor will undertake some or all of these steps as part of its normal fiscal and management oversight activities. Interpretable cost information is as essential to federal managers as it is to researchers, and will often accompany invoices submitted by the grantee organization. But major gaps and classification problems can remain that are distinct to the evaluation perspective.

The substance of cost measurement is also fairly complex--and relies even more on evaluation expertise than does the data collection process. In setting a conceptual basis for measurement, one begins by identifying the categories of cost where separate dollar totals may be needed. These categories most often include:

- The ***fixed costs of project start-up*** (e.g., management planning and coordination with DOL/ETA, line staff recruitment and training, office set-up, implementation of non-evaluation accounting and reporting procedures);
- ***Client outreach and recruitment*** costs (e.g., labor costs for the demonstration unit, adverstising expenses);
- ***Client intake and enrollment*** costs (mostly labor costs);
- Costs for each ***type of service*** delivered (e.g., purchases from contractors, staff time spent on direct services, funding for income support payments);
- Ongoing ***administrative and supervisory*** costs within the demonstration unit (e.g., support staff time, office materials, PC rentals) and other parts of the host agency (e.g., agency director's time); and
- ***Evaluation*** costs (e.g., staff time for data collection and transfer, recruiting costs for an experimental "control group").

The evaluators in charge of cost data collection face a major challenge in deciding where to draw the line between different cost categories--which costs to put where when sorting out figures from various cost accounting systems. This step is the "bread and butter" of cost data collection, and an effort so detailed and idiosyncratic that a general description or discussion of the matter is impossible. Obtaining these data requires more "digging" and specialized attention to individual sites than any other part of a descriptive evaluation's quantitative research agenda.[34] Much of the required information goes beyond the dollar figures themselves to an understanding of what those figures mean--where each figure came from, what comprehensive concept it measures, and where it fits in among the many facets of cost analysis. If demonstration costs are important, sponsors should always get an expert in evaluation cost analysis to frame and oversee the effort.

5.6 Secondary Data on Local Circumstances

Finally, non-demonstration data on local circumstances come from a wide variety of sources, most of them available to the general public in automated form on CD-ROM or by download from the Internet. These sources include:

[34]For example, time-use surveys are often needed to allocate staff labor hours and dollars to various categories of demonstration activities.

- Census figures and projections of population characteristics, including family income, race, and education;

- Local labor market characteristics from the Bureau of Labor Statistics, such as the local unemployment rate, hourly wage rates, and industry/occupational mix; and

- General community data from the Area Resource File, which draws on a number of county-level sources, including the Census, to describe the demographic characteristics of each county's population.

All of these sources provide highly localized information at the county, SDA, and/or zip code level, as well as statewide and national totals for comparison purposes.

Many key indicators from these sources are routinely published by the cognizant data collection agency or program oversight body. Where not available in published form for the localities of interest, additional data analysis may need to be done by the evaluator using individual-level records in the source file. All of these secondary data sources tend to lag real events by 6 to 12 months (e.g., information on late 1999 does not become available until mid- to late 2000). In most instances, the costs of direct data access are minimal for the user, though the labor costs of learning to use each new file or each set of published reports may not be.

Given the convenience of secondary information--and its potential for contributing to high-quality evaluation in a number of areas (as discussed later in this section)--evaluators should be careful not to under-invest in this tool. Many demonstration projects acquire good information on local characteristics when selecting sites but then do not integrate that information into their analysis of demonstration operations or participant activities and outcomes. This creates an unnecessary reduction in the pay-off from this type of valuable data.

5.7 Narrative Summaries of Operational Events

Based on the qualitative information discussed earlier, descriptive results written in narrative form "tell the story" of a demonstration's progress in expositional style, providing a 1- to 10-page synopsis for each demonstration site. Using text to convey information offers almost all audiences a user-friendly format and conveys more of the richness and texture of a demonstration's operations than can summary statistics or other quantitative information. As text documents, narrative summaries come in two general forms:

- ***Descriptive accounts of where the demonstration stands at a point in time***, focusing on *who* has done *what, where, when,* and *how*? (*Why* is a normative question left to the operational and outcome evaluations.)

- ***Chronologies of implementation events and milestone achievements*** that have taken the demonstration from its inception to its current position.

The best narratives combine both of these perspectives beginning, perhaps, with a chronology followed by a more detailed account of demonstration's current status. By putting everything in one place, reports of this structure show how the sequence of events in a demonstration's evolution led up to the current situation. Moreover, when thinking about multiple rounds of reporting (see below), this structure cumulates very efficiently yet remains fairly brief and free-standing in each round. Specifically, to get from one report to the next, the evaluator need only add the most recent events to the existing chronology and adjust the point-in-time summary to reflect the changes wrought.

In many demonstration projects, narrative accounts of implementation are supplied by the organizations running the intervention in the sites. As noted earlier, these organizations are the primary source of operational information on the demonstration--but they may not be the best chroniclers of that information. Any of several possible disadvantages of self-reporting may offset the ease and economy of that route. In particular, local progress reports tend to:

- Provide "spotty" accounts of the topics of interest to sponsors while giving only the vaguest indication of how complete their coverage really is. As a result, the user cannot tell if a topic not mentioned is (1) inapplicable in that site, (2) not important at the moment, (3) underwent no interesting changes since the previous account, or (4) simply was not tracked and reported when it should have been. In the face of such uncertainty, operator-provided accounts often leave readers less satisfied than more systematic accounts.

- Have little consistency across sites, or even over time within a site.

- Be late or light on content, particularly when local staff are most busy--the very time when complete, reliable reporting is most important.

- Not be objective and even-handed in all instances. There is a natural and understandable tendency for those involved in executing a difficult new initiative to dwell more on achievements than setbacks.[35]

Given these hazards, investing in an independent narrative account by federal staff or an outside evaluator makes good sense. Though the ultimate source of information remains the same--interviews and documents from local demonstration administrators--the independent route almost always provides a more dispassionate, balanced narrative than site-generated reports. A central presence in report preparation also ensures that each report will provide basic information on a set of established topics, eliminating gaps or the need for supposition on the part of the reader. Through its completeness and even-handedness, narrative summaries filed by outside parties--preferably professional researchers--will bring DOL/ETA results of maximum worth.

[35]Indeed, one could argue that putting problems and frustrations behind you and focusing on achievements is a desired—or even necessary—approach for those charged with making a previously untried policy approach work.

5.8 Summary Tables of Participant Activities and Characteristics

Many different summary tables can be created from the three types of quantitative information discussed earlier--participant data, expenditure data, and contextual data. Such tables can greatly enrich narrative accounts of demonstration progress. This not only makes a descriptive evaluation more detailed, concrete, and defensible, but ensures that it is--at least on factors susceptible to quantitative measurement--comprehensive of the entire demonstration experience.

Statistical tables, and the text that surrounds them, usually deal with one or more of the following aspects of demonstration operations:

- Counts and characteristics of individuals encouraged to participate in the demonstration, broken down into those who never applied, those who applied but were not admitted, others who applied but did not participate (i.e., voluntary drop-outs during the intake process), and participants.
- Tallies of participants engaged in various demonstration activities, currently and cumulatively over the life of the project to date.
- Participants' employment and personal circumstances as they leave the demonstration and, if available from agency records, at a check-point in the post-demonstration period.
- Time intervals taken by participants in moving from one activity to another during the demonstration intake, participation, and exit processes--including the share who fail to reach various points in the sequence.
- Resources expended in running the demonstration, for various cost categories.
- The non-demonstration circumstances of the demonstration sites--their local labor market, demographic, and policy characteristics.

Postponing discussion of the last two bullet items (expenditures and circumstances) until later brings the remaining, participant-focused indicators on this list into focus.

Among the participant indicators, a demonstration's federal sponsor needs to determine which tables are worth generating for a specific demonstration project, depending on the policy goals of the intervention and the sponsor's areas of greatest concern regarding demonstration success. Input from evaluation professionals can be quite helpful at this point, associating specific goals or concerns with different types of summary data. Evaluator input becomes especially important when not all tables of interest can be managed, due to budget or other constraints, and priorities must be set. Research experts can play a vital role at this point by making sure the needs of later evaluation activities are not neglected, since different summary statistics play larger or smaller roles in later evaluation components. For example, cumulative tallies of participants in various demonstration activities have more to contribute to an operational evaluation than to an outcome evaluation (see sections 6 and 7 below). Client status at exit

works the other way by providing a crucial starting point for outcome evaluation. Still other descriptive tables strongly support ***both*** types of normative assessments, including measures of the background characteristics of applicants and participants. The Unemployment Insurance Self-Employment Demonstration illustrates how descriptive data on participants can be used in a variety of ways to support both operational and outcome evaluations, up to and including a rigorous examination of demonstration impacts (see Benus et al., 1994).

Some statistical tables have distinctive features of their own that researchers should take into account when presenting the data in tabular form. For example, statistics on ***participant inflows and characteristics*** for the entire participant population could mask important aspects of intake that apply to only a portion. This type of variation is most likely to occur across the different phases of demonstration intake--early, middle, and, late. For example, outreach at the start of a demonstration may bring forth the most eager applicants only and/or suffer from a general scarcity of candidates (if the outreach message takes time to reach and influence the entire target audience). Or, the initial pattern could swing the other way, with large numbers of individuals taking interest in a here-to-fore unavailable option for bettering their lives. Either way, unusual selection can substantially alter the share and characteristics of individuals applying for or accepted into the demonstration in the early months. Variation may continue during the middle portion of intake if the demonstration's limited capacity to process applicants runs up against an outpouring of interest among those slow to encounter or react to the demonstration offer. The resulting delays and oversights in application processing could lead certain types of applicants (e.g., those with the best outside options, those with the least determination) to enroll in the project in lesser numbers than at other times. Finally, the closing segment of demonstration intake could encounter the "bottom of the barrel" or "beating the bushes" phenomena, leading that participant cohort to be less able or less motivated than earlier entrants. Any of these developments would constitute an evaluation finding of some consequence; evaluators need to structure intake tables to be sure to reveal them when they occur.

Timing may also affect the ***number and types of individuals participating in various demonstration activities*** at any time, again pressing home the need for cohort-by-cohort analysis. In addition, service types may need to be gathered into broader categories in demonstrations with many service types or where certain services are very similar. Where possible, descriptive tables on services should include measures of service intensity (hours per week in a given activity) and/or duration of participation in different demonstration components, to gauge the rough "quantity" of services received in each category and to consider the ***pace*** of participant movement through the demonstration and its components. This dynamic look at the demonstration intervention could be further enhanced with tables showing the most common ***sequences*** of service receipt, along with the background characteristics of those who follow them.

Reports of the ***exit and post-exit circumstances of demonstration participants*** have to be handled with particular care at the descriptive stage of an evaluation. While reporting exit status for participants with differing background characteristics (e.g., workers on temporary versus permanent lay-off, high school dropouts versus graduates) is fine, good research practice dictates that exit results ***not*** be broken down according to the types of demonstration services participants receive. Displaying the data in this fashion immediately invites interpretation of differences in client status as measures of which service type is most effective (i.e., helpful to participants). A misguided leap of this sort inflicts perhaps the greatest damage possible on an

otherwise well-run descriptive evaluation. This simple-minded perspective ignores all the other influences on post-demonstration outcomes--including those that may matter far more than demonstration service receipt--influences like the prior experience and baseline characteristics of demonstration participants, which could easily differ between those receiving service type "A" and service type "B." Thus, variations in observed outcomes by service type may have nothing to do with the services themselves and, as a result, the cross-tab presented in a descriptive table may say nothing at all about service effectiveness.

As long as the cross-tab is made available, however, that inference will be inevitable on the part of some readers, including some of the most influential consumers of evaluation research: the largely lay audience of policy makers and program operators who ultimately will choose the service types to pursue in the future. No amount of caveats or explanations will change this instinctive interpretation. Even those fully aware of the hazards of causal interpretation may succumb to the (perhaps unconscious) inclination to remember tables of outcomes and services in just that way. This is especially likely when real information on service effectiveness is not yet available--a situation that almost always applies to descriptive evaluations. All too often this simplistic--and, at that point, only--picture of the relationship between services and outcomes becomes cemented in people's minds before the evaluator ever intended to say (or felt comfortable saying) anything about service impacts. True impact findings, when they appear later, are often ignored--the question having already been "settled" with the wrong set of findings. The only way to avoid this result is to not release service/outcome cross-tabs as part of a descriptive evaluation. And the best way to do this is probably for evaluators to simply not examine outcomes by service type at that point, leaving the question wide open. Only at the normative stages of a study, in the outcome evaluation, should the question of causality and impact be broached and the service/outcome relationship revealed for the first time.

Similarly, summaries of the ***time intervals spent in various demonstration activities*** must not be compared to participant circumstances at and after project exit, lest premature judgments be made regarding the desirability of shorter versus longer courses of treatment. No such concerns apply to summaries of the ***characteristics of participants who reach intermediate checkpoints in the intervention process*** such as enrollment, assessment for services, participation in their first service activity, or successful completion of an activity.

5.9 Financial Summaries of Project Expenditures

Most evaluations of demonstration or pilot projects include at least one--and potentially as many as three--analyses of demonstration costs:

- Budgetary analysis;
- Cost-effectiveness analysis;
- Benefit-cost analysis.

These last two analyses address normative questions--What does the demonstration produce for the money spent? Is it worth its cost?--and will be discussed in section 7 as components of an outcome evaluation. Budgetary analysis simply describes the money spent by various organizations in pursuit of demonstration objectives and represents the final element of descriptive evaluation.

Preferably, the budget analysis breaks demonstration costs into a number of functional categories. The categories differ from study to study based on the activities involved in delivering the intervention. Exhibit 5.1 above (pages 45-48) lists the generic types of costs that could be reported in this fashion.

5.10 Timing of Reports

A final aspect of descriptive evaluation concerns the timing of reports. At one extreme, a study could produce just one descriptive report, compiled after demonstration operations end in all sites. Such a report would provide complete information on demonstration events and hold great interest to readers wishing to look back on the demonstration to see what happened. And, as noted previously, the information gathered in a descriptive evaluation is absolutely essential to subsequent operational and outcome analyses.

Whatever its value in its own right, a single after-the-fact descriptive analysis cannot inform sponsors of demonstration progress while the intervention is still up and running. This suggests that a final, comprehensive descriptive study be combined with at least one "mid-stream" report on some or all of the same topics. But when, and on what topics, should a mid-stream report be prepared?

Generally, earlier is better for descriptive summaries of demonstration operations and participant characteristics. These indicators provide the first tangible evidence that a demonstration is "off the ground" and serving its intended population, perhaps the two most essential results for federal sponsors to demonstrate prior to the end of the full evaluation. Early reports are also the best way to allay fears about local agencies' ability to function in their assigned capacities or--even more importantly--to detect under-performance and gaps in service delivery that require remedial attention or a reevaluation of demonstration objectives. Ideally, an initial descriptive report would cover the first 1 to 3 months of demonstration operations and become available in month 3 or 4. This could be accomplished by relying solely on data from local agencies' participant tracking systems--except where the MIS itself is experiencing difficulty, an important finding in its own right.

Depending on the length of the intervention--and the timing of start-up across multiple sites--1 to 3 additional descriptive reports could be useful. For short-term interventions of 6 to 9 months, there is little to be gained from another analysis prior to closure, since events will be too far along once a second report arrives (in, say, month 7) to address trends emerging since the first report. If sites begin operations on a staggered basis, however, it is essential that more than one descriptive report be prepared in the early-going, to examine ***each*** site's first 2 or 3 months of operations on a quick-turnaround basis. Separate reports by site often work best in this situation by keeping the focus on start-up problems and reducing reporting lags (by staggering the evaluator's workload). If summaries can be generated at will from automated table production software in the site's, sponsor's, or evaluator's computer system, summary statistics could even be compiled and examined once a month.

Longer field periods create the opportunity for more than one midstream report per site, but not necessarily the need. If a front-end report indicates substantial problems, or if the sponsor has special concerns about a particular demonstration agency, "check-ups" at 3- to 6-month intervals would make some sense. Qualitative feedback from a site could also influence the timing of reports, if new problems are hinted or rapid progress is made operationally to address an old problem. On the other hand, sites

with smoothly running demonstrations probably do not need to be re-examined more often than once a year (if the intervention lasts that long).

More complete descriptive reports that include operational narratives, fiscal information, and/or community descriptors require substantially more time and resources to prepare than automated computer summaries. At most, this investment should occur only annually during the course of demonstration operations. For most demonstrations, this means 1 or 2 (or 0) interim reports, plus a final descriptive report a few months after the demonstration ends. These forms of analysis serve mainly to document and state a project's overall accomplishments; in principle they could wait until the final report even in a long-term demonstration. Usually, the desire for interim reports among constituents of a study and the advantages of examining and writing about developments while the information is "fresh" lead to fairly comprehensive annual descriptive reports during the operations period. Annual accounts of demonstration evolution hold perhaps the most interest when an intervention progresses through several distinct stages over time--as would be the case, for example, in a demonstration intervention consisting of job search assistance followed (if no job is found) by additional skill training and subsidized employment.

6. Operational Evaluation: Lessons on Program Execution

This section begins an examination of the questions that typically motivate tests of new policy ideas in a demonstration format:

- Will this policy or program work?
- Whom will it help, and how much?
- Should we do more of it?

These are normative questions asking, ultimately, whether the demonstration intervention is worthwhile. Demonstration and pilot evaluations can include two types of normative analysis:

- Operational evaluation (also known as process evaluation or implementation analysis); and
- Outcome analysis.

These two research types share common, normative roots and, hence, are discussed jointly at the beginning of this section. The role and methods of operational evaluations complete the section, with a similar discussion of outcome evaluations appearing in section 7 below.

6.1 Normative Analysis of Demonstration Operations: Goals and Topics

Webster's Dictionary defines "evaluation" as "the act of ascertaining the value of [something]." When used in studying government programs or policies, "value" is a bottom-line concept concerned with what a program or policy intervention is truly worth. The formative and descriptive evaluations discussed in the two

preceding sections were not true "evaluations" according to this definition, since they eschewed any consideration of worth to focus instead on factual events. In contrast, operational and outcome evaluations hinge on the question of worth.

When one attempts to put a value on something--in this case, the accomplishments of a DOL/ETA pilot or demonstration project--one must first adopt normative standards that can say which consequences of a demonstration are good and which are bad, and how good or bad. This "good/bad" yardstick can be applied to either an examination of demonstration operations or an examination of demonstration outcomes, or both. In general, normative policy research has three goals:

1. Decide whether a government intervention is working out as planned, both in its program operations and its interactions with participants;

2. Judge whether the intervention has produced enough social value--i.e., benefitted enough citizens--to warrant its continuation or, in the case of a pilot or demonstration, its extension to other localities.

3. Find ways to improve the intervention in future applications.

Goal 1 involves studying the intervention itself and fits naturally within the concept of an operational evaluation" Goal 2, on the other hand, concerns aspects of a demonstration project that can only be achieved outside the program itself, in society at large. Since external events that flow from an intervention are often referred to as the intervention's "outcomes," the term outcome evaluation seems appropriate for this type of research. Both types of normative assessment can contribute to normative Goal 3, identifying ways to ***improve*** the intervention in future applications.

Operational evaluations describe and interpret what happens as government agencies (and, in some instances, their non-government partners or contractors) create and administer a demonstration project's intervention. These studies are often called "process analyses" or "implementation studies" in recognition of their emphasis on the ***process*** of demonstration activities and the ***implementation*** of the program design to be tested. The term "operational evaluation" encompasses everything that happens when a test program is created, not simply the processes followed or the steps required for initial implementation.

As a normative exercise, operational evaluations consider the ***effectiveness*** of the efforts expended in local sites to make the demonstration happen. They examine the organizations involved, the structure of their operating relationships, and the steps taken to create and administer the test intervention. The questions answered concern the achievements and desirability of those actions, not simply their character. As a result, the goals of an operational evaluation can be far ranging or focused in particular areas. Potential objectives include:

- Understanding the exact nature of the "treatment"delivered by the demonstration;

- Documenting how that treatment was produced;

- Identifying problems encountered in the process of generating the treatment and saying how they were resolved;

- Determining whether the intervention as implemented matched the intervention as planned--were goals such as effective collaboration among participating agencies, achievement of recruitment targets, and delivery of services on a timely basis achieved?

- Ascertaining the strengths and weaknesses of the treatment and of the inputs used to generate the treatment (organizational structure, operating procedures, community involvement, funding, etc.); and

- Developing hypotheses about the intervention's likely effects on participants, and reasons why certain outcome goals of the intervention may or may not be met based on operational factors.

All stages of demonstration development and operations could be included:

- ✔ Initial planning, staffing, and organizational team-building;
- ✔ Client targeting and outreach procedures;
- ✔ Early implementation and program adjustments;
- ✔ Interaction with other organizations, as referral sources and service delivery agents;
- ✔ Intake procedures, client flow, attrition during intake;
- ✔ Demonstration staff activities;
- ✔ Service providers (roles, experience, incentives, accountability, etc.);
- ✔ Service delivery process;
- ✔ Service receipt (patterns, duration, etc.);
- ✔ Participant exit procedures and services;
- ✔ Close-out at the end of demonstration operations; and
- ✔ Management oversight and control.

Inputs to this assessment come from several of the same sources mentioned in the discussion of descriptive evaluation in section 5. Already having examined the origins and content of these data, attention here focuses on the ***techniques used to analyze and report normative results*** on a demonstration project's operations.

6.2 Ways of Inferring What's Good or Bad about Demonstration Operations

Operational evaluations seek to determine what is good--and what is bad--about the way demonstration interventions are organized and run. If a descriptive evaluation has been done, gathering data and laying out the basic facts about the demonstration intervention, an operational evaluation can move immediately to a normative analysis of the facts and data. If there is no descriptive study, the operational evaluation must first gather the data and present the facts, including all of the types of data discussed in section 5 except cost data. Many pilot and demonstration studies combine descriptive and normative investigations of project operations in a single "process" or "implementation" study, though they are conceptually distinct. In these instances, the factual investigation comes first, as described in section 5, followed by a more judgmental assessment of project operations that looks for good and bad components.

What tools does one use to decide if a particular feature of a demonstration's operations (e.g., expedited client screening, contracting-out job placement services) helps or hinders progress toward a policy goal? Mostly, one looks for ***suggestive indications*** of successful and unsuccessful practices, not black and white findings of good or bad. Though it may seem unsatisfying, this framework has at least two advantages:

* Almost always, data on demonstration operations are insufficient to ***prove*** that a demonstration would better meet its goals were this practice modified, or that procedure reassigned to a different agency, or so on[36]; and

* Running and refining new policy interventions is a highly nuanced business, where trial and error, "feel," and aspiring for nothing more than change in the right direction guide most operational refinements.

In light of these limitations, methods for inferring possible operational hindrances and helps in achieving a demonstration's policy goals include:

- ***Examining demonstration descriptions and planning documents*** to find possible strengths and weaknesses in the intended intervention, based on logical coherence, operational feasibility, potential implementation obstacles, and past experience with similar policies and procedures.

[36]There is at least one clear means of proving that a particular operational practice or arrangement adds to (or detracts from) demonstration success, though the necessary data are rarely available. This entails comparing a project's impact on participants among sites that use different operating procedures or follow different service emphases. Such an analysis must first control for differences in participant characteristics and local economic conditions among sites and then trace remaining differences in impacts to specific operational features. Such an analysis requires an extraordinarily large amount of participant data, a credible impact analysis approach, and highly disciplined use of the many operational indicators that might "explain" variations in impacts across sites. The rigorous scientific methods needed to confirm causal relationships at this level may yield insignificant findings in many situations, whether the analysis covers many sites with large participant samples or not. Bloom et al.(1993) provides an example that met all the prerequisites but still yielded inconclusive results in the National JTPA Study.

- ***Talking to the designers of the test intervention*** to sort out the mechanisms by which they expect the treatment to influence participants and gauging the realism of this hypothesized "pathway of change."

- ***Doing "spot checks" of case files*** to ensure that the progression of treatment activities accords with the intended approach for most participants.

- ***Considering*** whether demonstration policies and operations are founded on ***assumptions*** about local conditions (labor markets, program/policy history, demographics, community settings) that simply do not apply to one or more demonstration site(s).

- ***Hearing from federal oversight officials*** at DOL/ETA regarding (a) the largest challenges faced--both centrally and locally--in creating and operating the intervention, (b) the elements of the intervention most likely to contribute to program success, (c) the greatest mis-steps or struggles occurring during administration of the intervention, (d) resolution of major problems, (e) operational accomplishments of local agencies, (f) reasons for incomplete success, and (g) suggestions for next time.

- ***Hearing from State and local staff--and demonstration operators***--regarding the same set of factors.

- ***Observing demonstration intake and service delivery*** while visiting study sites to see if the execution of the intervention hits rough spots or break downs in practice in ways that might limit its effectiveness.

- Reaching out to ***other sources in the community*** for independent views of the demonstration's accomplishments and organizational/operational weaknesses, effectiveness in interacting with other agencies and community groups, and the importance of local factors in all that occurred.

- ***Learning*** which elements of the intervention--and which of its operational practices--were, ***from the perspective of participants***, most helpful. Which were most difficult to deal with? Most wasteful or frustrating?

- ***Using agencies' tracking system data on individual participants*** to look for intake glitches and bottlenecks, shortfalls against recruitment targets, intermediate events on the hypothesized "pathways of change" that are not taking place, delayed program exit compared to the norm, and (if available) reasons for exit that suggest problems with client management procedures or individual service types.

- ***Comparisons across sites*** to reveal variations in the above factors that may signal outstanding operational achievements or unusual difficulties in certain sites compared with the demonstration-wide norm.

As noted earlier, several of these methods revisit the factual information from the descriptive evaluation, with an eye to the operational practices that seem most promising or threatening to demonstration success. Other techniques on the list overlap components of formative evaluation (see section 4): reviewing written documents, critiquing project staff expectations, and questioning the assumptions underlying the intervention's design. In these areas, operational evaluation simply updates a formative study's pre-demonstration findings by taking into account the final intervention design and impressions formed by the evaluator in doing field work and analyzing demonstration participation data.

Other analyses on the list require data collection or research methods not yet discussed:

- Identifying normative issues through on-site observation and case file review;
- Convening focus groups or administering small surveys to gain information on participant experiences and opinions;
- Expanding the use of participation data to check whether the demonstration's recruiting targets were met, compare the intervention's observed path of influence with staff expectations, find glitches and bottlenecks in service delivery, and (if possible) examine reasons for project exit; and
- Using cross-site variation in demonstration procedures and achievements/difficulties as clues to what worked and what did not work when applying the intervention.

Each of these areas will be examined in turn before closing the section with suggestions on the timing of operational reports.

6.3 On-Site Observation and Case File Review

Seeing the intervention in the field can be enormously educational as a "reality test" of assumptions or assertions about demonstration operations gathered from other sources, or in sparking new hypotheses about the role of various procedures that can be tested elsewhere. At least some observation of the intervention "in the flesh" should be part of every operational evaluation. However, this goal faces three major challenges in many demonstration settings:

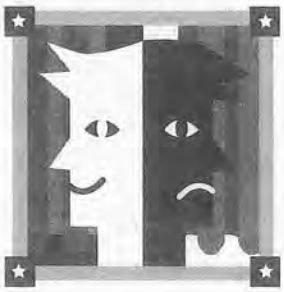
1. The "treatment" may not be visible, or only partially visible. This is true, for example, for treatments that offer of a re-employment bonus or pay wage subsidies through employers.
2. Time and sample sizes for on-site observations can be sharply limited and geographically constrained, ruling out the collection of statistically representative data. Thus, site-visit-based results cannot be interpreted as applying to the universe of all demonstration sites. This often presents the potential for over-generalizing or over interpreting the available data as though it spoke to all demonstration operations. It is vital that operational evaluators resist this temptation.
3. The mere presence of an observer may alter staff / participant interactions and the dynamics of service delivery.

Review of written case files can solve the first problem, if the demonstration sponsor stipulates that even invisible--but important--transactions between demonstration staff and participants (e.g., telephone conversations) be recorded there. A "paper trail" of this sort can be especially informative in testing a new intervention, when both the nature of the treatment and its demands on staff time are highly uncertain. If the burden of documentation seems too large in a longer-running demonstration, specification of certain time periods and/or participants for intensive case file tracking can enhance everyone's understanding of the treatment and its limitations. Ideally, evaluation staff would revisit case files for a fixed set of individuals in every round of site visits, to view "fresh" information each time and have the opportunity to question demonstration staff on ambiguous or missing information before time erases all memory. Consistent record-keeping by demonstration staff--and a consistent data extraction protocol for evaluation field staff--are even more important.

Little can be done to offset the office-based, sporadic sampling of participants that characterizes this type of data collection. However, some sort of systematic sample design may be possible, either by observing participants during specified and uniform time intervals in each demonstration office, or by stating the type or mix of participants needed and asking office staff to identify a sufficient number of case folders and/or office visits of that sort. Usually, observing services provided by outside vendors will involve added travel and expense but is justified when a demonstration relies on outside providers for many of its services or all of its referrals.

In most instances, observing service delivery unnoticed by participants or office staff is not an option, for ethical and practical reasons. Thus, any direct exposure of researchers to the "live" intervention will be noted by the key actors in the process; hence, it runs the risk of "observer effects" impinging on the phenomena under study. Several steps can be taken to reduce this risk. First, demonstration staff can make participants aware of the observer as early, and in as neutral a way, as possible, to eliminate any confusion and diffuse suspicions. Second, no more than one observer should witness a given treatment event, since the more "foreign" individuals one introduces into the service delivery setting the greater the risk of procedural or behavioral distortions. Finally, observers should never speak to anyone, staff or participants, while participants are present, or convey any physical reaction to what they are seeing. Experienced field researchers have the greatest chance of upholding these guidelines and should fill the role wherever possible. Advance training of observers is essential regardless of experience since the challenges of remaining unobtrusive can be highly context-dependent. As with case file reviews, consistency of data across observers and sites dictates the development of a uniform, pre-set protocol for recording all on-site observations.

6.4 Participant Focus Groups and Opinion Surveys



Another way to look at what is working and what is not in a demonstration is to consider the participant's point-of-view. In a results-oriented, consumer-focused era of social policy reform, the views of clients are an essential ingredient in any assessment of program effectiveness. Potential clients who decide ***not*** to participate in the project may also have lessons to impart about what is missing. Thus, in judging the success of demonstration operations, one would like to consider the following participant (and non-participant) viewpoints:

- Sources of initial information about the demonstration;

- Reasons for participating (or not participating);
- Life situations/attitudes that conflict with assumptions about participants built into an intervention's design;
- Participants' understanding of the intervention's provisions and special incentives;
- Parts of the intervention that functioned smoothly for participants, and parts that did not;
- Personal goals while in the demonstration;
- Aspects of the intervention that helped achieve those goals / aspects that stood in the way;
- Retrospective views of whether participation was worthwhile; and
- Suggested changes for future programs.

Participant viewpoints may also suggest hypotheses to be examined in the participant outcome evaluation (see section 7 below).

In general, researchers use two techniques to gather participant views: focus groups and "customer satisfaction" surveys. ***Focus groups*** are structured discussions between research staff and a small number of demonstration participants (5 to 10, typically) in a face-to-face setting. One researcher serves as facilitator, posing questions for group discussion. Participants then react to the questions, stating their own views and/or reacting to the views of others. By observing (and, with permission, tape recording) the conversation, other members of the research team are able to pull out the major themes of the discussion on the various topics covered--and to gauge the consistency of views among participants. Ideally, only a few questions (3 or 4) are introduced over the course of a session so that each topic can be explored in depth (with all participants contributing) and participants manage to stay clear about--and focused on--the topic at hand. The Evaluation of Job Corps' Pilot Project to Include 22- to 24-Year-Olds, for example, used focus groups to gather feedback from demonstration participants on their Job Corps experiences (see Executive Resources Associates, 1987).

To broaden the basis of conclusions, focus group studies should include two or three sessions—with separate sets of participants—on any given set of topics if at all possible. In multi-site demonstrations, two groups per site is a good number, and at least one group in each site is essential. For consistency, the same protocol should be followed in all sessions on a given topic, maintaining the same questions in the same order.

A final, complex aspect of focus group research concerns the selection of discussion participants. The small number of individuals involved, and the voluntary nature of participation, rule out statistical sampling on a representative basis. Instead, focus group samples must be chosen judgmentally. Since the goal of focus groups is to surface any major client concerns about the demonstration intervention, the best focus groups include a wide range of participant backgrounds and service receipt patterns. Thus, as a group, focus groups should include a range of locations (sites), time periods (early versus late in the demonstration period), and participant types (in terms of work history, age, education, etc.), as well as different types of program experience (in terms of enrollment duration, service patterns, and status at exit). Such a mix is best attained if researchers specify the bases--number and types of participants--they want

to cover in each session and ask local demonstration staff to identify an appropriate set of potential participants. The list needs to include at least twice the desired number of participants since many of those invited will be unable or unwilling to attend. The best focus group sessions last about an hour and are held at a safe, central location outside the demonstration's offices.[37] Depending on the population served, transportation assistance, on-site child care, and evening sessions may be needed to achieve good attendance. Payment for participation is another possibility.

Small-scale ***customer satisfaction surveys*** can also help evaluators assess a demonstration's effectiveness from the participant point-of-view, as can including questions on satisfaction with demonstration services in a broader survey of participant outcomes (see section 7, below). This approach requires good contact data on participants and telephone interviews with a broad, statistically representative sample of participants. It sacrifices depth for breadth compared to focus groups, garnering a much wider assessment of "customer" viewpoints on similar but often somewhat different topics. A strong example of the use of customer satisfaction questions in employment and training surveys appears in the evaluation of DOL/ETA's Unemployment Insurance (UI) Self-Employment Demonstration noted earlier.[38]

The depth/breadth trade-off is a difficult one to weigh. By soliciting views from many more participants,[39] surveys give researchers a sense of how widespread a client issue is before surfacing it in an evaluation report. But surveys may miss other issues--or potentially important predicted outcomes--that might emerge from focus group discussions. Surveys also do less than focus groups to articulate the exact nature of an issue and its perceived source(s). In other words, focus groups are more eloquent and surveys more balanced. Most evaluation sponsors prefer the former (if they have to choose between the two), though either option is defensible in almost all cases. The most important aspects of using either tool are to sharply define what each data collection mode can do and not do, and then present and interpret results in that light. The temptation to over-interpret both types of data is considerable; hence, the most important guidelines for their use concern what ***not*** to do:

[37]To be fully informative, focus group participants must feel free to express concerns about the way the intervention is being run. This level of candor is much harder to achieve in a demonstration office than at a neutral location.

[38]Benus et al. (1994).

[39]Sample sizes for customer surveys usually fall between 200 and 500 respondents. For broad statistical measures, such as the frequency of various sources of initial demonstration information, moving from a sample of 200 to a sample of 500 reduces the margin of error in reported results by about 35 percent (e.g., from $\pm$ 3 percentage points to around $\pm$ 2 percentage points). Further increases in sample size yield more gradual reductions in margins of error (e.g., moving from 500 respondents to 800 reduces margins of error by only 20 percent) and are probably not worth their cost. A better way to use added survey funds would be to add additional waves of interviewing (e.g., a very early wave to feed results back into up-front program refinements) or add topic areas and/or questions to the coverage of the interview.

Never interpret focus group results as a general portrait of participants as a whole, since the individuals selected for focus group outreach--and especially those who show up at a meeting--can depart substantially from the full participant pool.[40]

Never imagine that a summary of survey responses gives the whole story for any client issue, or that it measures the basic presence of an issue with complete consistency and reliability.[41]

Each of these limitations can be addressed by the alternate, complementary technique. Thus, a demonstration evaluation that emphasizes customer satisfaction as a measure of success should include ***both*** types of client research: survey data on issues identified through focus groups discussions,[42] and complementary insights of focus group participants to expand on survey-based results. A useful "bridge" between the two different measurement techniques can be achieved by having focus group participants fill out a very short questionnaire at the start of each session which includes one or two key "customer satisfaction" measures from the survey.

6.5 Operational Issues Revealed by Participation Data

An operational evaluation can draw on the same individual-level data on program participation used in section 5 for descriptive evaluation. However, to use these data to explore normative questions about demonstration operations, evaluators must introduce new data elements and interpretive frameworks. Most of these extensions are straightforward, requiring only that researchers:

- Compare ***total enrollment*** in each site to that site's enrollment target, to see which sites ***exceeded plan*** and which ***fell short***. All other things equal sites with greater enrollments relative to their goals are more successful sites, at least at the beginning step of the service delivery process. Shortfalls in enrollment should be traced if possible to factors measured elsewhere in the descriptive and operational evaluations: slow demonstration start-up, shortages of intake staff, an unusually small local target population or an unusually strong economy, etc. Where targets still seem reasonable, the analysis should highlight operational factors that reduced participation and draw lessons for future implementation.

[40]Selected cases tend to over-represent extremes in the distribution of any characteristic since they are deliberately chosen to provide breadth of background and experience. Within this group, those who show up may be (1) more organized or public-spirited than non-participants, (2) less pressed for time (e.g., the jobless, those without children), or (3) simply less satisfied with their demonstration experience and looking for a way to voice their displeasure.

[41]As is well known, measurement error and survey non-response can skew the results of any survey.

[42]Both survey topic selection and question wording can benefit from this synergism, addressing some of the liabilities of surveys generally.

- Look for ***glitches and bottlenecks in service delivery***, once participants are enrolled in the demonstration. Changes over time in the duration of project enrollment for successive cohorts of participants or in the aggregate mix of services at a site could signal a problem of this sort. Where these indicators are present, the next step is to identify any operational shortcomings that may have contributed to the slow-down, such as demonstration staff shortages, contractual issues with vendors, or the need for special participant certifications (e.g., a drivers license, physical exam, or skill competency test results) before initiating a specific activity. External changes and funding issues also need to be considered as possible explanations, such as when the sole supplier of a particular service leaves town, or a non-demonstration program used as a source of no-cost services restricts this option due to budget cuts.

- Where available, use MIS data on ***reasons for demonstration exit***--and ***status at exit***--to surface and illuminate operational issues of policy importance. How a demonstration or pilot project removes cases from its rolls sometimes speaks volumes about how it functions generally. For example, exits many months after the last service activity ended suggest that the end stage of treatment lacked focus. Or exits to return to school or to relocate, if plentiful, may indicate that the demonstration did not provide much of the kind of assistance participants thought they needed. Alternatively, large amounts of missing data on reasons for exit or status at exit would suggest that the demonstration too often lost touch with its participants later in the service sequence. None of these scenarios--nor any of a number of others that might arise-- necessarily implies negative results for participants, but each does signal a potential problem with demonstration operations that may suppress participant outcomes over the long run.

Using participation data to ***track the "pathways of change" by which a demonstration intervention influences outcomes*** involves more complex analysis. In this task, an evaluator first needs to record the series of events by which the intervention's designers intend demonstration policies to influence participant outcomes, as described above. Then, conceptual milestones along this "pathway of change" need to be matched to concrete measures of client progress from participation data files. For example, receipt of skill certification in a new trade could be the first milestone along a treatment "path" that starts by retraining dislocated workers in high-growth occupations and then follows up with job placement in high-growth industries. If this demonstration's participant tracking system shows more clients undertaking job search than in skill training as their first activity and records very few new skill certifications, the intended "pathway of change" has been disrupted almost before it started. One might then question whether the intervention can work as intended to achieve its ultimate goal for participants--reemployment at high wages. Researchers can then seek operational explanations for this break in the flow as part of their normative assessment of demonstration operations. It may be, for example, that demonstration procedures allow for immediate testing without training for participants with strong backgrounds, followed by up-front job search, but that staff too often took this short-cut for less job-ready clients and even neglected to establish skill certification in many cases. Findings of this sort would not only add to one's understanding of operational limitations, they would also provide a powerful tool for interpreting long-term participant results and demonstration impacts in the outcome evaluation (see section 7 below).

6.6 Cross-Site Comparisons

To this point, all of the techniques mentioned for judging a demonstration's or pilot's operations apply to individual demonstration sites. In multiple-site projects, evaluators can gain further insight into demonstration functioning by considering where one site stands in relation to another on operational factors. At a minimum, operational outcomes in outside sites provide a benchmark for assessing results in any one particular site. For example, consider a site where the average participant waits three weeks following enrollment before starting his/her first demonstration service. Is this lag long or short, good or bad? A tough question. But if one sees that each of the other five sites in the demonstration has an average lag of under two weeks, three weeks qualifies as long and, therefore, indicates a potential operational problem in that site. This suspicion can perhaps be confirmed using additional cross-site comparisons to see, for example, if participant attrition between enrollment and the start of services is also higher in the site with uncharacteristically long lags.

One can also use cross-site comparisons to see if one site differs from others across a ***range*** of operational indicators. If so, the unusual site can justifiably be called an "exemplary site"--or, if it differs consistently from other sites in a negative direction, a "challenged site." At which point, the evaluator should re-examine ***all*** findings for the site in light of this distinction to see if there are other normative implications. For example, such a reappraisal could reveal that a site with a fairly typical job placement rate at exit did not handle exits particularly well given its unusually favorable results in other areas of operations (e.g., certifications attained). Without a framework of what is typical in other sites, the areas of relative weakness in a given site could never be identified.

In making these assessments, ***analysts must be alert to what cross-site comparisons cannot do***. While such comparisons can highlight the sites that experienced the best results in various areas of demonstration operations, they cannot reveal whether one local demonstration ***agency*** did a better job than another, since each agency operated in a different environment and dealt with a different mix of clients. For example, an agency retraining displaced workers in Seattle would be expected to show better operational results than its counterpart in Appalachia, even if its operational procedures and staffing were no stronger. Many other factors of this sort intervene in determining operational outcomes. For example, in Seattle good public transportation might make participants more regular in visiting the demonstration office, more participants might have telephones in their homes to follow-up on job interviews, more displaced workers might be highly-educated, and a strong economy might increase exits to employment--all accelerating re-employment independently of how the demonstration is run. For these reasons, and because of other local factors:

> *Evaluators and sponsors should never interpret a comparison of operational indicators across states as a "report card" on agency performance, if unadjusted for other factors that could influence state-by-state results.*

Such a table may merely identify the agencies with the most favorable operational outcomes, not necessarily the agencies that did the best job with what they had to work from. Given the likelihood of this mis-impression on the part of research "consumers," the safest course is to simply *not present* site-level operational results side by side in tabular format, instead confining any comparative statements across sites to the text where appropriate caveats can be attached.

6.7 Timing of Reports

The timing of reports from an operational evaluation parallels that of a descriptive evaluation, since both draw on the same data sources and factual information. Generally, it will take a bit longer to supplement descriptive summaries of demonstration operations with normative assessments of their functioning. As a result, an operational assessment will tend to lag the corresponding descriptive report by 1 to 3 months. The basic cycle should remain the same, however, (see the end of section 5 above) and produce:

- An analysis of participation patterns, using local MIS data, within 6 months of demonstration start-up.[43]
- (for longer-running interventions) Full-scale reports every 12 months during the operational period.
- A final, comprehensive report roughly 6 months after demonstration operations end.

7. Outcome Evaluation: Lessons on Participant Results

Outcome evaluations "tell the rest of the story" for demonstration participants: what happens to the people or organizations served by the demonstration once demonstration services end. For people served by employment and training demonstrations, the post-demonstration outcomes of interest include employment status, earnings, income, dependence on income payments from the government (e.g., TANF, Unemployment Insurance benefits), job stability, job quality, and work supports (transportation, child care, etc.) for one or more years following demonstration exit. For pilots and demonstrations that act on organizations (e.g., local offices administering Unemployment Insurance benefits), the outcomes of interest might include such things as weekly flow rates, lags in processing, error rates, and customer satisfaction.

Outcome evaluation seeks to identify the good and bad aspects of participant outcomes following demonstration exit. Like operational evaluation, the goals here are normative:

- Determine how much the demonstration intervention helped participants, and
- Decide whether the intervention's costs are more than offset by its contributions.

This section describes the goals and techniques used in analyzing participant outcomes in demonstration and pilot evaluations. It includes ways of tracking outcomes, measuring impacts, and comparing benefits and costs, and concludes with an assessment of required sample sizes for statistically reliable findings.

[43]Involves multiple, staggered reports if start-up is staggered across sites.

7.1 Types, Goals, and Topical Coverage

Outcome evaluations track participant results in one or more of three dimensions:

- ✔ Outcome ***levels*** following demonstration exit;
- ✔ Outcome ***changes*** over time, beginning at demonstration entry and continuing past exit; and
- ✔ Outcome increments, or ***impacts,*** attributable to the demonstration.

The first two dimensions--outcome levels and outcome changes over time--are straightforward conceptually. Examples can perhaps best define their character. Findings on outcome levels take the form "61 percent of demonstration participants held jobs six months after demonstration exit." A typical result for outcome changes might be "Error rates in local Unemployment Insurance offices dropped 18 percent between the start of the demonstration and a month following its end."

The concept of outcome increments, or impacts, is more subtle. An outcome "increment" is defined as that part of a participant's outcome ***that can be attributed to the demonstration,*** and would not have occurred otherwise. Thus,

> *A demonstration impact (sometimes called a demonstration effect) represents the amount contributed to outcomes by the demonstration itself, on top of what the participant would have managed on her/his/its own.*

This amount, added at the margins, is the full measure of what the demonstration contributed and, therefore, a core indicator of its value. However, outcome levels and changes do not convey this value. Outcome ***levels*** are influenced by many factors besides the demonstration intervention, including participants' background characteristics and starting points (e.g., age, prior work experience) and local labor market conditions (e.g., the local unemployment rate). It is impossible to say from data on outcome levels alone which part of an overall outcome stems from "pre-set" factors of this sort and which comes from the demonstration intervention. ***Changes*** in outcomes come closer to measuring the influence of the intervention alone, though they also reflect the natural evolution of participant outcomes and local economic conditions over time which would have taken place even without the demonstration.

Put in a nutshell, a demonstration's impact--when properly measured--represents the ***contribution*** of the intervention to better or worse outcomes ***on top of everything else***. As will become evident, this shift in perspectives creates major challenges for an evaluation. It is no longer sufficient to know that outcomes for demonstration participants are good, or that they are improving. One must also know ***why*** outcomes are good or improving if one seeks to place credit where it is due--with the demonstration or with other factors.

In its detail, outcome evaluation can encompass a wide range of goals:[44]

- Describe the status of demonstration participants after they exit the demonstration, in terms of employment, earnings, dependence on government payments, and other labor market outcomes.
- Note the subgroups of participants (e.g., high school graduates) that fare better than others.
- Show the trend over time in participants' labor market outcomes, contrasting the pre-demonstration, in-demonstration, and post-demonstration periods.
- Quantify improvements in outcomes between two specific points in time, one prior to demonstration entry and the other following exit.
- Note the subgroups of participants that improve their situations most over that period.
- Estimate the intervention's impact on participant outcomes--the extent to which measured outcomes or outcome changes stem from the demonstration intervention itself, as opposed to other factors.
- Note the subgroups of participants for which impacts are largest.
- Test hypotheses about demonstration impacts taken from the operational evaluation.
- Gauge the outcome achievements of the demonstration relative to its goals.
- Compare the beneficial impacts of the intervention to its costs to see how much taxpayer money was spent to achieve the demonstration results.
- Decide if the demonstration intervention produced enough social value--i.e., benefited enough citizens to a sufficient degree--to warrant its continuation and/or extension to other localities.
- Provide a factual basis for advocating--or opposing--the demonstration intervention as national policy.
- Save taxpayer money by rejecting interventions that are not cost-effective.

To answer these questions, an outcome evaluation often needs to examine demonstration outcomes of many different types, and from a variety of perspectives. Whether measuring outcome levels, changes, or impacts, some of the most common topics of interest in DOL/ETA pilots and demonstrations concern:

[44]The discussion in this section applies to demonstrations that act on individuals; parallel goals and topics could be stipulated for demonstrations that act on organizations (e.g., demonstrations that attempt to integrate local DOL/ETA programs), though these are less common.

- ***The size of the "complete treatment"***--the combination of demonstration services and any similar services participants receive from outside the demonstration.

- ***Labor market behavior*** of participants following exit--labor force status, cumulative weeks in labor force, employment status, cumulative weeks/hours of work, job duration, job characteristics (e.g., hourly wage, fringe benefits).

- ***Career upheaval*** since job loss (if that triggers demonstration entry)--length of initial joblessness, reduced job quality (hours, hourly wages, benefits), occupational shifts, industry shifts, relocation, changes in immigration status.

- ***Family income and expenditures*** since exit--participant earnings, other earnings, receipt of government transfer payments (UI benefits, TAA, food stamps, TANF, etc.), other income, child care costs, child support payments.

- ***Social integration and progress*** of participants following exit--welfare participation, living arrangement, marital status, recent childbearing/fathering, housing/neighborhood quality, community involvement, educational attainment, criminal activities.

- ***Adult and child well-being in participant families*** following exit--health status, health insurance coverage, personal confidence/outlook, child care issues, school progress, child behavior problems.

- ***Costs of the intervention*** per participant--to DOL/ETA budgets, to other government agencies, to all governments (taxpayers), to participants (e.g., out-of-pocket work expenses).

- ***Cost effectiveness*** of the intervention in relation to other policy options--the government dollars spent per unit of participant/social gain compared to the next-best policy option.

- ***Net gains/losses***, in dollars--for DOL/ETA, all governments (taxpayers), participants, participant families/communities, society as a whole.

- ***Timing of gains and losses***--does government pay now, but gain later? are participants helped initially, but not in the long run?

- ***Winners and losers***--across different government programs, types of participants, and localities.

In sum, anything that happens to project participants as a result of the intervention--or to taxpayers (who fund the project) or other citizens (who experience "spillover" effects)--belongs in a comprehensive assessment of a demonstration's outcomes.

7.2 Collecting Outcome Data from Administrative Records

Monitoring participants beyond demonstration exit requires new data not included in other evaluation components: *long-term follow-up data on individual demonstration participants.* These data begin where participant tracking information from local demonstration agencies leave off: when a participant leaves the project and sets out alone in the labor market. Participant status at this point (i.e., at exit) tells the earliest

part of this story if available from site MISs. Key measures here include whether a participant is working and, if so, in what type of job. Except for the possibility of sites adding short-term follow-up data beyond exit (e.g., employment status 13 weeks later) to their records, the rest of the story must come from other sources.

Almost all outcome evaluations collect one or both of two basic types of follow-up information beyond exit:

- Program participation and earnings data ***from State and local administrative records***, including benefit receipt and quarterly earnings amounts from State Unemployment Insurance agencies, and welfare participation data from State or local TANF offices.[45]

- Non-economic outcome measures--and more detailed employment information--***from personal interviews with participants.***

While both types of data have their limitations (see below), they tend to be the only options available to track participants' situations once they break off contact with the demonstration.[46]

Several principles apply to the collection and use of ***administrative data*** in outcome evaluations:

- ***Collect individual-level data for as long a time interval as possible***, to maximize the length of follow-up and provide the best possible perspective on the past. In particular, evaluators need ensure that the outcome variables seen as key in the post-demonstration period (e.g., quarterly earnings) are tracked for the in-demonstration period and some of the pre-demonstration period as well. In-demonstration data play a vital role in benefit-cost analyses, while pre-demonstration data help define the participant population and strengthen many types of impact analysis (see below).

- ***Include in data requests all individuals known to have a connection with the demonstration, no matter how tenuous***, for anyone whose Social Security number (SSN) is known. This means everyone who receives demonstration outreach because they are on a list of eligible individuals (e.g., on the list of people filing for Unemployment Insurance benefits over a specified time

[45]Occasionally, demonstration interventions emphasize progress in other, non-labor market domains and, thus, can be better evaluated if other types of administrative data are collected. For example, the Lifelong Learning Demonstration promoted continuing education for incumbent workers and gathered key outcome information from the enrollment records of post-secondary schools and training institutes (Bell et al., 1996).

[46]Studies of pilot and demonstration projects that act on organizations rather than individuals—e.g., a demonstration of technical assistance to high schools building internship programs—may also collect data on an organization's clients from the same sources mentioned here, as measures of customer service and customer outcomes organization-wide. Other follow-up data sources tend to dominate these studies, however: financial and operational reports from the organizations themselves, quality-control data for the programs they administer, and private sector institutional measures (from Dunn & Bradstreet etc.).

interval), those who actually apply to the demonstration (from agency records), those who participate, and those in the eligible pool who are not even recruited. Each of these groups can play an important role in some part of the outcome analysis (see below), and increasing the number of cases in an administrative data request has almost no impact on total data collection costs.[47]

- ***Collect data in overlapping waves*** to (1) avoid missing any older data that have been removed from States' source files, (2) protect against gaps in data at the "seams" between waves, (3) provide ample time to check the data and obtain replacement files if an extract is faulty, and (4) provide an added check of data completeness.

- In submitting data requests, ***consider the reliability of the SSNs through which data matches will be made*** between files at either end of the exercise. Data quality matters equally in the demonstration database that supplies the requested SSNs and in the State system(s) where resident SSNs provide access to the right individual-level earnings and welfare data. An error at either end will result in a data mis-match--appending and analyzing earnings or welfare information on the ***wrong*** person--or no match where there should have been one. To the extent possible, evaluators need to learn more about the origins of both sets of SSNs and devise ways to test their reliability.[48]

- ***Understand the limitations of analysis variables collected for administrative purposes***--e.g., the types of jobs not included in Unemployment Insurance wage records, the potential for several monthly welfare payments to appear in a single monthly entry, and lags between an event (e.g., receipt of UI benefits) and its appearance in the source data set.

- In preparing data for analysis, ***focus on determining why an earnings or welfare record was not received*** in a particular quarter or month for a given individual. Was it because the person had

---

[47]Most of the costs of data requests and transfers, for both the evaluator and the source agency, are invariant with the number of individuals requested (i.e., the number of SSNs submitted). Transferring information for just a few hundred demonstration participants in a site means paying the full cost of negotiating a data-sharing agreement, submitting a file of SSNs, drawing the extracts, and—at the evaluator end—learning how to interpret individual variables, checking data reliability, and developing a system that converts raw variables into an analysis file. The added computer processing costs of expanding the list of SSNs involved are trivial in relation to these fixed costs.

[48]One way to test the reliability of SSN matches is to compare the values of other variables that appear on both of the source files. Date of birth works best in this role, since it is highly distinct among individuals (though occasionally identical), unchanging over time, and—in numeric form—easily matched by computer. The most obvious candidate, the individual's name, lacks these last two properties but may be worth using when date of birth is not available.

no earnings or welfare benefits that period, or because of incomplete data?[49] One cannot meaningfully analyze outcomes from administrative data until one knows that non-reporting means non-receipt (at least in the great majority of cases).

Other ***operational challenges*** must be surmounted to successfully analyze administrative data, including collaborative work between the demonstration sponsor and the evaluator to:

- ***Obtain access to administrative data at the individual level*** using SSNs from both files to match individual cases. Many of the same confidentiality issues that arose in section 5 with regard to demonstration agencies' MIS data on participants apply here as well. As there, successful arrangements with administrative data suppliers often require that the research team certify that none of the individual identifiers or substantive variables provided will be released or reported individually. Fortunately, this provision does not run counter to an outcome evaluation's research agenda. DOL/ETA's oversight role for certain State and local programs (e.g., the Unemployment Insurance system) may help to ensure access.

- ***Obtain accurate SSNs*** for all individuals with administrative data needs, for use in record matching and extraction. In most instances, SSNs come from demonstration application forms and program records of local demonstration agencies. If possible, the research team should emphasize the importance of capturing this variable accurately and, if necessary, train local staff on collecting SSNs and other identifying variables (e.g., name, date of birth).

- ***Negotiate a data transfer schedule and format*** with technical staff at the source agency. This schedule must take account of data management cycles at the source agency--how soon old records are moved off the master file to archives, and how long it takes for current records to become complete. Finding technical staff with the right expertise and focus on the demonstration can be difficult and very often needs to be reenforced over the course of data collection (especially if personnel change) by workingwithdata agency managers.

- If essential to an agreement, ***be prepared to reimburse State and local agencies*** for any costs they incur in filling evaluation data requests. This can require substantial resources (up to $50,000 for any one agency) that must be included in the evaluation budget.

- ***Identify someone at the source agency who can help define and interpret the substantive variables*** on the file. Misunderstood data, or serious frustrations when documenting variables, can jeopardize both the schedule and reliability of outcome analyses.

---

[49]For both earnings and benefit payments, key outcome findings center on (1) receipt or non-receipt of a particular type of income and (2) the average amount received over a series of quarters or months. Quarters/months of receipt, and dollar amounts above $0, stand out in administrative data, since they are the very outcomes State data systems are designed to record. Non-receipt and $0 amounts, on the other hand, have to be ***inferred*** from the ***absence*** of a record in the extract. This inference can only be made by ruling out missing data as an explanation for the absence of a record.

- ***Test data transfer procedures in advance***, checking the quality and completeness of the "test" files and--if necessary--repeating the exercise until all data transfer systems are working smoothly. Especially where data are archived (or even deleted from source data systems) on a set schedule, flawed data transfer mechanisms risk serious losses of information when used, untested, for "real time" requests.

7.3 Collecting Outcome Data from Participant Surveys

Personal interview data cover many more, and more complex, outcome measures than administrative data and, thus, raise a number of special research and operational issues. Unlike administrative data, interview data on demonstration participants are extremely unlikely to exist independent of the evaluation and so must be generated from scratch as part of the study.[50] The collection of research data using personal interviews is a research discipline in its own right and cannot be covered adequately here. At present, suffice it to say that the interview data used to evaluate labor market interventions are as expensive as they are essential, costing $100 to $400 per completed interview for each of as many as 20,000 respondents. Personal surveys must also be planned far ahead, and may take 12 to 24 months to carry out (including time to prepare data for analysis).

This combination of importance and cost makes surveys the one area of demonstration research where funds must be spent both copiously and with the greatest possible assurance of success. As a result, setting up and executing surveys demands great care and long lead times. Despite these precautions, evaluation surveys occasionally fail due to either absolute limits in data collection technology or flaws in their design and execution. Should this happen, an outcome evaluation will almost surely fall short of its originals goals.

To attain the highest level of certainty possible in funding a survey, a demonstration's sponsor must give careful attention to the two most ***avoidable*** sources of survey difficulties:

- Selection of a data collection organization that is inadequately equipped for the assignment, in expertise and/or resources; and

- Allotting inadequate time for survey planning and design.

The first of these threats, inadequate data collection capabilities, arises most often when sponsors commit too little funding to survey data collection in relation to what they want to see accomplished. In this situation, inexperienced survey organizations are both the most likely to offer to conduct the (under funded) survey and the least able to adapt when serious resource constraints arise. To avoid this dilemma and field a reliable survey instead, demonstration sponsors need to recognize the iron law of evaluation research:

[50]There may be times when JTPA's standard 13-week follow-up survey could play a role in an outcome evaluation of a JTPA pilot or demonstration project.

If a participant survey cannot be adequately funded, it should be eliminated or reduced in scope, and the consequent shrinkage in evaluation scope accepted as appropriate to the circumstances.

The formula for avoiding inadequate design time is not so straightforward. A few things are clear, however:

- ***A study should never attempt to design and field a participant survey in less than 12 months.*** The evaluation team must complete a number of large, complex tasks to move from no preparation to complete readiness, and some of those tasks can be neither compressed nor overlapped.[51] There is high potential for an irreversible error if one tries to wedge the whole process into 8 or 10 months.

- ***Sponsors should engage a separate evaluation team to prepare the survey instrument and OMB clearance package if, for some reason, the main evaluator cannot be brought on board in time.*** This move, too rarely made but often enormously profitable, provides both ample lead time for survey preparation and "two sets of eyes" for refining the questionnaire and sample design. It can be most valuable when planning a survey of demonstration participants as they enter the demonstration--a point when many demonstration initiatives do not have their long-run evaluation capabilities in place.

- ***The study team, including sponsor staff, should work closely with the Executive Office of Management and Budget (OMB)--and far in advance of any formal request for data collection clearance.*** This means putting OMB staff in touch with the study's design and survey components as early as possible, well before submission of an official clearance package.

The rest of the survey discussion focuses on design elements that maximize the utility of survey data assuming on-time, high quality data collection. Three fundamental questions arise in this regard:

❶ What kinds of information do outcome evaluations need from surveys?

❷ What sample designs work best for survey-based analyses of outcomes and impacts?

❸ When should surveys be conducted relative to demonstration entry and exit?

7.4 Selection of Survey Outcomes and Interview Questions

No single discussion could fully examine, or even broadly illustrate, the many ways outcome measures and corresponding interview questions are put together to conduct surveys of demonstration participants. There are as many variants as there are outcome evaluations, and in each case a thousand reasons arise for building the specific survey questionnaire selected. It is possible, however, to note the range of possible survey outcomes of interest to DOL/ETA evaluations and to suggest some criteria for picking survey outcome measures and question wordings.

[51] Major survey preparation steps include questionnaire design, sample design, review and approval by the Executive Office of Management and Budget (OMB), interviewer hiring and training, and a field test of the instrument with possible further adjustments.

Exhibit 7.1 lists many of the ***outcomes of potential interest*** in a DOL/ETA demonstration or pilot evaluation. Not all of these measures will apply to every program innovation tested, though many are likely to come up frequently. For many evaluations, only a very long personal interview could cover all relevant topic areas in sufficient detail--an interview exceeding the upper-limit of respondent tolerance of about 40 to 50 minutes. Thus, most demonstration evaluations must prioritize among the outcomes of interest to a participant survey. How should decisions on interview content be made in particular instances? ***In general, one should only consider a survey outcome measure that is unobtainable from other, less expensive sources such as administrative records, and which has at least one of the following characteristics:***

- ✓ It corresponds to one of the demonstration's policy goals (e.g., earnings);
- ✓ It might be affected by the intervention (e.g., family composition);
- ✓ It has important implications for participants' lives (e.g., job stability); ***or***
- ✓ It has important implications for DOL/ETA (e.g., frequency of use of employment and training services).

Clearly, sponsors need to compare outcomes to their goals for the demonstration. Less obvious, but just as important, are participant outcomes ***not*** related to the project's goals but which may be affected by the demonstration in secondary--and perhaps unanticipated--ways. Finally, since participant outcomes matter in their own right, even outcomes with little likelihood of changing due to the intervention need to be part of an outcome evaluation if they are important to either participants or sponsors.

Crafting the right survey question or questions to measure each outcome requires highly specialized knowledge. The process of questionnaire design should not be attempted without expert input from survey specialists in the topic area of interest. Things about survey questions that only experts will know, including precedent and past performance, play a large role in the questionnaire design process. Thus, when designing outcome questionnaires, one should begin with the way(s) previous surveys have asked about the concept of interest--particularly past evaluations of labor market interventions (both demonstration studies and ongoing program evaluations). The wording and format of previously-used questions need to be adopted whenever precedent shows that:

- Respondents understand and respond to the question without difficulty;
- Few respondents say "I don't know" to the question or refuse to answer;
- The format of responses (e.g., unit of measurement, time interval covered) is well-suited to outcome analysis, ***and***
- The content of responses provides reliable measures of the intended concept.

If no existing questions meet all four of these criteria, the demonstration research team must develop new ones. ***It is absolutely essential that any new question development for participant interviews be guided by survey questionnaire specialists with expertise in the topic area*** and, if possible, by the recommendations of multiple experts. Thorough pre-testing--and consequent question refinement--are also essential in this scenario, with so many of an evaluation's resources and outcome findings riding on the result.

Exhibit 7.1

Outcomes Measures from Participant Follow-Up Surveys: Common Examples

Labor Market Outcomes

Job search activities
Labor force participation
Employment status
Earnings
Number of jobs held over X months
Weeks worked in X months
Average job duration
Usual hours worked per week
Hourly wage
Fringe benefits
(e.g., health insurance, paid vacation, pension plan)
Industry and occupation
Union status
Use of work supports
(e.g., transportation assistance, on-site child care, job coach)

Participation in Income Support Programs
(if not available from administrative records)

Unemployment Insurance benefit receipt
Unemployment Insurance benefit amount
Food stamp receipt
Food stamp amount
TANF receipt
TANF amount
Supplemental Security Income (SSI) receipt
Supplemental Security Income (SSI) amount
Medicaid eligibility
Live in public housing
Using a rent subsidy or housing voucher

(continues...)

Exhibit 7.1 *(continuation)*

Outcomes Measures from Participant Follow-Up Surveys: Common Examples

Other Economic Indicators

Earnings of other family members
Child support income
Total family income
Poverty status
Health insurance coverage
Child care costs (out of pocket)

Education and Training Achievement/Activities

Highest degree/level of education
Education and training activities
Education and training assistance received
(from demonstration and non-demonstration sources)

Non-Economic Status/Well-Being

Marital status
Number and age of children
Living arrangement
Self-perceived housing quality
Self-perceived neighborhood quality
Most recent residential move (date and state of origin)
Number of residential moves in X months
Health status
Personal confidence/outlook
Community engagement
Problems with child behavior/school progress
Criminal arrests, convictions, and detainment/incarceration
U.S. citizenship/immigration status

7.5 Survey Sample Design

In evaluation applications, survey sample designs can be quite idiosyncratic: each must be "custom fit" to the particular intervention and evaluation approach used. Developing a basic sample design for interviewing demonstration project participants goes most smoothly if sponsors and/or evaluators work through a process of elimination, asking themselves the following questions:

Are there participants who do not belong in the outcome evaluation (at least not in the evaluation of outcomes measured in the survey)?

For participants who do belong in the evaluation, should an interview be attempted with every individual or just with a subset?

If interviewing is restricted to a subset, should survey respondents be selected at random, or on some other basis?

The possibility of including non-participants in the survey sample design will be considered later in the section.

Depending on a study's overall sample size (see discussion at the end of the section), ***survey analyses may benefit from omission of the earliest and/or latest cohorts of participants***. These individuals may have atypical outcome experiences, either because they have greater or lesser skills and interest than the typical participant or because they receive an atypical intervention during the start-up or wind-down phases of a demonstration. Fortunately, the three earlier types of evaluation--formative, descriptive, and operational--look for evidence of start-up and wind-down effects of this sort. Because follow-up interviews typically do not begin until at least 12 months after demonstration entry (see below), these lessons often are available in time to guide survey sample selection. Depending on exactly what is learned about the special qualities of the earliest and latest cohorts, either or both of these groups might be dropped from the interview sample as uncharacteristic of the larger demonstration.[52]

---

[52]This assumes that the first and last cohorts are of no particular interest in their own right. Sometimes, these groups do hold particular interest. For example, the earliest cohort of participants in the Lifelong Learning Demonstration (the pilot-within-a-pilot sample mentioned in section 4 above) was the only group that received encouragement to return to school *just ahead of the spring semester* rather than just prior to the *fall* semester (Bell et al., 1996). This seasonal distinction held enough interest that follow-up data were collected for all cohorts.

Often, personal interviews are attempted for only a subset of the participants of interest. Weak contact data may account for some omissions, but more often reductions occur for cost reasons: many evaluations simply cannot afford, or cannot justify, the high level of spending required to interview all relevant participants. Interviewing a random subset of relevant cases works just as well as interviewing them all when conducting outcome and impact analyses--except that the margin of error in measuring survey-based outcomes increases with subsampling. The end result is survey findings that are somewhat less conclusive than they might have been had more cases been interviewed. At least in the near future--as in the recent past--virtually all demonstration evaluations will confront this trade-off to some degree, increasing uncertainty to reduce costs. The tougher question concerns the ***extent*** to which this trade-off should occur, a matter also addressed later in the section.

Finally, ***there are times when an outcome evaluation should focus more survey resources on one particular subset of participants than is likely to occur when picking respondents at random.*** This situation arises when a particular subset or type of participant holds greater policy interest than is indicated by its relative share of the overall participant pool. This can happen, for example, when only a subset of participants seek demonstration assistance as happened in the Pennsylvania Reemployment Bonus Demonstration (Corson et al., 1992). In such situations, it makes sense to oversample the most vital segment(s) of the pool--here, the individuals who applied for the remployment bonus. Oversampling can also be based on participant background characteristics such as years in the labor force, industry affiliation, family composition, or educational level.

Evaluators accomplish oversampling by attaching a different probability of selection to different types of participants, setting probabilities for priority groups above average and for everyone else below average. They then select survey cases at random but use the new, unbalanced probabilities to do so[53]--increasing the chances of interviewing someone in the favored group ahead of someone in the participant pool at large. This allows for more confident statements about the favored group than would be possible otherwise, but reduces certainty with regard to other subgroups and--to some extent--the participant population as a whole. Adjustments of this sort must be made with great care and under the direction of sampling statisticians or others with sampling expertise.

---

[53]Conventional random selection implicitly attaches an equal probability to every participant, consistent with its goal that each participant subgroup should constitute the same share of survey cases as it does of the overall participant pool.

In addition to the ideal sample of participants, there are reasons to consider ***interviewing non-participants*** when conducting an outcome evaluation. Potential candidates for interviewing include people recruited by the demonstration who do not apply to the demonstration, those who apply but do not qualify, those who qualify but do not participate, and individuals who resemble demonstration participants but who have no connection to the demonstration. Depending on the intervention, these groups can provide:

- Useful reference points for understanding participant outcomes;
- Follow-up information on individuals who choose not to participate in the demonstration, including differences between various types of non-participants (non-applicants, excluded applicants, etc.); and
- Benchmark measures to be used in calculating demonstration impacts.

All of these uses appear later in the section. Regarding survey sample design, the critical role of non-participants in the measurement of demonstration impacts suggests that further examination wait until the impact measurement discussion below.

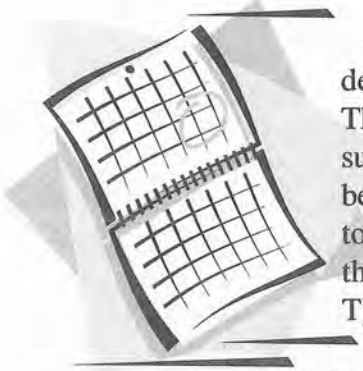
7.6 Interview Timing

In an outcome evaluation, time begins at demonstration entry. With staggered intake of participants within a site, and--in most multi-site demonstrations--the staggering of start-up across sites, the date of entry will vary considerably in calendar time. In "demonstation time," however, all participants have a common reading of "0" on the day of demonstration entry and stay in alignment, relative to entry, each succeeding day (day 1, day 2, . . .). Most outcome analyses work off this counter, not calendar time. Thus, when reporting on participants in the demonstration and post-demonstration periods, each individual must be at the same place as all others ***relative to his/her date of demonstration entry***--say, 60 days past entry, or 9 months past entry. It is worth noting that this time scale, initialized at entry, also plays a larger role than a third "clock" one could consider for demonstration research, a clock that starts ticking the day of demonstration ***exit***. Three factors lead evaluators to favor time-since-entry over time-since-exit in analyzing demonstration outcomes:

- Participants' experiences while in the demonstration can, at some level, be thought of as "outcomes" in their own right. In this view, in-program events are the "outcome" of gaining admission to the demonstration program and like other outcomes evolve over time relative to the day of entry.
- All participants occupy a common point in life at the moment of demonstration enrollment, from both the program and personal standpoints. At that moment, they all (1) are interested in receiving the demonstration's treatment, (2) have met a uniform set of project eligibility standards, and (3) understand the intervention and its options in roughly the same way. None of these conditions holds at demonstration exit, when (1) some participants may want to continue

their treatment beyond the scheduled end-point while others "drop out" early, (2) situations with regard to initial eligibility factors (e.g., receiving Unemployment Insurance benefits) may have changed in different directions, and (3) each participant has just gone through a different type or length of treatment.

- For the impact and benefit-cost analyses, what happens during the demonstration--net gains (or losses) in educational attainment, earnings, hours worked, etc. relative to what would have happened absent the intervention--is every bit as important as what happens after. Any additional hours spent in school or at work still mean sacrificing time at home, any earnings increases still produce more material well-being, and so forth, whether these changes occur early due to the time-use adjustments imposed by the intervention or late because of the treatment's carry-over effects into the labor market.



Usually, demonstration time is kept in months, not days, with the month of demonstration entry designated as Month 0, followed by Month 1, Month 2, etc. This is almost always the way one thinks about the timing of participant follow-up surveys, asking "How many months beyond demonstration entry should participants be interviewed?" Depending on the length of follow-up needed for the intervention to have its full effects, a survey design framework may also need to allow for more than one interview per participant, one in Month M and another in Month T (M < T); even three rounds of interviews are sometimes included.

In picking the month of interview relative to demonstration entry, one also chooses the ***length of time covered by analyses of the survey data***. For example, a survey scheduled to occur 12 months after demonstration entry most often provides information regarding Month 12 and--for outcomes measured continuously over time--any preceding month beginning in Month 0 and continuing through Month 11. Confining time coverage to just this interval and no more follows from several considerations. First, an interview conducted in Month 12 simply cannot collect information about events that have not yet occurred, in Month 13 and beyond. Also, as a matter of choice, outcome evaluation surveys generally do not ask about months prior to Month 0. This is because the demonstration intake process in Month 0 provides the better opportunity to collect data on Months -1, -2, etc., ***preceding*** demonstration intake, most often through scaled-back self-administrated questionnaires added to the demonstration application process (see section 5 above). More extensive interviews about the pre-demonstration period rarely make sense for employment policy test initiatives unless essential in measuring the impacts of the demonstration on participants (see below).

So ***how does one choose the appropriate "follow-up" interval for participant follow-up surveys, and when are multiple rounds of interviews needed?*** Any decision to use short follow-up intervals and/or several rounds of interviewing needs a powerful justification, given the costs of survey data collection. One guiding principle is to:

Limit participant follow-up interviewing to as few rounds as possible, subject to other analytic and data quality considerations.

Of course, the "other considerations" are crucial and often difficult. They include:

- The need to ensure accurate reporting by respondents by using only short recall periods (e.g., events over the last 12 months rather than the last 24);
- The risk of losing track of participants while waiting for the next interview date;
- The urge to produce findings on participant outcomes while key questions about the intervention are most in the public eye; and
- The importance of long-term findings in evaluations of long-term labor market strategies. In many employment and training evaluations, the intervention takes some time to evolve--6 to 12 months--and full responses to the intervention only emerge over 1 to 5 years past that point.

The first three of these considerations favor frequent interviewing, generally at intervals no longer than 12 months--and, ideally, every 6 months or so. In national longitudinal surveys, the 6-months-or-less standard is achieved regularly only by large-scale, general-purpose surveys such as the Survey of Income and Program Participation (SIPP). Evaluations of individual pilot and demonstration interventions for limited populations cannot possibly justify data investments at this level. However, participant interviews every 12 months beginning in Month 12 after demonstration entry are perfectly reasonable and have been used in many demonstration evaluations. Even so, most demonstration studies have had to settle for less frequent interviewing and longer recall periods; for example, the Evaluation of the New Chance Demonstration interviewed participants 18 and 42 months after entry, while the Summer Training and Education Program (STEP) Evaluation targeted months 18, 30, and 54.[54]

Fortunately, experience suggests that participants in DOL/ETA programs and demonstrations provide reliable information on employment-related events even when asked about them 12 to 18 months later, if asked in the right way.[55] The recall abilities of this population 20 or 24 or 30 months after the fact have not been tested systematically, though limited experience suggests some extension past 18

[54]See Quint et al. (1997) and Grossman and Sipe (1992).

[55]Specifically, surveys should gather information on jobs, unemployment spells, and schooling using a continuous retrospective calendar, asking first if the respondent is currently engaged in any of these activities and working back from there. Approximate start and stop dates are recorded for each spell, for as many spells as have occurred during the reference period, starting from the most recent and working backwards through time. Experience has shown that this form of questionning—applied to an 18-month reference period—detects short-term or casual jobs that might be overlooked otherwise and produces spell data that match up fairly well at the "seams" between consecutive rounds of data collection. (For example, what a respondent says about her/his Month 18 activities when interviewed a second time in Month 36 accords reasonably well with what s/he said in an 18-month interview.) See Kornfeld and Bloom (1997), who use data from the National JTPA Study that follow this structure.

months should be possible.[56] Longer intervals between interviews need not create a respondent tracking problem, as long as the data collection organization continues to contact sample members regularly in the interim--say, every 6 months--to update its records of addresses, telephone numbers, and alternate contact persons.

Other than to interview earlier and incur the costs of added survey rounds, there is little that can be done to address policy makers' frequent desire for survey-based results sooner rather than later. Making the most of interim results from administrative data--which can be collected over and over at little cost--can alleviate some of this pressure, especially for employment and training interventions where evaluators can track many of the most important outcomes (employment, earnings, Unemployment Insurance benefits, welfare receipt) using administrative sources. Where this is not enough, more money simply will have to be spent if earlier--and therefore more frequent--reporting of survey-based outcomes is essential.

Issuing survey results early--based on a partial set of interviews or by skipping careful data checking and analysis variable construction--is ***not*** a good idea. Obviously, "unclean," corrupted data can produce misleading findings which, if released, can result in highly visible and embarassing retractions--or even loss of faith in the reliability of the evaluation in general. The final cases interviewed are also important: without them, the survey cannot fully represent the population of interest, nor will it include enough respondents to hold sampling error to an acceptable level. These added cases are part of the overall sample precisely because findings can be wrong without them; it follows, then, that findings can ***change*** when the last interviews are added.

The final factor on the list--the need for long-term follow-up--puts even more pressure on survey costs. Whatever interview cycle is adopted a, say, four-year study of survey-based outcomes requires twice as many rounds of interviewing as the evaluation of short-term strategies where the whole story can be told in two years. Fortunately, for labor market interventions it is rarely essential to track survey-based outcomes over the long term even when demonstration effects could last many years, for two reasons:

- The availability of low-cost follow-up data on employment, earnings, and welfare participation, from administrative sources, can often answer key questions about the long-term benefits of a demonstration intervention. This was the case for the National Supported Work Demonstration and the AFDC Homemaker-Home Health Aide Demonstrations, for example.[57]

- Where that is not the case, the long-term path of survey-measured outcomes can often be projected from three--or even just two--rounds of follow-up interviews. Suppose, for example, that interviews are conducted 12 months after demonstration entry--by which point all in-demonstration activities have been completed--and 24 months after entry. For continuously measured indicators such as hours worked or TANF receipt, the month-by-month trajectory of outcome levels between Month 12 and Month 24 can be extrapolated into future months using mathematical models. This technique has often been proposed for benefit-cost analyses that depend on survey-measured outcomes.[58]

---

[56]For example, the National JTPA study conducted initial follow-up interviews as many as 30 months past program enrollment, and collected 60 months of retrospective employment data at baseline for some sample members.

[57]See Couch (1992) and Bell et al. (1995).

[58]An application of this method appears in Bell and Orr (1994).

No single "bottom line" emerges from this constellation of considerations, nor any simple formula for determining the interval between survey interviews. The best advice is to be flexible, rather than locked into past customs, when setting follow-up intervals for participant surveys. Here are some ***customs that demonstration planners should look to break***, if doing so will provide a more informative and affordable survey outcome analysis:

Inflexible Custom 1: Use only multiples of six when scheduling participant follow-up interviews(Month 6, Month 12, Month 18, etc.).

There is nothing magical about intervals of six, or even of three (Month 3, Month 6, Month 9, Month 12, etc.). While earnings records from administrative data cover three months at a time (a single calendar quarter), other follow-up data need not line up in the same way.[59] A shift away from Months 6, 12, etc. in scheduling follow-up interviews for a demonstration can sometimes both shorten recall periods and generate findings sooner, often at no real cost to the evaluation. This "win-win"situation arises, for example, when three survey waves are moved from Months 12, 24, and 30 to Months 10, 20, and 30.

Inflexible Custom 2: In any survey wave, all sample members must be interviewed in the same follow-up month (e.g., Month 18 exactly).

This dictum is probably the most costly of any current survey custom. For the majority of demonstrations that enroll participants over an extended period of time, conducting interviews on a particular anniverary of enrollment for all sample members (e.g., Month 12) spreads out survey data collection over many months. For example, a demonstration that enrolls participants over 14 months must run its follow-up survey 14 months as well to catch every participant exactly X months after demonstration entry. This inflexibility raises survey costs considerably relative to a more condensed interview period[60]--and, in the end, may not even produce the uniformity of interview dates desired.[61]

---

[59]Indeed, survey follow-up data ***will not*** line up with quarterly earnings data even if all interviews occur at six- or three-month intervals following demonstration entry. Rarely will demonstration entry occur rignt at the start or end of a calendar quarter (i.e., on or near January 1, April 1, July 1, or October 1 of a given year). Hence, 3X months of interview data will rarely align in time with precisely X ***full calendar quarters*** of earnings information.

[60]The fixed monthly costs of operating a survey, which are non-trivial, are paid many more times over an extended interview period than is necessary in a compressed interview schedule.

[61]Every survey interview occurs somewhat later than the date it enters the interviewing queue for the first time, since interviewers have to locate the respondent, catch them at home at a convenient time, and gain their cooperation. Each of these steps can be a time-consuming exercise and will vary sharply in length across respondents. Thus, even if all interviews are ***first attempted*** on a uniform anniversary of demonstration entry, they will be ***completed*** at many different points relative to entry—any time between 1 day and 3 months after that anniversary.

Far better to wait until ***all*** participants are nearing or past the desired follow-up month and do all interviews then, in the span of just a few months. For questions that pertain to the day of the interview (e.g., current employment status, current marital status), a bit of variation in the reference date ***relative to demonstration entry*** will do little harm--and the original, uniform reference date (say, 20 months following demonstration entry) was somewhat arbitrary in any case. For outcomes measured continuously over time based on spell data (e.g., hours worked each week), researchers can simply ignore the "extra" months of data collected for the earliest demonstration entrants (those interviewed well past the intended anniversary date). As an added bonus, a compressed survey of this design will provide a "preview" of longer-term participant results for the subset of respondents that has progressed furthest past demonstration entry when the survey begins. And because the survey still ***ends*** at the point originally planned (X months after project entry for the last cohort of participants), these "bonus findings" are available at the same time as all findings would have been obtained under the conventional plan.

Inflexible Custom 3: Never go more than 12 months between interviews.

As noted earlier, waiting 18 months or more to collect self-reported labor market information will not seriously jeopardize data quality if the evaluator updates contact information on respondents in the interim and asks retrospective quesitons in the proper, proven manner.

7.7 Assessing Client Status and Progress following Demonstration Exit

Survey and administrative data on participants provide a rich base of information for reporting the status and progress of demonstration clients following exit. These components of outcome evaluation focus on participant outcome ***levels*** and ***changes over time*** as indicators of demonstration success and use those indicators to draw normative conclusions about the test intervention. While not the most powerful tools for this purpose, these elements do provide useful information without complex analysis. (A stronger tool for making normative judgments about a demonstration's participant outcomes--the demonstration's ***impacts*** on participants--is discussed subsequently.)

Analysis of outcome levels involves little more than straightforward compilation of participant characteristics and circumstances in the post-demonstration period. The kinds of descriptive tables introduced in section 5 for summarizing participant characteristics before demonstration entry serve equally well here to summarize participant characteristics after demonstration exit. ***To look at changes*** in participant circumstances, before versus after the demonstration, one simply combines these two tables into one, with separate columns for the pre- and post-demonstration periods. An essential requirement is that a given characteristic (e.g., number of children) or status indicator (e.g., labor market status) be measured in the same way in both time periods, drawing on data from the same data source or one closely comparable. Without this assurance, differences in measurement technique may obscure or compound true changes in outcomes.

The final step adds a third column to the table indicating the numeric difference between the "before" and "after" figures. This sharpens the cross-time comparison and provides a direct estimate of change. Standard statistical tests can then determine which of the measured changes are real and not just artifacts of random variation in the data.[62] Changes that have both statistical significance and substantive significance (i.e., are big enough to matter) become indicators of key trends in participant circumstances following demonstration entry,

changes perhaps caused by the demonstration or perhaps not. Each of these analysis steps can be repeated for various subgroups of participants (e.g., high school drop-outs, women, Hispanics) of strong interest, or for longer or shorter post-demonstration intervals.

Having run this analysis, what is one to make of it? The importance of ***findings on outcome levels*** can best be understood in terms of the basic asperations of a demonstration's sponsors. Typically, sponsors desire that:

> *Demonstration participants perform at an acceptable level in the labor market and experience an adequate level of economic security through employment following demonstration participation.*

Where the sponsor agency or Congress provides a specific standard for what is "acceptable" (e.g., at least 26 weeks of work per year at 30+ hours a week) and/or "adequate" (e.g., family income that exceeds the poverty line), outcome levels in relation to that standard constitute the major mark of success--or failure--of the demonstration. Even without explicit standards, the demonstration's sponsor always needs to know the level of outcomes experienced by participants once they leave the project. Certainly, outcome levels are ***the only*** thing that matters to ***participants***: either they have reached a particular level of labor market success and quality of working life or they have not. Once they have reached this level, participants probably feel little need to understand what factors took them there. As the "bottom line" for constituents, then, DOL/ETA should also take note of outcome levels to see if they have any implications for the demonstration intervention or policy more broadly.

For example, an outcome evaluation might quickly determine that--whatever its other virtues--a demonstration that provides work experience and transitional employment to low-skill, childless adults leaves almost all participants with incomes below the poverty line. If this outcome is simply unacceptable, policy makers will have their answer:

> *Exclusive reliance on work experience and transitional employment does not get the job done; something else will have to be substituted or added.*[63]

---

[62]The most convenient test first calculates individual-level measures of change for each participant, then checks whether the average value of this change measure differs significantly from 0 using a standard t-test.

[63]While not needed to pass judgment on the overall demonstration relative to its aspirations for participants, information on the demonstration's contribution to or ***impact*** on observed outcomes would help in devising the next treatment strategy to attempt. For example, if the original policy tested adds to income—but not enough to raise participants above the poverty line—DOL/ETA might consider a new treatment strategy that combines work experience and transitional employment with, say, wage supplements. In contrast, if the demonstration intervention were known to have no impact on income, the next attempt to alleviate poverty in the target population should leave it out.

On the other hand, if demonstration participants live above the poverty line following the demonstration, decision makers may wish to continue the intervention as the one ***known*** way to attain that result--whether the demonstration policy itself made any difference or not.[64]

Analysts can also draw normative judgments and policy implications from ***findings on changes in outcomes over time*** when a demonstration sponsor makes ***improvements*** in participant circumstances an explicit goal of the intervention. For example, suppose DOL/ETA runs a demonstration project that provides English language training to immigrants with the goal of increasing their job security. The explicit goal of the intervention might be to move participants into jobs with an average employment duration twice that of their previous jobs. If so, researchers could use data on individuals' job tenure to calculate and compare average pre- and post-demonstration tenures, judging the intervention a success if the latter is at least twice the former. The absolute level of this outcome would not matter in this assessment, nor would the incremental contribution (i.e., impact) of the demonstration intervention to that result.

7.8 Inferring Demonstration Impacts

The final--and, for the purpose of passing judgment on a demonstration's success, most telling--way to examine participant outcomes is to infer the impact, or incremental contribution, of the intervention on those outcomes. If calculated correctly, an impact measure answers the questions at the heart of many, if not most, pilot and demonstration tests of new policy ideas:

Does the test intervention help participants?
If so, in what ways and how much?

When viewed as economic policies, government decisions to intervene in free markets can be justified only if they produce benefits for some subset of citizens, including the collection of citizens referred to as "society as a whole."[65] Otherwise, it would be more efficient, and more helpful to citizens generally, to spend the same resources on other projects--or on tax reductions.

---

[64]This conclusion may seem odd to some readers, particularly those steeped in the tradition that all government interventions must justify their existence. Yet it is fully consistent with a common short-cut for making social policy decisions: "If it ain't broke, don't fix it." One can see why this is the case by considering a world where policy makers study the without-demonstration world and the demonstration intervention in ***reverse order***. With this reversal, the target population ***starts out*** with work experience and transitional employment services in place and living above the poverty line. Here, the alternative policy option—eliminating work experience and transitional employment for this population to see if the same income goal could be met without it--***is never even considered***, since few people care whether work experience and transitional employment contributed to the result or not. It is sufficient to maintain support for the intervention that the desired social result takes place with the intervention and might ***not*** under any other policy configuration.

[65]One could base policy interventions on other criteria such as political appeal or moral power, dropping efficient use of resources as the top criterion. However, historically formal evaluations of social policies have been based almost exclusively on economic efficiency, on the assumption that non-economic factors will enter the policy debate as government officials and elected representatives consider continuation or expansion of a test intervention once evaluation results emerge. DOL/ETA has long emphasized economic returns as the preferred measure of success for labor market interventions, beginning in the 1960s with studies of the Manpower Development and Training Act (MDTA). See, for example, Borus (1964), U.S. Department of Labor (1970), and Smith (1970).

So how does one decide if a demonstration intervention helps participants? Almost always, the first step is to focus on a demonstration's impact on the average participant.[66] If the average participant benefits (or, conversely, is harmed) by the intervention, then on net the participant group as a whole has benefited (or has been harmed).[67] The challenge, then, is to compare the average outcome in the participant group with what ***would have happened*** on average to the same people absent the demonstration. The first of these concepts, participant outcomes with the intervention in place, is covered by the analysis of outcome levels just described. But researchers cannot measure the second concept--outcomes for participants if there were no intervention--so directly, since it does not exist. Because the world without the intervention never takes place, it must be simulated rather than observed. Evaluators call this simulated world the ***"counterfactual,"*** since it provides a counterpoint to the observed, factual world.

To calculate a demonstration's impact, researchers apply a formula that compares observed "with-demonstration" data with simulated "without demonstration" counterfactual data. The same formula is used regardless of the outcome measure involved--i.e., whether the analysis focuses on job stability or high school graduation or family income or anything else:

Estimated impact = average participant outcome - average counterfactual outcome.

A more formal, mathematical statement of this formula appears in Appendix A.[68] Of course, the estimated impact is a statistical estimate with a margin for error in either direction and so only ***approximates*** the true average impact of the demonstration. By taking account of this margin of error, one can determine with a high level of certainty whether the true average impact of the demonstration is greater than 0, less than 0, or too close to call.[69]

---

[66]A demonstration can affect different participants in different ways. For example, the JOBSTART Demonstration reduced the annual earnings of male high school dropouts by around $300 per year but had no effecst on the earnings of female dropouts (Cave and Doolittle, 1991). No doubt, individual participants ***within*** each of these subgroups were also affected to differing extents.

[67]Impacts cannot be measured reliably for individual demonstration participants, given the overwhelming difficulty of constructing a reliable "counterfactual" (see below) for each individual participant. This makes some form of group comparisons inevitable. While comparisons of group averages are by far the most common, researchers sometimes compare the ***entire distribution*** of observed participant outcomes with their hypothetical distribution in a postulated counterfactual world. Such an analysis makes sense when looking for impacts on, for example, the distribution of family income, where comparison of the share of families in various income ranges adds substantial insight to a simple measure of average income gain. Similarly, this technique can be used to determine if a dislocated worker intervention changes the distribution of unemployment durations, as well as the average duration.

[68]Appendix A also explains how the formula can be extended to control for differences between the participant and counterfactual scenarios unrelated to the demonstration.

[69]Appendix A explains how this is done. At least one evaluation of a DOL/ETA demonstration met the required level of proof for a ***negative*** result. The JOBSTART Evaluation (Cave and Doolittle, 1991) found a statistically significant reduction in the earnings for male dropouts using a highly reliable research method, random assignment (see below).

The greatest technical challenges in performing an impact evaluation spring from two aspects of the counterfactual:

- Findings on participant impacts cannot be produced without a counterfactual; and
- No one has ever seen the true counterfactual, leaving to evaluators the unenviable job of constructing it reliably out of "thin air."

The hardest--but inevitable--part of these challenges flows from the word "reliably" in the second bullet. Measured as the average participant outcome minus the average counterfactual outcome, impact estimates are only as good as the weaker side of relationship. Inevitably, this is the counterfactual side. The next three subsections discuss ways of defining a reliable counterfactual to a demonstration intervention.

7.9 Representing the Counterfactual with a Comparison Group

As with average outcome measures for participants, simulated average outcomes for the counterfactual should reflect a mixture of results for a varied set of individuals. Evaluators accomplish this by measuring outcomes for a varied set of non-participants and treating those outcomes as data on the counterfactual world. The individuals in this group--while differing among themselves--must be united in their non-participation in the demonstration and their collective resemblance to the participant group.[70] Impact analysis compares the average outcome for the non-participants to that of participants to calculate the demonstration's impact. The former group--the source of the counterfactual--is known as the demonstration ***"comparison group."***

To succeed in impact evaluation, a researcher must first find a comparison group whose baseline characteristics differ as little as possible from those of participants. Note that this ***does not mean*** that comparison group members must have had no contact with the demonstration. Rather, they cannot have received the demonstration intervention itself. If they had other contacts with the project--say, through outreach and intake--but never got the treatment, they are not disqualified as comparison group members.[71] Neither are individuals who receive services similar to the demonstration's from ***other*** community sources such as the Employment Service or the local welfare office. These types of assistance occur normally in a world without the demonstration, since other forms of employment and training assistance almost always are available in the community.

---

[70]The number of individuals required for a comparison group will be discussed in the final poriton of this section. It does not have to equal the number of participants in the demonstration.

[71]As explained in a moment, some of the best (or at least the theoretically most compelling) comparison groups are defined by interactions with the demonstration's outreach and intake processes.

How does one identify (and later collect data for) a reliable comparison group? For employment and training interventions, many comparison groups have been tried over the years[72] including:

- ***Non-applicants*** in the demonstration sites (including those eligible for, and those not eligible for, demonstration participation);
- ***Individuals who drop out during the intake process*** (those who drop-out prior to completion of the process, and those screened-out by demonstration staff);
- ***No-shows*** (those who enroll in the demonstration but then fail to show up to receive services);
- Experimental ***"control group" members*** (individuals picked at random just prior to enrollment and excluded from the demonstration for research purposes);
- ***Comparison site residents*** (those living in non-demonstration sites who meet the demonstration's eligibility criteria); and
- ***Pre-post comparison cases*** (the demonstration participants themselves, observed over an interval ***before*** demonstration entry).

These comparison groups differ on a number of factors that can affect the group's ability to portray what would have happened to participants in a world without the demonstration. As is well known, randomly selected control groups do the best job in this regard and, hence, deserve first consideration when designing an impact analysis. Unlike all the other options on the list, experimental control group members do not differ from participants in any systematic way, since both are drawn atrandom from the same pool of qualified applicants.[73]

---

[72]See Bell et al. (1995) for a comprehensive review of the comparison groups used in evaluating DOL/ETA programs and demonstration projects since 1964.

[73]This description glosses over the possibility that the participant group might not include all of the individuals selected at random for project enrollment (the experimental "treatment group"). Some enrollees never participate in the demonstration, a group typically called "no-shows." Like other non-participants, "no-shows"assignment experiment creates matching treatment and control groups, the former comprised of both participants and no-shows. To take advantage of this overall match, experimental impact analyses begin by comparing the average outcome of participants plus no-shows to the average outcome of control group members. The difference between these averages incorporates both demonstration impacts on actual participants and 0 impacts on no-shows. A standard statistical adjustment called the "no-show correction" divides this difference by 1 minus the no-show rate to produce a measure of the demonstration's average effect on true participants. For example, an average treatment/control earnings differential of $400 per year translates to $500 per year when one divides by 1 minus the no-show rate, if the no-show rate is 20 percent: $400 / (1 - 0.2) = $400 / 0.8 = $500. This means that participants gain an average of $500 per year because of the demonstration, while treatment group members gain only $400 on average, a mix of $500 effects on the 80-percent participant sample and $0 effects on the 20-percent no-show sample. The no-show correction, introduced by Bloom (1984), is very widely used in employment and training impact evaluations and does not require any assumptions about the process by which some treatment group members participate while others become no-shows.

However, experimental evaluations present special operational challenges that can increase demonstration costs (see below), or may not be feasible in certain situations;[74] in those cases, researchers must turn to other, second-best comparison group options.

Exhibit 7.2 breaks down and reorganizes several of these comparison group options by diagraming the client intake process for a typical DOL/ETA demonstration. The comparison groups shown (see the left and right columns of the exhibit) are ***internal*** to the demonstration,defined by the intake process itself. (***External*** comparison groups--which come from different sites and/or time periods than the demonstration--are discussed in a later subsection.) The exhibit highlights the two most powerful determinants of similarity and difference between potential internal comparison groups and the demonstration participants they will represent:

- Proximity to participants during the intake process; and
- Reasons for non-participation, self-selection versus program selection.

The central column of the exhibit shows how individuals flow through the various steps of the intake process. All potential participants begin as members of the general population in a demonstration site. This pool is, of course, huge and extremely diverse, but it narrows sharply through a series of steps that isolate sequentially (moving down the column):

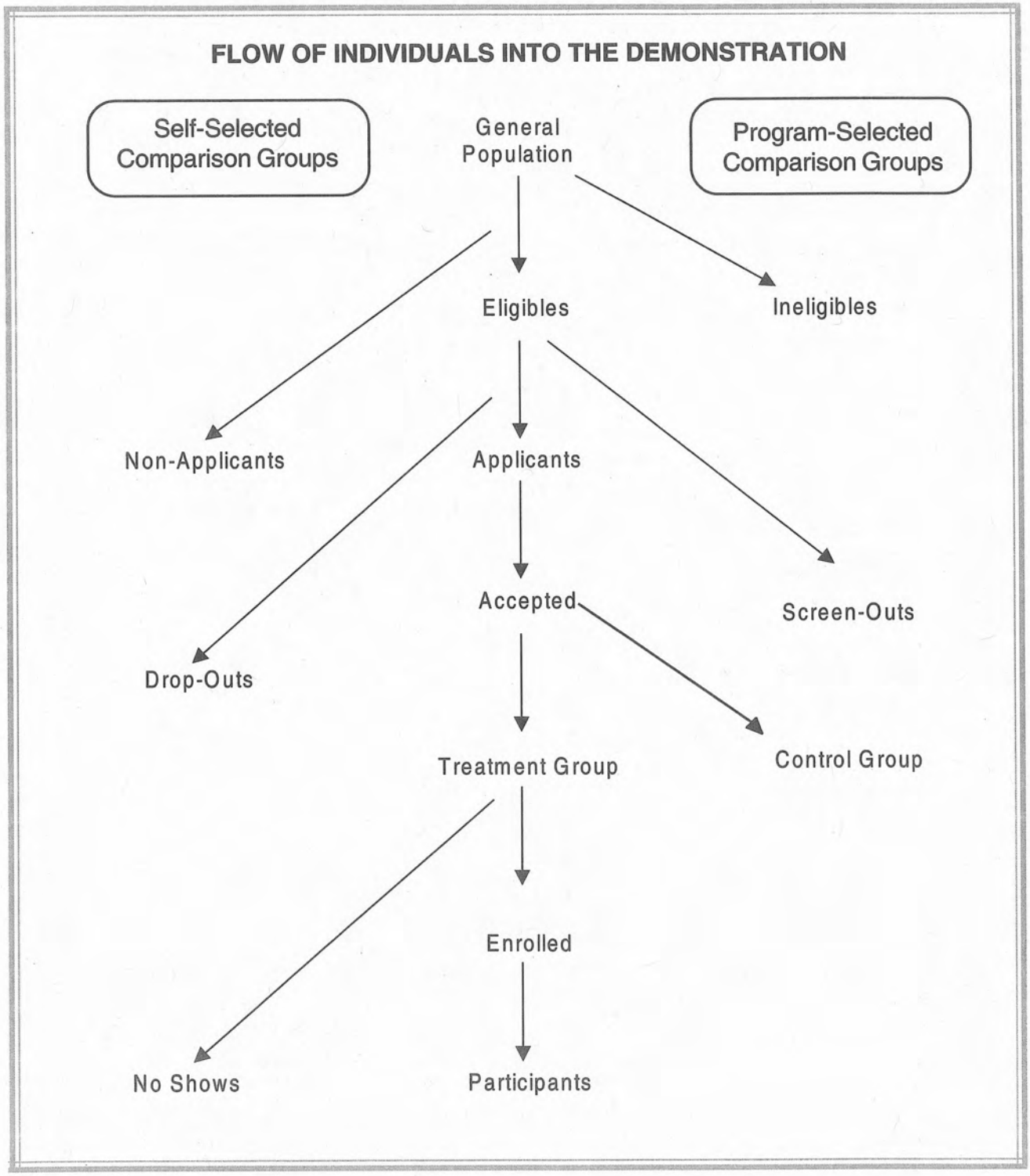
- ***Eligibles*** as a subset of the general population;
- ***Applicants*** as a subset of the eligibles;
- Individuals ***accepted*** for demonstration entry as a subset of all applicants;
- ***Treatment group*** cases as a subset of accepted cases;
- ***Enrolled*** individuals as a subset of all treatment group members (if enrollment is not automatic following random assignment to the treatment group), and
- Actual ***participants***.

---

[74]Orr (1998) examines the feasibility of random assignment in different settings in detail. He also provides an in-depth discussion of the research methods and trade-offs evaluators face when conducting random assignment impact evaluations. Time does not permit a further examination of these issues here.

Exhibit 7.2

Where to Find "Internal" Comparison Groups that Did Not Participate in the Demonstration - Some Possibilities



This narrowing takes place as portions of the general population "peel off" from the main flow to become non-participants of one type or another. Those who choose to leave the flow of their own volition appear on the left side of the exhibit and include ***non-applicants***, applicants who ***drop out*** during intake, and ***"no-shows"*** who fail to show up for demonstration services once enrolled. All three of these groups are defined through the process of "self-selection" at the initiative of the applicant. In contrast, the non-participants on the right side of the exhibit are defined by "program selection"--issues surrounding demonstration eligibility and intake staff discretion. Program selection can create up to three potential comparison groups: individuals categorically ***ineligible*** for demonstration participation, those found eligible but ***screened out*** during intake as less than prime candidates for demonstration services, and those assigned to a randomly selected non-participating ***"control group"*** for research purposes (see below).

7.10 Strengths and Weaknesses of Internal Comparison Groups

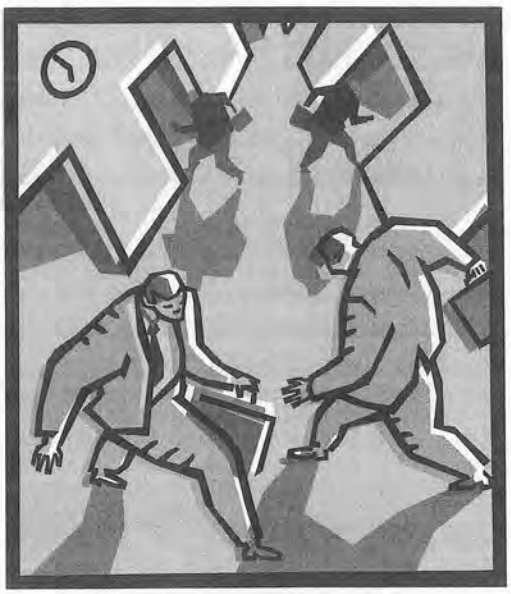
Of the many options for selecting an internal comparison group, some retain a fairly close ***proximity to demonstration participants*** throughout intake, in both character and location on the diagram, while others do not. No-shows, for example (bottom left in Exhibit 7.2), follow the same path as participants through five intake steps: like participants, they are eligible applicants accepted into the demonstration and assigned to the treatment group, then subsequently enrolled. They differ only in their participation patterns following enrollment. The fact that these two groups flow down many of the same paths during the intake process implies that they have a good deal in common, with the essential exception that no-shows do not actually participate in the demonstration while participants do. At the other extreme, ineligible individuals (top right in the exhibit; e.g., high school honor students in a demonstration of school-to-work services for non-college-bound seniors) may have little in common with participants. This is consistent with the fact that ineligibles leave the main intake flow at its very first step, five steps ahead of where participants and no-shows part company. Other potential comparison groups lie between these two extremes, both locationally and, one would expect, in their correspondence with the participant group on baseline factors. This suggests a general rule for picking an internal comparison group:

Rule 1: Non-participants who most closely follow the path of participants during the intake process--and end up most proximate to them in an intake flow diagram--make the best comparison group candidates.

Reasons for exclusion can also exert a powerful influence on the relationship of non-participants to participants. By definition, all non-participants differ systematically from participants on at least one selection factor-- the factor that led to their exclusion from the demonstration. This being the case, the challenge in conducting impact analysis is to determine which intake factor is least likely to affect outcomes in its own right. By choosing a comparison group that differs systematically from participants on just that one factor, demonstration sponsors and evaluators can minimize the distortion, or statistical bias, of the impact comparison caused by pre-existing differences between the two groups. Two different "families" of exclusion factors are highlighted by Exhibit 7.2: self-selection and program selection. Examining reasons for exclusion at this level provides some useful lessons for comparison group selection.

Program selection (other than random assignment to the treatment and control groups) creates the largest systematic difference between participants and non-participants. Program selection criteria, particularly those that define categorical eligibility, tend to involve major factors that can vary greatly between the selected and rejected groups. For example, the San Diego Immigrant Training Demonstration Project admitted only working-age Latin American immigrants without functional English-language skills.[75] Each of these factors—age, country of origin, and communications skills—could sharply distinguish participants and ineligible non-participants on various labor market outcome indicators, independent of the effects of the demonstration. Their combination would surely create a large amount of program selection bias if one compared the two groups to measure demonstration impacts. Discretionary exclusion by demonstration intake staff could also sharply differentiate the excluded group—screen-outs in this case—and the participants. In the AFDC Homemaker-Home Health Aide Demonstrations, for example, intake staff clearly "creamed" by selecting the best-educated and most work-experienced individuals from the pool of categorically eligible applicants.[76] These same factors have been found to strongly influence subsequent earnings levels, independently of the intervention, and thus also could cause substantial program selection bias were screen-outs used as a comparison group.

On the face of it, then, the program-selected comparison groups appear to be seriously flawed (again excepting randomly-selected control groups). Surprisingly, however, program selection does not pose the greatest threat to reliable impact estimation using internal comparison groups. This is because all differences between participants and ***program-selected*** non-participants stem from individuals' ***observable, external characteristics*** such as age, work experience, criminal record, and so forth. This is true even when demonstration staff use subjective factors in the screen-out process, such as judgements about a candidate's "suitability" for or ability to benefit from demonstration services. By definition, anything systematic that influences staff conclusions about, for the most part, total strangers must come from external cues of some sort—both objective and subjective, relevant and discriminatory, obvious and subtle.



[75]See San Diego Consortium & Private Industry Council (1995).

[76]See Bell and Orr (1988).

With external factors, an evaluator can at least hope to measure and control for participant/comparison differences when calculating impacts.[77]

In contrast, ***self-selected*** non-participants differ from participants on both observable factors and ***factors that may be entirely internal to the individual***, including such characteristics as motivation and resourcefulness in meeting the logistical challenges of going to work or training regularly (e.g., transportation arrangements, child care). By definition, these factors are unobservable to intake workers until an applicant actually enrolls in the project and begins receiving treatment. For the same reason, they cannot be measured by evaluators and used to adjust impact comparisons, as can program-selection factors. So while the internal differences that lead to self-selection may be of lesser consequence initially, one would expect them to persist longer—and do more damage—than program-selection differences when estimating demonstration impacts.[78] This suggests a second rule for selecting internal comparison groups:

Rule 2: Program-selected non-participants make better comparison group members than self-selected non-participants, if data on the factors that determine categorical eligibility and discretionary exclusions are available at the individual level for both participants and non-participants.

The combination of Rules 1 and 2 points clearly to the most-promising comparison groups generated internally by the demonstration: program-excluded groups that part company with participants as late as possible in the intake flow. A glance at Exhibit 7.2 shows these to be the screen-out group and—if possible—a randomly-selected control group. How well will these groups simulate counterfactual outcomes for the participant sample? By construction, an experimental control group provides an ideal "mimic" for the hypothesized without-demonstration outcomes of participants, since control group members differ from treatment group members only through chance variations in the random selection process.[79] With enough individuals in each group, even these differences are eliminated, making the control group an exact replica of the treatment group without the demonstration treatment.

---

[77]This is accomplished by first measuring external factors in a consistent fashion for both groups and then including those factors (as right-hand-side *X* variables) in the regression equation used to measure demonstration impacts described in Appendix A.

[78]Bell et al. (1995) develop this argument in detail, consider its implications for impact estimation, and test its validity using data from the AFDC Homemaker-Home Health Aide demonstration evaluation.

[79]As explained earlier, the distinction between treatment group members and actual ***participants*** (the omission of no-shows from the latter group) can be offset during impact analysis by using the widely-accepted "no-show correction" technique.

Screen-outs cannot approach this ideal, since they systematically differ from participants on the selection criteria used by intake workers when making discretionary picks for the demonstration. Some of the observable factors that cue intake staff are likely to remain unmeasured and unremoved from the impact comparison while others (like intake workers' ratings of applicant "potential") may be measured inexactly or inconsistently across cases. In the one known application of the screen-out approach (Bell et al., 1995), the impact estimates suffered from selection bias during the demonstration period but not after demonstration exit.[80] While this represents only the beginning of the evidence in either direction, in principal:

✓ *Screen-outs constitute the most promising internal comparison group when random assignment of accepted applicants to treatment and control groups is not possible.*

Orr (1998) describes the circumstances in which random experiments are—and are not—feasible. In employment and training program evaluation, there are few of the latter:

✓ *Randomized experiments are almost always feasible when evaluating the impacts of DOL/ETA pilot and demonstration projects, if time and budget allow and other factors (e.g., a community-wide "saturation" treatment) do not intrude.*

7.11 Other Comparison Group Options

External comparison groups add further options to the menu of techniques available for measuring demonstration impacts. These include:

- Comparison-site designs, which compare participants in demonstration sites to individuals in non-demonstration sites who meet the demonstration eligibility criteria;
- Pre-post analyses, which use changes in outcome levels—pre-demonstration versus post-demonstration—to measure demonstration impacts;
- Strategies that combine two or more comparison groups, either in a single analysis (e.g., pre-post comparison-group studies) or use them separately to obtain multiple but related impact estimates.

[80]Interestingly, a no-show comparison group worked even better than the screen-out group, showing an acceptably small degree of selection bias in both periods. Like Bell et al., other researchers have tested various comparison group impact estimates against unbiased experimental estimates to determine how well each comparison group matches up to a randomly-selected control groups as a representation of the counterfactual. None have looked at internal comparison groups, however. Results are mixed, in part because none of the other studies established a clear standard for measuring what level of reliability is acceptable relative to the experimental norm. The main lesson from these analyses is that comparison group strategies in general are quite unreliable and can yield widely varying impact conclusions for the same demonstration using identical data sources. (Barnow, 1987, and Bell et al, 1995, provide overviews and assessments of this literature.) Most of this earlier work used the National Supported Work Demonstration as its test case; see LaLonde (1986), Fraker and Maynard (1987), Couch (1992), and Friedlander and Robins (1992).

A final impact analysis strategy supplements the comparison group approach by confirming (or failing to confirm) that the hypothesized "pathways of change" for demonstration participants are active.

Comparison-site designs broaden a demonstration evaluation geographically in search of individuals that match demonstration participants without participating in the project. When compared to internal comparison groups, out-of-site "external" comparison groups have one distinct advantage: like randomly selected control groups, they ***could***—in principle—include individuals who are ***indistinguishable from participants on both program and self-selection factors***. Presumably, the traits that combine to lead certain people to demonstration participation in the research sites exist as well—and in the same combination—for certain people in other locations. In that sense, going out-of-site removes the selection bias problem: the right people are there and have not received the demonstration treatment. The problem is finding them. If, as suggested earlier, it may not be possible ***after the fact*** to measure and control for all the factors that influence participation, one cannot expect the same factors to successfully ***predict*** who in a different community would participate if given the opportunity. Yet, identifying that select group—or at least coming close—is crucial to the success of a comparison site design.[81]

On other factors, an out-of-site comparison group seems less compelling conceptually than certain of the internal comparison groups discussed earlier. Most obviously, an off-site comparison group faces a different labor market environment than demonstration participants, driving a wedge between likely outcomes in the two sites that has nothing to do with the demonstration. One way to address this hazard is to include a great many demonstration and comparison sites in the design, selecting at random which communities go into each group and implementing the intervention in only those places assigned to the demonstration group. To be safe from chance but potentially extreme differences between the two sets of sites, this design must include many more demonstration communities that many projects can afford to serve or study.

Comparison site designs also create a number of ancillary problems not present when using internal comparison groups. First—even if the number of demonstration sites is not expanded—comparison site designs imply data collection in at least twice the number of locations needed otherwise. If random selection is not used, this approach also introduces the complex matter of picking a "matched" non-demonstration site for each demonstration site, usually from among hundreds of options. Third, zeroing in on even a broad pool of demonstration eligibles in comparison sites may require more nuanced baseline data on eligibility factors than exists in most surveys or administrative files. Friedlander and Robins (1992) obtained mixed results in the one known test of the empirical reliability of comparison site designs.

[81] In particular, a comparison group that includes the exact equivalents of participants intermingled indistinguishably with 10 or 20 or 50 times that many people ***not*** exactly like participants (i.e., people who would ***not*** participate if the demonstration if it came to their town) has little to recommend it over an internal comparison group that includes essentially the same people in the *same* location (other than the tiny slice of "participant-equivalent" individuals, all of whom, in the demonstration sites, are actual participants). Either way, the average comparison group outcome for impact analysis is dominated by the types of people who constitute the great majority of both comparison groups: individuals ***not*** just like participants. Depending on the demonstration participation rate among eligibles (which usually is quite low, around 5 or 10 percent), little is gained -- and possibly much is lost--by switching to another site.

The argument over ***pre-post comparison designs*** hinges almost entirely on conceptual issues. Treating each participant's pre-demonstration experience as his or her own counterfactual, while appealing, involves several risks. First, changes over time for an individual may reflect nothing more than life-cycle or learning patterns that have nothing to do with a particular demonstration or pilot intervention. Researchers can model and remove trends of this sort from pre/post comparisons whenever the outcome measure is observed for two or more periods prior to demonstration entry.[82] Trend adjustments work best when the data cover several pre-demonstration periods and include the periods with the most predictive power—those just before demonstration entry. Done in this fashion, pre/post comparisons have a good bit of conceptual appeal, since they compare ***the very same*** people in a demonstration and non-demonstration world. Views differ, however, on how far this takes one. For example, if person-specific factors dominate outcomes at any point in time, pre-post comparisons should work well—particularly if the factors involved remain fairly stable over time, such as gender, race, and years of schooling, or have predictable consequences when they do change (e.g., age, labor market experience).

If instead "timing is everything"when it comes to demonstration participation and new directions in the labor market, problems arise. The most well-known problem is the "pre-program dip," the tendency for workers to seek employment services only when they experience an important setback in the labor market caused, for example, by layoff, demotion, or ill health. As has been documented many times, this tendency virtually assures that earnings and other employment outcomes move in unconventional ways just prior to and just after demonstration entry—whether the demonstration intervention itself made any difference or not. Usually, employment rates and earnings decline steeply in the three to six months prior to entry, then rebound quickly with or without assistance as workers find new jobs or recover from illnesses. While analytic techniques have improved in this area, many pre-post impact analyses still struggle to adjust correctly for "pre-program dip" and other transitory labor market "shocks."[83]

Interestingly, demonstration projects may suffer less from this problem than ongoing programs. For at least some demonstration participants, it seems likely that the factor triggering entry is not a change in their personal circumstances but ***the availability of the demonstration treatment***. It is possible, for example, that the proven success of the Center for Employment Training's (CET) welfare-to-work program in San Jose—and the publicity surrounding it—induced some welfare recipients in the CET replication sites (Rogers, 1996) to change a long-stable, "steady-state" relationship between welfare and work and enter the demonstration just because it is there. If so, pre-post analysis should provide a fairly accurate sense of how CET services affect welfare and labor market outcomes. This pattern seems even more likely in the Lifelong Learning Demonstration, which deliberately sought out people in "steady state" in the labor market—mature, incumbent workers—when promoting a return to school (Bell et al., 1996).

---

[82] One pre-demonstration observation is not sufficient to establish a trend; two is the bare minimum, and three or more provide much more robust trend estimates.

[83] Though not in a demonstration context, Benus and Byrnes (1993) provide several examples of the kinds of methods one might employ to address "pre-program dip" in the context of re-employment services for dislocated workers.

Hybrid or multi-layered approaches to picking comparison groups have often been proposed in the absence of an experiment as a way around the limitations each non-experimental method. The most common and promising strategy would combine the elements of pre/post and internal comparison-group analyses to simultaneously control for person-specific effects and general changes in labor market conditions over time. In this approach, pre-demonstration data on the outcome measure (e.g., quarterly earnings) for participant and comparison group members plays a crucial role, since controlling for it removes general trends in the economy and both trend and level differences between participants and comparison group members prior to the demonstration. While a clear improvement on straight pre/post analysis, this hybrid should only be used if evaluators expect to obtain reliable time-series data on outcomes for both participants and non-participants. This is but one way in which one comparison group can offset the weaknesses of another, allowing evaluators to draw conclusions from several helpful—if not individually infallible—sets of results. Evaluators outside the experimental realm should look for others, as suggested by the context and policy questions to be addressed.

As a final option, all types of impact analysis could benefit from an extension of the ***"pathways of change"*** idea introduced in earlier sections.[84] Here, the evaluator looks for evidence that intermediate events along the expected "pathway" from demonstration inputs to participant outcomes are taking place for important numbers of participants. One gains most from this exercise when a participant/comparison group comparison suggests that the final step in the chain has been reached (e.g., positive impact is found on participant earnings) but doubts remain regarding the reliability of the estimation technique. If intermediate links in the chain are missing (e.g., participants rarely finish training, impacts on weeks worked or hours worked per week are trivial or negative), one should question any apparent effects at the far end. Alternatively, should one confirm each step on the path, the apparent effect on earnings is not certain but—having ruled out a potentially important source of contradictory evidence—can be reported with greater confidence.

The above discussion—while extensive and enlightening—does not provide a single ***"bottom line" for conducting impact analyses of DOL/ETA pilot and demonstration projects***. A more general framework is needed, one that can adjust to the special conditions that arise when studying different test interventions under a variety of circumstances.[85] One simple, seven-step version runs as follows:

1. In all cases, consider first designs that represent the without-demonstration "counterfactual" by using an experimental control group—a subset of qualified applicants selected at random and excluded from the project.

---

[84]Section 6 explained how operational evaluations can benefit from "pathways of change" analysis, tracing a demonstration's or program's intervention through a series of steps to final participant outcomes. Some studies make the question of linkages a major focus of the research (eg., the Evaluation of the Job Training for the Homeless Demonstration lists program linkages as one of its five major research topics; see James Bell Associates, 1998).

[85]Heckman et al. (1998) reach a similar conclusion, noting in their abstract that "[T]here is no inherent method of choice for conducting program evaluations. . . [Decisions] should be guided by the underlying economic models, the available data, and the questions being addressed."

2. With the advice of impact evaluation research experts, either (a) select the best possible experimental approach or (b) determine there is no way to manage random assignment from a technical or financial standpoint, or (c)—as occasionally happens—conclude that an alternative impact estimation approach would provide more useful or better quality results than an experiment.[86]

3. If the experimental approach is ruled out, favor designs with program-selected internal comparison groups (e.g., screen-outs)—and, for interventions likely to attract participants in "steady state" labor market situations, pre/post comparison-site designs.

4. Be sure of data quality and availability—and adequate sample sizes (see below)—for ***both*** the participant and comparison groups before committing to the design. Quality assurance is particularly important for variables slated to play a central role in overcoming the limitations of the selected comparison group (e.g., control variables for a regression analysis (see Appendix A), pre-project data on outcome trends).

5. If interest in an intervention's net contribution to participant outcomes dominates the research agenda motivating a demonstration test, pursue the best available impact estimation strategy even if it appears weak. Where several promising methodologies exist, try each to see if they obtain consistent results.

6. Do not over promise regarding the extent or conclusiveness of the information the study will provide on demonstration impacts—not to policy makers, internal funders, other interested parties, or one's self.

7. Of perhaps greatest importance, do not over-interpret impact findings once they emerge.

With regard to these last two points, impact findings are almost always helpful and worthy of investment, yet—at the same time—flawed to some degree by the limitations of research technology. When sponsors or evaluators oversell or misinterpret the findings, they undoubtedly will be put to uses that magnify those flaws.

---

[86]For two sides of this question, see Heckman et al. (1998) and Orr (1998).

7.12 Using Impact Results to Calculate Demonstration Benefits and Costs

Impact analysis measures a demonstration's effects on participants. Policy makers will also care about consequences for ***society as a whole***, which includes—in addition to participants—government agencies providing demonstration services and running related programs, the taxpayers who finance government services, and participant families. The final element of outcome evaluation, called ***benefit-cost analysis***, extends the study to these perspectives, showing how much society gains or loses from a demonstration intervention and the parts of society affected in each case.

As its name implies, benefit-cost analysis compares an intervention's benefits to its costs. Costs include not only the budgetary costs to the government agency or agencies delivering demonstration services (see section 5 above), but also any negative—i.e., unwanted—effects elsewhere in society. These include, for example, increased costs for other government programs (e.g., Pell grants for higher education, child care subsidies for working parents) and sacrifices made by families when participants return to work (e.g., loss of at-home production, out-of-pocket costs for transportation and child care). Social costs also arise when outcomes expected to change favorably (e.g., participant earnings) move in the opposite direction (e.g., earnings decline[87]).

Comparisons of a demonstration's achievements to its costs take two forms, each answering a different policy question about the intervention:

Cost-effectiveness analysis asks "What is society paying, per unit, for the results obtained?" Or, conversely, how much is accomplished for every $1,000 expended?

In essence, cost-effectiveness analysis tells policy makers the "bang for the buck" achieved by a demonstration intervention. A typical example would be a finding on how much was spent for each participant placed in a job by a school-to-work transition project such as the Transition to Work Demonstration Projects.[88] Questions of this sort are answered by comparing total demonstration spending on the intervention with different indicators of achievement taken from operational and/or outcome evaluations (e.g., number of participants who completed training, demonstration impacts on employment rates). There are many ways to configure these comparisons in ratios of different types, ways too numerous to be explored here. Instead, the focus turns to a more comprehensive and unified analysis framework for assessing a demonstration's pluses and minuses, benefit-cost analysis.

---

[87] This occurred, for example, in the JOBSTART Demonstration (Cave and Doolittle, 1991).

[88] This project, whose full name was the Evaluation of Transition to Work Demonstration Projects Using a Natural Supports Model, helped students with severe disabilities move from school to integrated employment. Its final report (Conley et al., 1995) did not include a cost-effectiveness analysis. Cost-effectiveness research, as well as benefit-cost analysis, were precluded by a decision not to collect financial data from participating agencies. Without this information, the evaluator could do little cost analysis other than to surface general issues and note some possible cost implications of the intervention using qualitative impressions and the views of various agency staff.

Benefit-cost analysis asks "Once all social benefits and costs are considered, is the test intervention worth what it costs?"

Not all outcome evaluations include a benefit-cost analysis, and perhaps only a minority of demonstration evaluations.

> *As the only way to check the return-on-investment from a demonstration initiative, or to reach a social "bottom line" on the intervention's merit,* ***all*** *outcome evaluations should include a benefit-cost study.*

As a relatively modest extension of an impact analysis, a benefit-cost study requires very little new data and no large-scale investments of time or staff resources. It also cannot be done credibly without a solid impact analysis to build upon. And in many cases, benefit-cost analysis reduces to a trivial exercise: if the main expected benefits of an intervention (e.g., increased earnings, more rapid re-employment, a lower high school dropout rate) do not materialize in the impact analysis, the final evaluation report can say immediately that the demonstration was not socially beneficial—i.e., that its benefits did not exceed its costs—without doing any new analysis.[89]

Benefit-cost analysis involves several analytic steps that reach beyond what is accomplished in an impact analysis:

> **<u>Step 1</u>: Produce impact estimates for additional outcomes, including the cost of the intervention.**

Added outcome measures for a benefit-cost analysis should be chosen in the same manner as the original impact analysis outcomes—as factors the demonstration could affect and that would matter if they should change. Attention is broadened here, however, to include all sectors of society, not just participants. Added benefit-cost measures typically include:

- the costs of services received from the demonstration and from other sources in the community offering similar services (e.g., a local welfare agency's Work First program),
- cost of administering income transfer payments (UI benefits, food stamps, etc.),
- taxes paid by participants, and
- lost leisure and home production.

[89] This happened, for example, in the youth component of the National JTPA Study (see Orr et. al, 1996).

Controversy sometimes erupts around this last item, depending on whether one views added work and school time as supplanting other ***socially-valuable*** activities, or just socially-unproductive leisure enjoyments. As long as the supplanted time would not have been spent in socially destructive activities (e.g., drunk driving leading to a fatal accident), this issue is easily resolved. From an economic perspective, if any individual gives up time that previously benefited him or her ***in any fashion*** without hurting society, the supplanted hours constitute a cost to that individual and, in consequence, to society as well and should be counted in a benefit-cost assessment.[90] This cost could be offset by attendant social benefits if time at work replaces socially destructive activities such as crime or binge drinking.

Step 2: Summarize each area of impact over the lifetime of a participant.

This step involves summing quarterly or yearly impact measures over time for the entire follow-up period.[91] Any cost or benefit likely to have continued past the end of the observation period (e.g., long-term earnings gains) must first be projected into later years, either by extrapolating the observed trend in impacts in the months just prior to data cut-off or by mimicking cross-time changes in long-run impacts shown in other studies of similar interventions with longer-term follow-up.

Step 3: Identify the part or parts of society to which each benefit or cost accrues.

When a demonstration or pilot takes place, many parts of society may gain or lose. For example, if a demonstration increases employment, various levels of government—federal, State, and local—could gain through lower income support payments and higher tax collections.[92] One can also think of these benefits as accruing to taxpayers, the people who fund the government. Other changes could occur for participants, who may experience increased earnings and greater self-esteem because of the demonstration. Finally, participants' families may benefit from higher incomes and lessened domestic strife.

All these consequences will have been measured by this point but must now be apportioned to the right sectors of society. The best technique for this step is to set up a spreadsheet with rows that represent different categories of benefits and costs and columns that represent various sectors of society. Check marks in the body of the table show the sector or sectors to which each benefit or cost accrues (assuming it occurs at all).[93]

---

[90]Putting the appropriate monetary value on each hour of lost leisure and home production is also difficult and controversial. Greenberg (1997) and Bell and Orr (1994) discuss this point, taking the view that an hour not spent in paid employment is worth at most the amount of income it would have generated if devoted to work (i.e., the individuals's hourly wage rate) and possibly a lot less (but more than $0).

[91]Appendix B explains how benefits and costs more than a year past demonstration entry must be translated into "present value" terms before including them in this sum.

[92]Costs also accrue to government, and to the other parts of society referenced here. For the sake of simplicity, these added effects have been left out of the illustration.

[93]More sophisticated versions of the matrix replace check marks with plus or minus signs to indicate whether a possible effect would be a benefit (+) or cost (-) to that segment of society. For an example, see Bell et al. (1996), exhibit 6.1.

Sometimes, a single benefit or cost item accrues to more than one segment of society, as happens when government and participants both incur higher child care costs. Also, a benefit to one part of society can represent a cost to another part; for example, reduced welfare payments benefit government/taxpayers but impose a cost on participants.

The various impact measures from previous steps are then entered into the spreadsheet in the appropriate cell (row and column). The matrix should also include effects of the demonstration that are not measured in dollars, such as increases in a numerical self-esteem scale or changes in marriage rates. Each of these effects are expressed in their own "natural" units. Even potential costs and benefits that researchers ***cannot measure at all*** (e.g., added taxpayer satisfaction when welfare roles decline) should appear in separate rows of the matrix as a reminder that, though unseen, these factors also may have changed in important ways.

Step 4: Analyze the benefit-cost figures

Analytic examination of benefits and costs proceeds in three stages. First, for each sector of society, the evaluator adds up all dollar-denominated benefits and subtracts from them total dollar-denominated costs. The end result, a single dollar amount either positive or negative, represents the "net dollar benefit"of the intervention to that segment of society. More precisely, it shows how much that segment of society gained or lost in aggregate from the participation of a single, average demonstration participant.[94] Second, one calculates "net dollar benefits" ***for society as a whole***, summing the figures for each segment. This one number, though not perfect (see below), encapsules much of what economic analysis has to say about a demonstration's worth. Third, net dollar benefits for individual social sectors and for society as a whole are related to measured ***non-monetary*** benefits and costs. Net dollar costs (negative benefits) constitute the price paid for any measured non-monetary benefits, since the two come together as a "package deal" through the intervention. If the non-monetary benefits seem worth this "price"—something for policy makers to gauge, not evaluators—the intervention could be judged a good use of taxpayers' money in economic terms.[95] ***Positive*** net dollar benefits also make the intervention attractive from an economic standpoint, unless important non-monetary ***costs*** show up that could outweigh these benefits; again, this is a judgment for policy makers to make, not evaluators.

Step 5: Acknowledge and explore the limitations of the analysis, then—if they make a difference—report alternative scenarios as well as the initial findings.

The benefit-cost approach described here, while very helpful and the best available numeric summary of a demonstration's worth, is certainly not perfect. Many of its limitations can be examined empirically, however, and attached as caveats to the main results. The first caveat is interpretational: the all-inclusive "net

[94] Recall from section 7.8 that the impact analysis measures the effects of the demonstration on the ***average*** participant. Summing over a lifetime, projecting future effects, discounting to present-value terms, and allocating the result to different segments of society does not change this: all benefit-cost figures derived from the original impact analysis remain in "per-participant" form. The same is true of any impact estimates generated as part of the benefit-cost analysis. Even the costs of demonstration services, which start out as aggregate totals, must be converted into per-participant format.

[95] This favorable view might not extend to political, institutional, or other perspectives on the intervention.

dollar benefits to society" figure from step 4 in fact does not include everything: it obscures a policy consideration of great importance, ***distributional equity***. The "full society" perspective implicitly treats all segments of society as interchangeable as recipients of benefits and costs—it does not care which sector of society gains or loses. Yet, at the same time, who gains—and at whose expense—permeates most social policy decisions.

To illustrate the issue, consider an intervention that achieves a positive net dollar benefit for society, but does so by imposing substantial net costs on participants offset by even larger gains for government/taxpayers. In a net social benefit calculation this will appear as an overall gain to society, yet most people would view it as a loss if the "losers"—the demonstration participants—were disadvantaged, low-income workers. The evaluator's responsibility in this situation, and in any benefit-cost analysis, is to see that the ***distributional*** consequences of the test intervention are not forgotten. This can be accomplished by putting higher emphasis on the sector-by-sector findings, where major distributional effects are not only visible but central to the situation. Researchers also should include disclaimers about the omission of distributional consequences when reporting findings for society as a whole.

Other limitations in benefit-cost analysis have more technical origins. First, because they do not lend themselves to numerical "summing up," tests of the statistical significance of various impact measures fall by the wayside when moving from impact to benefit-cost analysis. This means that all benefit-cost findings carry some amount of unacknowledged statistical uncertainty—uncertainty as to which measured effects are real (i.e., not equal to 0) and which estimated amounts provide dollar values reasonably close to the truth.[96] Mistaken dollar amounts make the greater difference, since the whole point of benefit-cost analysis is to compare and combine dollar amounts. Other technical issues that challenge the validity of the estimates include the assumptions made when projecting future benefits and costs from the intervention and valuing lost hours of leisure and/or home production.

To address these and other uncertainties in the numbers, benefit-cost studies should include a "***sensitivity analysis***" of the findings, which shows how vulnerable a study's conclusions are to mismeasurement and inaccurate assumptions. Analysts execute these checks by systematically varying the magnitude of the individual cost and benefit levels to reflect statistical uncertainty and unproven assumptions and then re-deriving the net-dollar-benefit results. Hopefully, many different scenarios are checked in this fashion as evaluators vary one factor at a time and then several at once. To insure reliable interpretation of the numbers, researchers should at a minimum construct two alternate scenarios: one pessimistic and one optimistic. The pessimistic scenario moves each input factor of important magnitude in the direction that will ***reduce*** net dollar benefits for society as a whole, using a degree of adjustment that seems both substantial (i.e., not too small) and conceivable (i.e., not too large). The optimistic scenario goes the other direction, adjusting inputs to ***increase*** net dollar benefits. If the overall findings of these analyses differ from the original results, the new, adjusted scenarios should be reported along with the original results and characterized as other possible outcomes of the intervention. If the original conclusions do not change, the evaluator needs only to report that a sensitivity analysis was done, indicate the factors that varied and the types of scenarios checked, and announce that the basic "story line" of the results is insensitive to changes in underlying estimates and assumptions over a reasonable range.

[96] This is not assured, and may be way off, due to statistical margins of error.

7.13 Sample Size Requirements and Survey Non-Response

The appropriate size of the participant group sample—and, for impact analysis, the chosen comparison group—has come up at several points in the preceding discussion. In outcome evaluation, two distinct sample size issues can arise:

- How many participants to serve in total, and then study (perhaps in conjunction with comparison group members) using relatively inexpensive administrative data, and
- How many individuals to include in a participant follow-up survey or surveys.

Given the investment required to expand demonstration intake and service delivery to more participants—and the substantial variable costs of survey data collection driven by sample size—these are important questions.

On a common-sense level, any demonstration worth testing should involve enough participants to assure that observed results do not reflect simply the quirks of a small group. Statistically, one needs confidence that measured project outcome levels, changes, and impacts are close to true project outcomes, a desire thwarted by the margins of error that apply to all statistical measures.

Margins of error in statistical measures arise from the possibility that the specific individuals studied depart from the norm for all participants by chance, a phenomenon known as "sampling error."[97] This error can be removed, or at least minimized, by studying large samples: just as flipping a coin 1,000 times virtually assures an even split between heads and tails, applying an intervention to a very large number of individuals almost guarantees that the group as a whole will experience typical outcomes following the intervention. Whatever is distinctive about any small subset of participants due to sampling error tends to go the other way—and, on net, to balance out—when analyzing a large number of participants. When drawing conclusions from a demonstration sample, just as in any other statistical analysis, there is safety in numbers.

There is also ***danger*** in numbers if large samples threaten the operational feasibility or affordability of a demonstration. Designing and counting on a research approach that hinges on a large number of participants has no value if that number, or something near, is not reached. Worse yet is an evaluation plan that—in a demonstration of ample size—proposes to conduct follow-up interviews with a large number of participants, then finds that the costs of data collection in relation to the overall research budget makes doing so impossible. Obviously, there is a lot to be said for coming to the right place, size-wise, when designing a pilot or demonstration project and its evaluation.

[97] Other sources of error in research statistics may also occur, including mis-specification of the outcome variable, measurement error, and survey non-response. All but survey non-response are unrelated to sample size and are not examined further here.

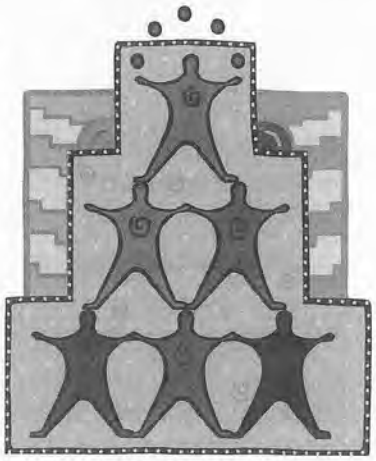
How does one draw a balance between data needs and budget and operational realities? By developing a formal statistical measure of the potential for sampling error and mistaken research conclusions. There are many different ways to do this, but the one that works best in demonstration and pilot evaluations centers around the concept of "minimal detectable effect." A minimum detectable effect, or MDE, indicates the size of impact that statistical analysis can detect. It might indicate, for example, that an earnings impact of $500 per participant per year is detectable—i.e., that when such an effect occurs, statistical analysis should find enough evidence of it to conclude that some positive effect occurred. In this situation, any true impact equal to or larger than $500 would likely be "found" by the evaluation, while smaller impacts have a greater chance of going undiscovered. For this reason, analysts and research sponsors would like MDEs to be as small as possible so as to avoid rejecting a beneficial intervention simply because of limited "detection" ability.

Each impact measure—for earnings, length of unemployment, hours worked, welfare benefits received, etc.—has its own MDE. In employment and training studies, researchers tend to rely almost exclusively on the MDE for earnings when making decisions about sample size. As the most crucial outcome—and one that encompasses many others—this makes sense. Several factors determine the size of an earnings (or any other) MDE:

- The variability across individuals of the outcome measure involved (e.g., duration of unemployment has more variation than annual employment rate);
- The standard of statistical evidence applied—conventional standards are quite demanding, insisting that certain types of mistakes in inference occur no more often than 1 time in 10;[98] and
- The sample size analyzed—in the case of impact analysis, the sample sizes of both the participant and comparison groups.

[98] The 1-in-10 standard applies to the mistake of concluding that a non-zero impact has occurred when, in fact, the true impact is zero. In statistical parlance, this result is referred to as a "type 1 error," and its rate is set by choosing a significance level of .10 for a two-tailed t-test of the null hypothesis that true impact is 0. Almost always, a 2-in-10 standard is used with regard to "type 2 error"—overlooking a real impact when it does occur.

Less outcome variability, lower standards of evidence, and larger sample sizes reduce MDEs,[99] allowing detection of smaller and smaller effects. Researchers have no control over the variability of outcomes and—with good reason—are loathe to reduce standards of statistical evidence.[100] ***This creates a direct trade-off between lowering MDEs and keeping analysis sample size—and therefore the scale of both the demonstration and follow-up surveys—within reason***. In demonstration evaluations featuring impact analyses, sponsors and evaluators usually settle on samples of 1,000 to 5,000 participants in total, and a roughly equal number of comparison group members.[101] This provides a good chance of detecting earnings impacts of over $400 per year (with 5,000 participants) or over $900 per year (with 1,000 participants) using administrative data on the full sample.[102]



Given the cost of data collection, participant survey samples tend to be smaller—from 500 to 2,000 participants and, for impact analyses, an equal number of comparison group members. These surveys sometimes have difficulty detecting impacts of modest proportions but will certainly discover major demonstration effects. The exact level of detection provided depends on the outcome measure used, since different survey measures have different person-to-person variability. With earnings and income support benefits now covered almost exclusively by administrative data, the most important survey measures for an impact or benefit-cost analysis vary from study to study. As a result, survey sample sizes that work in one evaluation may not in another; the only way to be sure is through careful statistical analysis in each case.

[99]The computation of specific MDEs involves a complex mathematical translation of these factors using sophisticated statistical tools. It cannot be characterized in a simple formula, though standard statistical software packages may include a computational tool. An evaluation sponsor should never try to make and interpret these calculations without the assistance of a professional statistician or econometrician with evaluation experience.

[100]Reducing the standard of statistical evidence would increase the risk of statistical tests "seeing" impacts that are not really there or missing impacts that do occur, or both. It also would lead to great skepticism regarding one's findings in the rest of the research community.

[101]For a fixed total sample size (e.g., 2,000 survey respondents), MDEs are lowest when participant and comparison groups are of equal size (e.g., 1,000 respondents each).

[102]These figures could differ by population, since some populations have more underlying variability in earnings than others. The figures here are based on a population of disadvantaged, frequently unemployed adult women eligible for JTPA Title IIA services. Disadvantaged men tend to earn more than women when they work, resulting in higher variability and larger MDEs than shown here. Tightly-defined populations with fewer labor market disadvantages (e.g., dislocated workers) could have lower earnings variability and, therefore, smaller MDEs. Numeric figures were derived from samples examined by the National JTPA Study (see Orr et al., 1996).

There is one constant in survey design that affects sample size (as well as other research considerations): ***survey non-response***. No survey ever collects data for all cases--individuals, households, organizations--of interest. Inevitably, some of those targeted for interviews do not respond: either they are not found by the process that locates and contacts intended respondents, or they do not agree to--or are not capable of[103]--supplying the requested information. The techniques--and pitfalls--of locating and obtaining cooperation from survey respondents are too numerous to review here; a whole literature on survey tracking and non-response addresses this dimension of program evaluation and provides better source material than can be provided here. However, this closing discussion does consider the *implications* of survey non-response for designing and executing demonstration and pilot evaluations.

From the standpoint of running an evaluation and its analyses, the inability to interview all intended survey respondents has three implications:

- Sample sizes for survey-based analyses drop below planned levels;
- More money is spent for each completed interview, reflecting the resources expended on unsuccessful interview attempts; and
- Available data do not necessarily reflect the mix of respondents targeted for interviewing, creating a risk of biased findings due to omission or under-representation of certain participant types.

The implications of the first two points are straightforward: evaluators must (1) select more "target cases" for interviewing than the number of cases needed for analysis (see above) and (2) include funds in the budget for some projected number of unsuccessful interview attempts as determined by the expected response rate and the number of completed interviews sought.

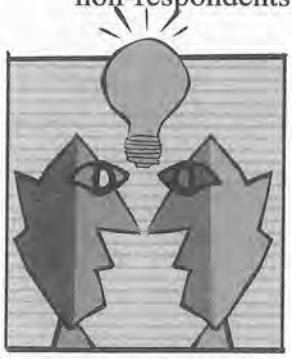
To obtain an appropriate *target* sample for interviewing--recognizing that there will be attrition through non-response--one guide is to multiply the number of completed interviews needed by the inverse of the expected response rate. Then, if the expected rate is exactly achieved, the survey produces exactly the number of completed interviews desired. For example, if one expects to complete 80 percent of all interviews attempted and wants 1,000 cases available for analysis, 1,000 should be inflated by a factor of 1.25, to 1,250 *attempted* interviews (1.25 = [1/.8]). In practice, it is best to add an extra margin of around 10 percent to this guide, since expected response rates are not always achieved. Continuing the example, this would result in 1,375 cases being selected for interviewing (1,250 + .10 [1,250]) and 1,000 interviews completed if the actual response rate reaches 73 percent.[104]

---

[103] Individuals who have died, become seriously ill, or encountered debilitating cognitive and/or communication disorders have generally lost the capacity to report the events and circumstances in their lives of interest to researchers, or at least to do so accurately. Other potential respondents may be "off limits" to data collectors while institutionalized or incarcerated. Similar circumstances may afflict organizations as potential survey respondents: businesses close, organizations disband, information sharing is constricted by legal issues, etc.

[104] Eighty percent response in this scenario would produce 1,100 completed interviews, the same 10 percent "bonus" built into the original target sample of 1,375.

The third implication of survey non-response--potential shifts in sample composition--is more complex. Not only will non-response result in more resources expended and fewer interviews completed, it may also change a study's findings in unwanted ways. If non-respondents differ from respondents in important ways, outcome and impact findings based on the survey--if not otherwise corrected (see below)--will suffer from "non-response bias." Put in blunter terms, the policy questions to be answered from the survey data will be answered for the wrong mix of people. The blend of people (or households, or organizations) who are untraceable or uncooperative in an interview situation often differ from demonstration participants generally on a range of factors related to outcomes. For example, they may be the most able and successful of the participant group and--because they have jobs--spend more time away from home or feel more pressed for time when asked to cooperate. Alternatively, the least successful participants may be hardest to interview, either because their living situations and/or access to telephones are more unstable than those of other participants or because their lesser skills and education make them more suspicious of interviewers or less capable of understanding survey questions. Other factors may impinge on one's success in interviewing respondents, further "driving a wedge" between respondents and non-respondents on factors affecting outcomes of interest to the survey.



In the face of these threats, evaluators have developed several methods for dealing with non-response bias when analyzing participant outcomes:

- The preferred approach is to **keep non-response low** enough that it cannot have an appreciable effect on measured outcomes. With response rates of 80 percent or higher, the missing 20 percent of the survey target population can have only a small influence on reported average outcomes--even if non-respondents differ sharply from respondents in terms of outcomes.[105] This "damping down" effect of the unusual attributes of non-respondents stems from their small place in the survey target population as a whole and ensures that even large distinctions between respondents and non-respondents do little to skew survey results when response rates are high. Primarily for this reason, survey clearance specialists at the Executive Office of Management and Budget seek response rates in federal studies of at least 70. Response rates below 70 percent are viewed as highly suspect in evaluation research unless extensive tests for non-response bias produce favorable results. Surveys with less than 60 percent response probably will not be credited by outside observers in any case since they approach the point where participants *not* interviewed are nearly as numerous as participants interviewed. Wherever possible, evaluators should try to attain response rates of 80 percent or higher when analyzing demonstration or pilot outcomes.[106] In practice, rates vary depending on the survey modality used, with office-administered data collection by demonstration staff exceeding in-home data collection by evaluators in terms of response rate, followed by telephone and mail-out surveys. This variation, though understandable from a practical point of view, should in no way alter the standard of acceptability one applies in analyzing the data once collected.

---

[105] For example, for measured outcomes of respondents to fall even 10 percent below outcomes for the entire target group, respondent outcomes have to be 36 percent below non-respondent outcomes on average. This contrast would need to be even greater to produce similarly-skewed results with a response rate above 80 percent. For example, with 85 percent survey response, respondent outcomes would have to fall fully 43 percent below non-respondent outcomes for 10-percent mis-measurement to occur. Regardless of the survey response rate encountered, evaluators should conduct hypothetical calculations of this sort to gauge the degree of bias possible in their survey-based findings.

[106] Baseline surveys should do better, particularly if administered in program offices as part of the demonstration intake process. Rates of 99 percent or more are not unheard of in this context; Orr et al. (1996) report a 99.5 response rate for baseline forms administered as part of the National JTPA Study.

- **Test for--and adjust for--non-response bias using an alternative data source** for which non-response is not an issue, such as administrative records on participant earnings from state Unemployment Insurance records. If the corresponding survey results are not appreciably skewed by non-response, one would hope (but could not confirm) that other survey-based results have little or no non-response bias. Similarly, a multiplicative factor that removes non-response bias from survey-based earnings measures could substantially improve biased results for other outcomes measured by the survey--particularly those outcomes closely associated with earnings such as weeks of employment and availability of health insurance.[107] Other survey measures that can sometimes be verified against external data include information on participant activities while participating in the pilot project or demonstration, receipt of government employment and training assistance over the follow-up period, and monthly welfare receipt and benefit amount.

- **Adjust the mix of respondents** to look more like the intended target population for the survey on characteristics measured at baseline without survey non-response bias. Most often, this approach **uses weights** to place greater emphasis on the types of participants under represented in the respondent data when calculating average outcomes for participants as a group. This approach has the advantage of applying to all outcome variables considered without investing in separate, new analyses for each individual outcome.

- Identify the most difficult-to-reach *respondents* using data on the number of contact attempts made before completing the interview, then **assume that outcomes for non-respondents are similar to outcomes in this "marginal respondent" group**--or that they depart even further from outcome levels for "non-marginal respondents" in the same direction. A direct test can then be constructed to determine if the absence of data on non-respondents significantly skews the results.

- **Collect a few key data items from non-respondents through an intensive "follow-back" survey,** to get a direct measure of respondent/non-respondent outcome differences. This approach works best for non-respondents who were reached in the initial survey wave but refused to cooperate--individuals who might be persuaded to take part if told they will be asked very few questions.

Groves and Wissoker (1999) provide nice examples of all of these strategies applied to the National Survey of America's Families.

[107] See Kornfeld and Bloom (1997) for an application of these principles to survey and administratively-derived earnings measures in the National JTPA Study.

References

Barnow, Burt S. 1987. "The Impact of CETA Programs on Earnings: A Review of the Literature," *Journal of Human Resources* 22 (2): 157-193.

Bell, Stephen H., Nancy R. Burstein, and Larry L. Orr. 1987. *Evaluation of the AFDC Homemaker-Home Health Aide Demonstrations: Overview of Evaluation Results.* Bethesda, MD: Abt Associates.

Bell, Stephen H. and Larry L. Orr. 1988. "Screening (and Creaming?) Applicants to Job Training Programs: The AFDC Homemaker-Home Health Aide Demonstrations". Paper presented to the Association for Public Policy Analysis and Management Annual Research Conference, Seattle, Washington.

_____. 1994. "Is Subsidized Employment Cost Effective for Welfare Recipients? Experimental Evidence from Seven State Demonstrations," *Journal of Human Resources* 29 (1): 42-61.

Bell, Stephen H., Larry L. Orr, John D. Blomquist, and Glen G. Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Bell, Stephen, Suzanne Reyes, Jane Kulik, Terry Johnson, and Mark Gritz. 1996. *The Lifelong Learning Demonstration: Evaluation Design.* Bethesda, MD: Abt Associates.

Benus, Jacob M. and Rhonda M. Byrnes. 1993. *The St. Louis Metropolitan Re-Employment Project: An Impact Evaluation.* U.S. Department of Labor, Employment and Training Administration, Research and Evaluation Report Series 93-B.

Benus, Jacob M., Terry R. Johnson, Michelle L. Wood, Neelima Grover, and Theodore Shen. 1994. *Self-Employment Programs: A New Reemployment Strategy, Final Report on the UI Self-Employment Demonstration.* Bethesda, MD: Abt Associates.

Berkeley Planning Associates and Social Policy Research Associates. *Responses to Defense Cutbacks: Evaluation of the Defense Adjustment Demonstration—Summary of Findings.* U.S. Department of Labor, Employment and Training Administration, Research and Evaluation Report Series 97-A.

Bloom, Howard S. 1984. "Estimating the Effect of Job-Training Programs Using Longitudinal Data: Ashenfelter's Findings Reconsidered," *Journal of Human Resources* 19 (Fall): 544-556.

Bloom, Howard S., Larry L. Orr, George Cave, Stephen H. Bell, and Fred Doolittle. *The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months*. U.S. Department of Labor, Employment and Training Administration, Research and Evaluation Report Series 93-C.

Borus, Michael E. 1964. "A Benefit Cost Analysis of the Economic Effectiveness of Retraining the Unemployed," *Yale Economic Essays* 4 (Fall): 371-430.

Cave, George, and Fred Doolittle. 1991. *Assessing JOBSTART: Interim Impacts of Program for School Dropouts*. New York: Manpower Demonstration Research Corporation.

Conley, Ronald, Rima Azzam, and Arthur Mitchell. 1995. *Evaluation of Transition to Work Demonstration Projects Using a Natural Supports Model.* Washington, DC: Pelavin Research Institute.

Corson, Walter, Paul Decker, Shari Dunstan, and Stuart Kerachsky. 1992. *Pennsylvania Reemployment Bonus Demonstration: Final Report.* U.S. Department of Labor, Employment and Training Administration, Unemployment Insurance Occasional Paper 92-1.

Couch, Kenneth A. 1992. "New Evidence on the Long-Term Effects of Employment Training Programs." *Journal of Labor Economics.* 10(October): 380-388.

Executive Resource Associates, Inc. 1987. *Evaluation of the Job Corps' Pilot Project to Include 22- to 24-Year-Olds*. Arlington, VA.

Fraker, Thomas, and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources* 22 (2): 194-227.

Friedlander, Daniel and Philip K. Robins. 1992. "Estimating the Effects of Employment and Training Programs: An Assessment of Some Nonexperimental Techniques." Paper presented to the American Economic Association Annual Research Conference, Anaheim, CA.

Greenberg, David. 1997. "The Leisure Bias in Cost-Benefit Analyses of Employment and Training Programs." *Journal of Human Resources* 32 (Spring): 413-439.

Greenberg, David and Mark Shroder. 1997. *The Digest of Social Experiments: Second Edition.* Washington DC: The Urban Institute Press.

Grossman, Jean Baldwin and Cynthia L. Sipe. 1992. *Summer Training and Education Program (STEP): Report on Long-Term Impacts.* Philadelphia, PA: Public/Private Ventures.

Groves, Robert and Douglas Wissoker. 1999. *1997 NSAF Early Nonresponse Studies: National Survey of America's Families (NSAF) Methodology Report No. 7.* Washington, D.C.: The Urban Institute.

Gueron, Judith M. and Edward Pauly. 1991. *From Welfare to Work.* Washington, DC: Russell Sage Foundation.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1998. "The Economics and Econometrics of Active Labor Market Programs." Draft paper prepared for the Handbook of Labor Economics, Volume III, Orley Ashenfelter and David Card, editors.

Hollister, Robinson G. Jr., Peter Kemper, and Rebecca A. Maynard. 1984. *The National Supported Work Demonstration.* Madison, WI: University of Wisconsin Press.

James Bell Associates. 1998. *Employment and Training for America's Homeless: Final Report of the Job Training for the Homeless Demonstration.* U.S. Department of Labor, Employment and Training Administration, Research and Evaluation Report Series 98-A.

Johnson, Terry R., and Daniel H. Klepinger. "Experimental Evidence on Unemployment Insurance Work-Search Policies." *Journal of Human Resources.* 29 (3): 695-717.

Kornfeld, Robert and Howard S. Bloom. 1997. "Measuring Program Impacts on Earnings and Employment: Do UI Wage Reports from Employers Agree With Surveys of Individuals?" University of Chicago Joint Center for Poverty Research, Working Paper #1.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review* 76 (4): 604-620.

Leiter, Valerie, Michelle L. Wood, and Stephen H. Bell. 1997. "Case Management at Work for SSA Disability Beneficiaries: Process Results of the Project Network Return-to-Work Demonstration." *Social Security Bulletin* 60 (1): 29-48.

Quint, Janet C., Hans Bos, and Denise F. Plit. 1997. *New Chance: Final Findings on a Comprehensive Program for Disadvantaged Young Mothers and Their Children.* New York: Manpower Demonstration Research Corporation.

Rogers, Kevin. 1996. "Random Assignment Evaluations in Workforce Development Programs: Lessons from the Ongoing CET Replication Study." Presentation to the Sectoral Employment Development Learning Project Focus:HOPE, Detroit.

Rupp, Kalman, Stephen H. Bell, and Leo A. McManus. 1994. "Design of the Project NetWork Return-to-Work Experiment for Persons with Disabilities." *Social Security Bulletin.* 57 +(2): 3-20.

San Diego Consortium and Private Industry Council.1995. *San Diego Immigrant Training Demonstration Project: Final Evaluation Report, 1992 – 1994.* San Diego, CA.

Smith, Ralph Ely. 1970. "An Analysis of the Efficiency and Equity of Manpower Programs." Unpublished Ph.D. dissertation, Georgetown University.

U.S. Department of Labor. 1970. *The Influence of MDTA Training on Earnings.* Manpower Administration Evaluation Report No. 8.

Wholey, Joseph S. 1994. "Assessing the Feasibility and Likely Usefulness of Evaluation." *Handbook of Practical Program Evaluation*, Joseph S. Wholey, Harry P. Hatry, and Kathryn E. Newcomer, eds. San Francisco, CA: Jossey-Bass Publishers.

Appendix A

Methods for Estimating and Testing Demonstration Impacts

To calculate a demonstration's impact, researchers compare outcomes of demonstration participants with simulated outcomes representing a hypothetical world without the demonstration. To represent this hypothetical, or "counterfactual," world researchers simulate what outcomes for participants would have been like had there been no demonstration. Section 7 of the text discusses how "counterfactuals" can be created for employment and training outcome evaluations and provides a formula for measuring the impact of a demonstration on its average participant:

(1) *Estimated impact = average participant outcome - average counterfactual outcome.*

As noted in the text, the same formula is used for all outcome measures of interest—post-demonstration earnings, receipt of welfare benefits, etc. A more formal statement of this formula is given here. This appendix also explains how impact analysis can be extended to control for differences between the participant and counterfactual scenarios for factors unrelated to the demonstration and outlines the methods used to determine whether measured impacts are likely to have been real.

The Basic Impact Formula

Formula (1) above can be restated using mathematical notation. Using *Ypi* to represent the outcome level of participant *i* after participating in the demonstration and *Yci* to represent her/his corresponding outcome in the hypothetical counterfactual world,[103] the demonstration's impact on a given individual is:

$$Mi = Ypi - Yci,$$

where $i = 1, 2, \ldots, Np$ indexes individual demonstration participants, with *Np* participants in total. The impact of the demonstration on the average participant, *M*, can then be expressed as:

[103]For expository purposes, we assume here that actual outcome numbers have been established in the counterfactual world for ***each*** demonstration participant. As shown below, this is not quite how impacts are estimated in practice, but describing the procedure in these terms makes clearer what impact analysis represents.

(2) $M = \bar{Y}p - \bar{Y}c$.[104]

The same impact estimator can be obtained using ordinary least-squares (OLS) regression. To estimate *M* here, one runs an OLS regression on the equation

(3) $Yj = C + M \cdot Pj$,

where *Yj* is the outcome measure (e.g., annual earnings) for person *j* and *Pj* is a 0/1 variable indicating whether person *j* participated in the demonstration (*Pj* = 1) or not. Here

$j = 1, 2, \ldots, Np, Np+1, Np+2, \ldots, Np+Nc$,

is a consecutive index of the *Np* observed participant group outcomes followed by the *Nc* simulated counterfactual outcomes.[3] For participant group outcomes

$Pj = 1 \qquad (j = 1, 2, \ldots, Np)$

while for counterfactual outcomes

$Pj = 0 \qquad (j = Np+1, Np+2, \ldots, Np+Nc)$.

Here, the coefficient on *Pj*—*M*—represents the average effect of participation, while *C* gives the average outcome level in the counterfactual world.

Extensions to Control for Factors Other than the Demonstration

The analysis gains through the more complex formulation in equation (3) when analysts are concerned that factors other than the demonstration could also influence the outcome, *Y*. To accommodate this possibility, variables measuring these factors need to be added to the regression equation to control for any systematic differences in these factors between the participant group and the set of observations used to represent the counterfactual. As discussed in the text, adding more "explanatory variables" to the regression becomes essential when the counterfactual world is represented by observations from a "comparison group" of real-life individuals who did not participate in the demonstration. These individuals generally will differ from participants in a systematic way in the distribution of these additional "explanatory" variables. For example, comparison group members may be systematically older or more educated than demonstration participants, or have more recent work experience. Adding these variables to the right-hand-side of the regression equation in (3) helps remove their influence, which otherwise might distort the estimated measure of the demonstration's impact, *M*.

---

[104]A bar (-) above a variable indicates the average of the variable across all *Np* individuals.

With the addition of new right-hand-side variables, equation (3) becomes:

$$(4) \quad Yj = C + M \cdot Pj + Ba \cdot Xaj + Bb \cdot Xbj + \ldots + Bz \cdot Xzj, \qquad \text{where}$$

Xa through *Xz* represent potential non-demonstration influences on outcomes such as age or education. The estimated impact of the demonstration on the average participant is still *M*, but now *M* will not be distorted by the influence of *Xa, Xb, . . . , Xz* on *Y* as would have been the case in the earlier, simpler regression (or, equivalently, in the original impact formula (2)).

Deciding Which Estimated Impacts Are Real (Tests of Statistical Significance)

As with any measure derived from a sample of individuals, the demonstration impact figure *M* is a statistical estimate with a margin for error in either direction. Hence, regardless of the formula or equation used to produce it, *M* only ***approximates*** the true average impact of the demonstration. By taking account of this margin of error, researchers can determine whether the true average impact of the demonstration is greater than 0, less than 0, or too close to call—i.e., whether the measured impact is real (the "greater than 0" and "less than 0" cases) or not. Such calculations are often referred to as "tests of statistical significance" for the impact estimate.

In the simplest case depicted by equation (2), researchers use a statistical "t-test" to determine whether the difference in means (average outcomes) between the participant and comparison group samples reflects a real demonstration effect or is an artifact of the data resulting from the statistical uncertainty of the estimate. The t-test tries to prove with great certainty (either 90-, 95-, or 99-percent certainty) that the true average impact of the demonstration differs from 0 in one direction or another. Computationally, this test compares critical values from the distribution of the standard t-statistic with $Np + Nc - 1$ degrees of freedom to *M* divided by its standard error to test the null hypothesis that the true average impact equals 0 against the alternative hypothesis that it does not equal 0. This procedure requires a two-tailed t-test with a .10, .05, or .01 significance level, depending on the level of statistical certainty desired. Ninety-percent certainty—i.e., a significance level of .10—is often used to decide which measured effects of employment and training demonstrations are real, although 95-percent certainty continues to be an important alternative—and more demanding—standard of proof.

If the test of statistical significance rejects the null hypothesis in favor of the alternative, the average demonstration impact is almost certainly greater than 0 or less than 0 (depending on the sign of *M*). If the null hypothesis is not rejected, the true impact may still differ from 0—or it may not; the data simply are not strong enough to make the distinction. Given this uncertainty, an evaluation that fails to prove there are non-zero effects has not shown that effects are zero. Rather, the appropriate conclusion is that—despite the care taken in the evaluation and the amount of information used— the demonstration's ability to affect key outcomes for the average participant remains in doubt.

Appendix B

Converting a Demonstration's Future Benefits and Costs into "Present Value" Terms

This appendix explains how the benefits and costs of a demonstration—when occurring more than a year after demonstration start-up—are translated into "present value" terms in a benefit-cost analysis.

In general, translating to "present value" means discounting, or reducing, the monetary value of a demonstration benefit or cost that occurs after the point in time in which demonstration impacts are to be valued. The need to discount stems from the fact that--from today's perspective--longer-term benefits and costs make less of a difference to demonstration participants and to society than immediate impacts, due to a phenomenon known as the "social rate of time preference." The social rate of time preference, like an individual's rate of time preference, grows out of the desire for economic gains to happen sooner rather than later. This preference for the present arises not from short-sightedness or over indulgence, but from the economic reality that $\$X$-worth of resources today can be invested in productive uses for 12 months and become ***more*** than $\$X$-worth of resources in a year. For this reason, if society—or an individual—is to receive $\$X$, it would rather have that amount arrive today than a year from now.

This concept can be illustrated best by an "everyday" example. Most people would prefer receiving a surprise deposit of $\$X$ in their bank accounts this year rather than next year. By next year, today's gift—if left in the bank—will be worth $\$(X+R)$ due to 12 months of accumulated interest, $\$R$, paid on the account. As a result, R becomes the appropriate rate of time preference for investors. Similarly, the social rate of time preference—or the "social discount rate" as it is sometimes known—equals the current rate of return on very low-risk investments. In benefit-cost analyses this rate, R, is usually pegged to the interest rate on federal government Treasury bills, roughly 4 percent a year plus the rate of inflation.

All future benefits and costs of an employment and training demonstration are then discounted to "present value" terms by dividing them by a multiple of $1 + R$:

for year 2 effects, divide by $(1+R)$,

for year 3 effects, divide by $(1+R)(1+R)$,

for year 4 effects, divide by $(1+R)(1+R)(1+R)$,

and so on. This adjustment correctly captures the value of a future demonstration benefit or cost as of year 1, the usual point of valuation in a benefit-cost study. This result can be demonstrated by considering a further example. Suppose a demonstration or pilot project produces a social benefit of $\$X$ realized in year 3. By the above formulas, this benefit is to be considered worth

$$\$X / (1+R)(1+R)$$

as of year 1. To see why, imagine that $\$X/(1+R)(1+R)$ in cash were available in year 1 and invested for two years at an interest rate of 100R. Such an investment would grow to a total of $\$X$ dollars by year 3:

Year 1 amount = $\$X / (1+R)(1+R)$

Year 2 amount = Year 1 amount plus interest at 100R percent
$= \$X / (1+R)(1+R) + [\$X/(1+R)(1+R)]R$
$= [1+R] [\$X/(1+R)(1+R)]$
$= \$X / (1+R)$

Year 3 amount = Year 2 amount plus interest at 100R percent
$= \$X / (1+R) + [\$X/(1+R)] R$
$= [1+R] [\$X/(1+R)]$
$= \$X .$

U.S. Department of Labor
Employment and Training Administration
Washington, D.C. 20210

Official Business
Penalty for Private Use, $300

Presorted Standard
Postage and Fees Paid
U.S. Department of Labor
Permit No. G-755