REVIEW OF ALTERNATIVE METHODOLOGIES FOR EMPLOYMENT AND TRAINING RESEARCH

Stephen Bell

Urban Institute

This report has been funded, either wholly or in part, with Federal funds from the U.S. Department of Labor, Employment and Training Administration under Contract Number K-6826-8-00-80-30. The contents of this publication do not necessarily reflect the views or policies of the Department of Labor, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

Employment and Training Administration Occasional Paper 2003-11

REVIEW OF ALTERNATIVE METHODOLOGIES FOR EMPLOYMENT AND TRAINING RESEARCH

Research and evaluation needs to be part of the justification, oversight, reappraisal, and ongoing improvement of every employment and training program. The U.S. Department of Labor's Employment and Training Administration (ETA) recognizes this and seeks to identify the best methodologies to use in looking at varied policy questions surrounding its programs. Guidance on research methods should also prove valuable to state and local decision-makers as they implement federally directed workforce investment activities and look for information on their results.

This report addresses these issues, based on the independent assessment of an expert evaluator of labor market policies and interventions. The report considers the best ways to measure national, state, and local programs' impacts on various client populations and the economy as a whole—and to judge their cost-effectiveness as taxpayer investments when both client and broader social benefits are considered. Though it pertains to programs and decision-makers at all levels of government, the discussion adopts the terminology of and emphasizes recommendations for the national viewpoint.

Three main messages stand out:

- Information on policies and programs taken from research and evaluation must be rigorous and thorough;

- Different methods work best to gather information of this sort in different policy and program settings; and

- All the research techniques involved pose large technical challenges and utilize substantial amounts of data, necessitating a major, sustained commitment of resources by the government to be successful in improving the nation's workforce investment system.

The discussion developing these themes covers several topics. Section 1 discusses the role of research in policymaking and the types of information gathering and evaluation activity to be considered. The best way to do research in different workforce investment program and policy areas comes next, in Section 2. While the ideal of a randomized experiment to compare matching groups of program participants and nonparticipants and measure program effects may sometimes be unattainable, it can and should always serve as the goal ahead of next-best evaluation strategies.

Section 3 applies this standard to the range of ETA-sponsored (and state and locally administered) interventions. It shows how random assignment experiments can be conducted to rigorously measure impacts in almost all cases if the determination is to get the best information to guide policy—even when barriers to the approach appear to arise from the very nature of the intervention involved. This section also explores what can be

done to provide scientific evidence on program effectiveness even in light of these challenges, and includes several strong recommendations for top-line, comprehensive evaluation of the programs and policies ETA and affiliated state and local agencies oversee.

Recommendations continue in Section 4 in two other important areas of strategic planning: making good evaluation choices when the best choice has not been achieved for some reason, and supporting improvements to existing tools for measuring the impacts, costs, and benefits of employment interventions in second-best situations in the future.

A. THE ROLE OF RESEARCH AND EVALUATION

When government agencies want to undertake activities to improve the well-being of American citizens—be it in the areas of employment, health care, national security, or any other—what kind of information do they need? That is the function of "research and evaluation" in the public sector: to give policy makers and program managers in government the information they need to effectively pursue the goals of their agencies and improve the lives of their customer populations. It is a priority relevant to other nations besides the United States, of course, so attention needs to be paid to the practices of other Western democracies in this regard.

Types and Value of Policy Information

We can begin thinking about employment and training research needs by identifying the types of information that might help government set and administer public policies focused on improved labor market functioning and better outcomes for American workers and their families. Seven broad types of research come to mind, described here from the national perspective but relevant as well to local and state decision-making:

A. Understanding of the needs and behaviors of the customers ETA (or other state/local agencies) is to serve, both employers and employees.

B. Knowledge of how the core U.S. system for social organization—the market economy—responds to the needs of workers and businesses, creating opportunities for gain but also potentially leaving problems.

C. Rationales for how government actions, if ideally realized, would improve the free market outcomes of workers and businesses by addressing specific problems.

D. Information on how policies and programs adopted in pursuit of these goals actually function—what happens in their administration, whom specific workforce investment initiatives actually "touch" in the labor market, either on the supply side (workers) or demand side (employers), what actions those private parties take once impacted by a policy or program.

E. Measures of "the difference" any particular policy or program makes, once it is implemented and plays out through the behaviors of individuals and assimilating, equilibrating mechanisms of the marketplace…information on the "value added" or impact of what ETA (or other workforce investment agency) has done in instigating a particular policy or program, held up against the standard of where things would have come out absent the intervention.

F. Accounting of gains against losses among the various impacts that do occur, for different constituencies (the who "wins," and who "loses" question) and for society as a whole, in what is known as benefit-cost analysis.

G. Indications of how existing policies or programs could be made to work *better* or more consistently in all parts of the nation, to move society closer to the goals sought when the activity was undertaken.

These seven types of information—and the investigative activities intended to produce them (the research and evaluation methods to be discussed in this chapter)—might be re-titled in shorthand as:

A. Basic Research
B. Market Analysis
C. Rationale for Intervention
D. Implementation Analysis
E. Impact Evaluation
F. Benefit-Cost Analysis
G. Best-Practice Improvements

The value of each of these research types for better public policy is summarized in Table 4.

Integrated Policy Evaluations as the Key to Accountability and Program Success

Once programs or policies have been legislated and their implementation initiated through regulation, the government's need for information types D, E, and F becomes paramount. In particular, *knowledge of what happens in carrying out a policy's intent—implementation analysis—and the measured consequences of those actions—the policy's impact on workers and firms and a benefit-cost assessment of its social value—provide the only sound basis for justifying a policy intervention using objective evidence.* This broadly shared principle (among labor economists, program evaluation specialists, and—increasingly—the Congress and executive agency leaders) lies at the heart of every point made in this chapter.

Table 4: The Value of Different Types of Program/Policy Information

Source/Type of Information	Value in Furthering Public Policy Improvements
A. Basic Research	Finding places where better outcomes have social priority Making better assumptions about household and business reactions to potential policies/programs
B. Market Analysis	Foreseeing the amalgam of individual reactions to policies and programs as the aggregate response by society Recognizing non-obvious connections between disparate behaviors, particularly those with the potential to affect everyone's outcomes through "systems equilibrium"
C. Rationale for Intervention	Providing the logic basis for specific government responses to identified needs Identifying actions that can improve on market outcomes given individuals' behavioral responses and social equilibrium effects Setting standards of achievement for judging a policy/program's success based on measurable indicators
D. Implementation Analysis	Checking if the mechanisms by which an intervention is to accomplish its goal are in fact operative in practice Understanding the nature of—and pointing to possible solutions to—any breakdowns in those mechanisms

[continued on next page…]

Table 4: The Value of Different Types of Program/Policy Information (continued)

Source/Type of Information	Value in Furthering Public Policy Improvements
E. Impact Evaluation	Measuring the contribution to social improvement of policies and programs as actually implemented, compared to where society would be without them
F. Benefit-Cost Analysis	Gauging contributions (benefits) against what society had to give up to get them when choosing to implement certain policy/program actions (costs) Deciding if "return on investment" is high enough to justify the particular use of government resources being studied, compared with other public sector investments or returning resources to taxpayers for private use Understanding which segments of society come out ahead or behind because of the policy/program, and by how much in economic terms
G. Best-Practice Improvements	Upgrading policy accomplishments within existing program and policy rules Suggesting different program or policy rules that might increase social returns (including "differential impact analysis" of competing policy models)

The "Three-Legged Stool" of Comprehensive Program Evaluation and Research

The emphasis in most of this chapter will be placed on information types E and F—impact evaluation and benefit-cost analysis. But effective integration of these elements with information type D—implementation research—is equally vital, and provides the highest overall payoff from program evaluation. Only with all three legs of the "three-legged stool" of comprehensive program evaluation in place can government agencies (i) know what they've done at the state and local level, (ii) determine the consequences of those actions, and (iii) decide which consequences are good and bad from the standpoint of different segments of society. This broad concept of the type of research and evaluation needed at ETA and other agencies deserves more discussion before zeroing in on the impact and benefit-cost "legs" of the stool.

The three-legged approach to measuring program accomplishments dominates all others *precisely because it does not leave out any of the three components*. It recognizes how one element helps another, and crafts each component to take advantage of these cross-cutting synergies. Findings from a descriptive analysis of program implementation gain value when evidence of program impacts is added. Similarly, to use impact findings well, the question of "impact of *what?*" needs to be answered. While one can start to answer this question based on the legislated parameters of an intervention, and get closer through examination of regulatory guidelines for implementation, ultimately there is no substitute for direct empirical measurement of "what happens" in local settings once the regulations go forward. This is the role of process analysis, and—while not examined in detail here—it is one of tremendous policy importance.

Why Process Analysis Is Essential to Evaluating Program Impacts

Good implementation analysis done as part of a larger, integrated evaluation pays attention not just to the intervention being evaluated but also to the *alternatives to that particular program or service for the customers who use it.* As will be discussed in greater depth later, impact analyses look at how results differ with versus without a particular intervention in place. To makes sense of this information, policy makers must comprehend the mechanisms by which results come about under *both* scenarios. What a new program or policy *replaces*—the market- and/or policy-driven mechanisms that would exist without it—is as important in determining impacts as its own characteristics. Implementation analysis that recognizes this dichotomy is much more valuable than process research that covers only one side of the equation…but will only be undertaken if a tripartite evaluation approach is adopted from the beginning.

The synergies work in the other direction as well. Looking for results in the right places is crucial to effective impact and benefit-cost analyses. Close scrutiny of the ways in which policies or programs are implemented can suggest new areas where they may have an effect. An impact analysis *not* informed by these process results might miss something important. It might also produce a skewed impression of the overall social value of the intervention when extended to benefit-cost analysis. Yes, one could gain good information on impacts, benefits, and costs without the process piece—but only by

being fortunate, rather than informed, on where those benefits and costs are likely to arise. No effort to assess program effectiveness should run this risk: reliable impact and benefit-cost analyses can only be achieved when the scope for possible program effects is reliably informed by thorough implementation analysis.

Moving toward Better Policies through Process-Informed Insight

A final contribution of descriptive, process analysis concerns its potential to *improve how programs and policies work by changing the ways they are implemented.* Federal agencies responsible for program success need good descriptive information on topic D above—the character of the interventions they are already administering—as well as topic G, how those interventions might be improved. Insights from looking at state and local program operations can often suggest areas where administrative changes might be helpful, though of themselves can never *prove* that this is the case absent additional impact research. Still, the basic principle is affirmed: process information provides a crucial complement to impact and benefit-cost analysis, and vice versa, *when all components are considered indispensable parts of a single integrated program evaluation strategy.*

A last point worth noting concerns the obligation of government agencies to undertake such studies. Unlike other informational types listed in Table 4—such as types A, B, and C on basic research, market analysis, rationales for intervention—implementation and benefit-cost research will rarely be undertaken by academics for purely scholarly reasons. While some examination of policy impacts is common in academic research, it is rarely as focused or as well-informed by other research components as it needs to be to guide real-world policy decisions. Both in its constituent parts, and especially in its integrated whole, multi-faceted evaluation of the nation's primary workforce investment policies must be undertaken by ETA and its state/local partners—and backed through Congress with major fiscal resources—if it is to take place at all.

Emphasis on Impact Evaluation and Benefit-Cost Analysis

Notwithstanding the tremendous value of evaluations that produce policy information in several of areas at once, some narrowing of focus for this chapter is essential. This step is taken even noting that ETA has an interest in—and, indeed, sponsors—research in all seven of the above-listed areas. Within this framework, several factors lead us to concentrate on information types D and E produced by impact evaluations and benefit-cost analyses. Most important is a desire to obtain *evidence of efficacy,* and accountability for results, for *ETA-sponsored policies and programs already in place* and administered by a diversity of state and local agencies.

Policy Development as a Secondary Goal

The emphasis on efficacy of existing policies diminishes interest in learning more about workers, firms, and market outcomes through basic research. This downgrades the role of information types A and B for present purposes.

The emphasis on existing interventions also reduces the need to formulate rationales for new government policies, as in information type C. Information types A, B, and C are most important *ahead* of implementing particular policies and programs. While quite valuable in leading ETA and other policy makers to attempt interventions with the most promise, they tell us nothing per se about the results then achieved—nothing on the efficacy agenda that is to be the focal point for this chapter.

Focusing on Accountability for Results

Information type D, drawn from research on the mechanics of policy implementation, comes closer to addressing questions of efficacy. It documents the *operational* achievements and failings of interventions actually taking place. These are "results" from the point of view of program managers and those who oversee implementation at the federal level, but not from the perspective of society in general…which, after all, initiated a government role to change *outcomes* for workers and employers, not simply to see its own programmatic activities take shape in the ways intended.

For society, the only "results" that count are the improved lives of workers and the greater business prosperity that result from the activities government executes. Accountability and measurement of results in this sense comes from impact evaluation, as information type E.

The next step—and the rest of accountability for results—comes from comparing the benefits of policies and programs to their costs through information type F, benefit-cost analysis. In many ways, this is an extension of impact evaluation, looking at impacts in *all* areas of social consequence and in a holistic way. Though not desired for its own sake, the cost—in tax dollars—of an agency like ETA doing something rather than nothing is a program "impact" every bit as much as the gains that action may produce for participants in the labor market.

Benefit-cost analysis makes a point of not leaving out any of the consequences of a policy or program, and of looking at them as a group to see net effects on all of society and its constituent parts (e.g., workers, business firms, taxpayers). But it is still a way—indeed, the most comprehensive way we know of—to see whether an existing government program or policy achieves its purpose for those elements of society for whom it was conceived and undertaken.

Suggestions for Improvement Welcome, but Evidence of Existing Results Paramount

The final type of information valuable to the policy process is "best practice" research, type G. This is essentially the after-the-fact version of the earlier, more forward-looking "intervention rationale" exercise (information type C). It examines essentially the same question—what might be done in the future to improve on what we have now—but does so based on experience with actual policy attempts rather than analyses and expectations for intervention approaches not yet tried.

Like information types A through D, best-practice research is not intended to give a read-out or "report card" on the accomplishments of existing policies as currently implemented—the purview of impact and benefit-cost analysis. Instead, it is looking for something that might be done differently in the future. While of great value, the type of information it provides cannot tell us whether we have succeeded and should feel satisfied *with policies and programs already in place*, or indicate whether to keep and/or expand these activities or to move away from them.

The research and evaluation tools that provide information of this latter type—types E and F on the list, information on which existing programs and policies produce better labor market results for businesses and workers, and social benefits that exceed their costs—are the focus of this chapter. We will talk from here on only about the methodologies ETA or local and state agencies might use to obtain information in these areas, concerning program impacts and benefits and costs. This information is obtained through a range of investigative activities alluded to by many names: evaluations, demonstrations, research, pilot studies, and others. Before moving on to describe how these activities can strengthen the Department's understanding of its programs and policies, it is worth clarifying what each of these terms means and how they relate to the impact and benefit-cost focus of the chapter.

Distinctions Among Demonstrations, Pilots, and Evaluations…and Their Value to Government Agencies

In earlier work supporting ETA's desire to improve the information available to employment policy decision makers, Bell[1] depicts pilots, demonstrations, and evaluations of ongoing programs as similar but distinct types of activities with different informational pay-offs. He first defines pilots and demonstrations as "special program or policy initiatives undertaken on a small scale to test the feasibility and effectiveness of a new…or improved policy idea" (p. 9). Demonstrations differ from pilots in his assessment by attempting to demonstrate, through the use of extensive data and formal research, that the intervention achieved its policy goals—not simply that it could be implemented in the real world (the primary goal of a "pilot").

Both pilots and demonstrations differ from "studies of ongoing national programs [and]…newly enacted national policies" (p. 10). These we might call "evaluations", though Bell in his earlier work defines evaluation methods as "research tools for compiling information and drawing conclusions about specific government programs or policies" (p. 9) applicable to all three types of intervention. He also refers to a fourth category, "research on labor market issues that provide vital background information without focusing on the consequences of specific program or policy approaches" (p. 10).

The term "research"—if it is to connote a specialized genre of labor policy-related investigation at all—might usefully be equated to the first two information types in Table 4, basic research and market analysis. These do not focus on (but can nonetheless inform thinking about) government interventions, but attempt instead to understand how private

[1] Bell (2001).

actors behave and interact in the labor market and the good and bad outcomes that may result from a social perspective. This body of analysis makes up the bulk of the writings of academic labor economists and social scientists concerned with family and work issues, employer skill needs and labor demand, the institutional/legal/historical aspects of labor unions, issues of workplace organization, and productivity/national economic output.

Based on this taxonomy of phrases, the proper label for the type of investigation and data analysis explored in the current chapter is "evaluation of ongoing ETA programs and policies", or evaluation for short. Though Bell's *Guide for Practitioners* revolves around a different type of investigation—examination of pilots and demonstrations—many of the methods highlighted there, and their strengths and weaknesses, feature strongly in the current chapter. This is because many evaluation methods apply to any government intervention, regardless of its origin or intent. However, there are important if subtle differences, particularly concerning the feasibility of random assignment as a way to measure the impact of ongoing programs rather than new pilot or demonstration initiatives.

The current essay also contrasts with Bell's earlier work by providing far less detail on the mechanics of using any particular evaluation tool. The goal here is to recommend broad evaluation strategies and priorities to ETA leadership and other state and local authorities in the workforce investment realm (as well as to evaluation researchers). It does not attempt, as did the earlier work, to instruct evaluation practitioners in the "nuts and bolts" of doing impact and benefit-cost analysis. Nor does it as extensively explore the advantages of getting involved in in-depth research projects from the standpoint of state and local—as opposed to national—agencies.[2] Even so, the reasons provided here for why evaluations are invaluable for national policy-making when done rigorously and comprehensively apply equally well to more local—and, in many ways, more programmatically directly involved—units of government.

Framework for Gaining—and Being Guided by—Information on Policy

As the final step in setting the stage for recommendations on ETA's policy evaluation agenda, we discuss the dynamics of policy decision-making in a world where research information plays a role, and the intertwined dynamics of evaluation planning and support needed to "feed" good program information into the system. Though somewhat abstract at times, this exploration seeks to ground later recommendations on evaluation strategies in a realistic appraisal of how evaluation findings might be used by policy makers at the state, local, and national levels.

The first thing to be said, of course, is that they might not be used at all! Many considerations influence policy choices in a pluralistic government, and evidence on "what works" may be well down the list even when available and considered conclusive. So we begin by noting that research—the systematic investigation of what works and what does not work among the range of options available to government organizations—

[2] See, for example, Bell (2001), 26-27.

is only worthwhile when policy choices are to be based on explicit, objective criteria that can be informed by data. If there is no agreement that policies should be guided, at least in part, by how their characteristics compare to agreed standards of what's desirable, evidence on those characteristics—including evidence on impacts, benefits, and costs—has little role to play.

Recent OMB Guidelines: Strengths

From time to time, the federal government affirms that evidence does in fact matter to what it wants to accomplish in American society.[3] This occurs most forcefully when it states the criteria to be used by various domestic and international departments to judge the value and appropriateness of their various projects and policies. This has taken place recently through the White House Executive Office of Management and Budget with the issuance of the Program Assessment Rating Tool (PART) to the executive departments in October 2002.

PART states what is needed to justify policies and programs as worthy of taxpayer support. It covers a wide range of indicators. The indicators that most nearly approximate the concept of program impacts, benefits, and costs on which the chapter focuses require that, to be justified, a policy have:

- A clearly identifiable purpose that addresses explicit needs and problems in society;
- An intervention design expected to make a significant *impact* in reducing the problem or need involved [emphasis added];
- Outcomes measures that can be tracked over time to see if progress is being made in ameliorating the problem; and
- Actual evidence of measured progress toward the outcome goal.

The last two of these standards come closest to seeking to measure impacts based on the outcomes of those affected by the intervention. But the PART also establishes other quite different grounds for judging the success and justification of a policy, including whether the intervention is: "optimally designed" to address the problem or need; supported in its goals by all partners to its implementation (e.g., federal grantees); coordinated well with related programs that have similar goals; using its funds entirely for its intended purpose; and performing on par or better than other programs with similar purposes.

[3] Principal reference—in this and subsequent discussions—to national organizations such as OMB and ETA and to the federal perspective is not meant to denigrate the importance of state and local considerations. Rather, federal terminology is adopted reflective of federal leadership in many areas of evaluation research and also for expediency's sake when making points that apply as well to other levels of government.

Recent OMB Guidelines: Omissions and Weaknesses

Other criterion-based standards for making policy choices might also be utilized beyond those in the OMB tool. These would base decisions on whether to support an intervention on factors such as:

- Congruence of implementation with the actions intended to be undertaken in the policy's design;
- "Through-put" of total people or firms served per unit of time;
- Progress toward identified goals per dollar of spending;
- Fiscal integrity in accounting for all funds expended;
- Level of customer satisfaction among those it is intended to help; and
- Distribution of benefits from the intervention across different segments of society.

Some of these criteria, as well as certain of OMB's PART standards, can be formulated in net impact terms consistent with the focus of this chapter. In particular, progress toward goals and the distribution of benefits across constituencies could be measured relative to what would have happened *absent the intervention*. Similarly "measured progress toward the outcome goal" in the OMB assessment tool could be defined in terms of *net outcomes*—improvements compared to what those same outcomes would have been without the policy—rather than *gross outcomes* or *outcome changes over time*.

This is a crucial distinction. Gross outcome assessment effectively credits the intervention for all that occurs in the measured domain, both good and bad. In contrast, keying off outcome *changes* over time implicitly assumes the world would have gone on unchanged if the intervention were not implemented. Neither of these perspectives is as attractive for judging the worth of a program as measuring progress against a world in which the intervention does not exist but where everything else—including (positive or negative) outcomes and trends in outcomes unrelated to the intervention—is the same.

The PART system apparently despairs of being able to apply this net impact standard, or else finds that standard less attractive, since it explicitly frames the question of progress as one of change over time: "Does the program demonstrate improved efficiencies and cost-effectiveness in achieving program goals each year?" In putting forth this standard, no mention is made that year-to-year trends might be adjusted for external constraints on policy achievements, such as changes in labor market conditions or increased demographic diversity of the clientele served. Though this standard more readily accomplished from an information standpoint—trend lines in outcomes and spending can

always be tracked—it is not the preferred way among employment policy evaluators for judging a program's worth.[4]

As stated above, the preferred method for gauging program success is net impact analysis of the consequences for different consumer groups and the translation of those measures into gains and losses for different segments of society through benefit-cost analysis.

Assumptions Versus Information

The distinction between movement toward an outcome goal over time and the true impact of an intervention relative to what outcomes would have been without it illustrates the core dilemma of policy formulation based on objective standards and "knowledge" of performance. The "knowledge" must come from somewhere, either real information or through preconceived expectations and assumptions.

Impact analyses try to determine what would have happened absent an intervention using real information—ideally, through the observation of a set of people or firms identical to those who participate in the intervention but not exposed to the "treatment". Monitoring of outcomes, either in absolute terms or through changes over time, establishes (whether consciously or not) the contribution of the intervention *by assumption alone*, using one of the assumptions noted previously: nothing good (or bad) would have occurred without the policy, or nothing different from last year would have occurred without the policy.

This illustrates a universal principal of "knowledge-based" policy choice: to apply it, one must establish a "knowledge" base either by assumption or through the gathering of information about the world—that is, by doing evaluation research as described earlier. In this sense, *research replaces assumptions with measurement as the basis for policy choice*. If this seems an obvious or perhaps belabored point, its importance in deciding how and when to fund evaluation research justifies the emphasis.

Pressures to Do Less than the Best

As ETA or any government agency plans its strategic agenda for program and policy evaluation over a five-year horizon, it will be torn between pressures to do (a) *no* formal impact evaluation of its core programs, (b) simple and easily-obtainable impact evaluations, or (c) difficult but rigorous, definitive impact evaluations. When is it worth it to move from (a) to (b), and from (b) to (c)?

There are no hard and fast rules if considerations other than practicing the best possible science factor into the decision…considerations such as finite budgets for research, finite patience in the political process for obtaining policy guidance, and substantial

[4] Bell (2001), 22, states the preference of many employment and training evaluators for net impact analysis above other policy assessment tools as follows: "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."

institutional and constituency resistance to ambitious research in many instances. But some rough guidelines can be evinced from the assertion that "research replaces assumptions with measurement" when making policy choices.

Some Rules of Thumb

If measurement is to replace assumptions as the "knowledge base" of policy decision-making, planners should consider two guidelines when making research investment decisions:

- Some measurement is always better than none, even relatively weak measurement, and

- The more rigor one seeks in measuring policy effects through impact evaluation, the more one substitutes information for assumptions.

The first of these propositions presumes that assumptions can always be reasserted ahead of measured evidence when the basis for measurement appear too fragile to support a policy decision. This it true in principle, though perhaps not often in practice. The principal is that evaluation findings *do not have to be to be factored into a policy decision simply because they exist*; they can be used where judged sound and useful and ignored otherwise. But in practice evaluation results almost always do get used, or at least promoted by those whose policy position they support—even when their reliability is questionable. The support they offer for one position or another, or their ability to satisfy an OMB or other agency reporting requirement, may outflank their scientific merit. The first rule of thumb, therefore, has to be taken with a grain of salt and thus has some risk attached as the basis for planning.

The second point stresses a more universal point: the ability to replace fragile assumptions with good information grows as the commitment to rigorous impact evaluation rises. This is true even when evaluation methods and the information produced remain less than perfect. Better methods for gathering information have to help, since there is no way around relying on either assumptions or information when making policy. This point allows us to turn the conventional planning question, "How much evaluation do I want to undertake and at what level of rigor?" around into something more telling: "How much do I want to have to rely on assumptions and assertions?… because that's where less evaluation leaves me."

Tensions Between the "Perfect" and the Possible

A final point on these theorems, before proceeding to a more concrete discussion of the evaluation strategy ETA might adopt for measuring the effectiveness of its major programs in the years ahead. Ironically, the two guidelines above at times work at cross-purposes in guiding the policy and research process, with the implications of one acting as an impediment to the other. For example, champions of the most rigorous possible research sometimes distain the "something is better than nothing" philosophy of the first maxim. Such a lenient view will, to their thinking, become an excuse for doing too few

truly rigorous state-of-the-art impact evaluations…if other weaker evidence is admissible. From this perspective, refusing to "admit to court" lower-quality information will hasten the adoption of high research standards in all policy realms.

Other onlookers take a more short-run, pragmatic view of the matter in arguing that allegiance to the second maxim stands in the way of being sensible about the first. They fear that "the perfect (rule 2) will become the enemy of the possible (rule 1)." In this conception, the determination of highly principled evaluators and evaluation sponsors to give policymakers only the best possible research evidence acts for the time being to starve the policy process of what information is available or can quickly be assembled. While less definitive, that information—according to the first maxim—is better than no information.

This conundrum cannot really be resolved. Both sides are right if the dynamics of setting standards and using evaluation research are as they posit. And no doubt both sets of dynamics apply to different actors in the policy and research realm: some government sponsors of policy assessment are too ready to settle for weak evaluation data, and might be moved to firm up their commitment to rigorous research more quickly if denied that "easy out." Others would make effective and appropriately circumspect use of existing, weaker evidence if given it more often, without losing their commitment to do better as often as possible.

The historical trend, particularly in employment and training evaluation—and often led by ETA—is upward with regard to expectations and implementation of rigor in evaluation research, notwithstanding the cross-currents just cited. This suggests that it remains appropriate to embrace both principles, aspiring to the best research *possible* with each new evaluation project while using the best research *available* when each new policy decision has to be taken.

It also suggests that extensive Congressional support for DOL-sponsored evaluation in the workforce investment arena will be needed to continue these standards and, over time, to obtain better policy information than previously has been available to judge program success. As noted earlier, this means highly reliable information on program impacts, benefits, and costs. The remainder of the chapter lays out a strategy for obtaining that information. It sets the bar high for funding comprehensive assessment of the nation's workforce investment system on an ongoing basis, arguing that any such expenditure constitutes money well spent on better policies and increased taxpayer accountability for workforce investment activities.

B. THE IDEAL APPROACH TO IMPACT AND BENEFIT-COST EVALUATION

Even the very best evaluations of program impacts and benefits and costs involve a good deal of uncertainty, making them far from infallible. The reasons for this will become apparent shortly regarding the specific scientific challenges faced by evaluators trying to measure program effects and the tools at their disposal. Perhaps, though, the fallibility of research evidence is easy to accept on face value when looking at policy interventions—what government agencies do is complex, depends heavily on administrative decisions and staff actions during implementation, moves toward its objectives only if the public responds in a certain way, and may lack clarity of intent or viability of approach to begin with. In such circumstances, one can hardly be surprised if researchers looking in from the outside—even when equipped with the most scientifically sophisticated and appropriate research tools—might often reach an understanding of what a policy or program has accomplished that is incomplete or off-base.

If the best that can be done remains dicey, it behooves government agencies—including ETA—to avoid doing less than the best whenever possible. Or, to frame the issue the way it most often arises, *the pressures to adopt a research approach that falls short of maximum reliability must be resisted at every turn as new studies are initiated.* This theme will return over and over in the chapter as the common pressures for compromise—including but not confined to resource limitations—are identified and precise methodological responses are urged on the Department. Right now, it provides the motivation for the subject matter of the next two sections: defining the ideal system of program/policy evaluation at ETA.

This is what—in the view of one evaluator who has focused a career on identifying the best means of measuring impacts, benefits, and costs for labor market interventions within U.S. institutional and political constraints—Congress, the Administration, and the agency itself *should do* to assure that ETA's programs and policies best serve the American people by advancing society toward the stated objectives of each intervention. The discussion first describes the one research technique thought capable of supporting a rigorous, comprehensive system of assessing program effectiveness—random assignment experiments—and then outlines a program of impact and benefit-cost research that can be achieved using this tool if sufficient will and funds come forth. It presents the best-case vision as a point of reference, a reference vital for both:

- Seeking the best means of policy and program evaluation in all cases, assuming we take seriously the desire to use scientifically-valid information on program effectiveness to make policy decisions; and

- Understanding what is sacrificed any time a less-than-best-available methodology for impact and benefit-cost evaluation is adopted.

A later section, the last in this chapter, looks at second-best options when randomization is either not attempted or not accomplished.

The Central Evaluation Challenge

As discussed in Section 1, impact evaluation and benefit-cost analysis form parts of a single larger enterprise, measuring how government actions change outcomes for individuals and society as a whole. Impact evaluations begin this assessment by measuring changes attributable to the action (the program or policy's "value added") occurring for the groups in society the intervention was intended to assist, perhaps older workers or small-business employers in the case of ETA policies. Benefit-cost analysis extends this assessment to consequences for other groups (e.g., taxpayers) and changes not beneficial to society but intrinsically a part of the overall action taken (e.g., greater private and government expenditures on transportation as more older adults stay in work). But always, the goal is to measure *the difference* a program or policy makes relative to what would have happened—either better or worse—without it.

The "What If Not" Question

In checking its existing policies and programs for favorable impacts and overall cost-effectiveness, ETA—like all government agencies—faces the central evaluation challenge: judging what the world would look like if the particular policy or program of interest *were not there*. That this is the heart of holding governments accountable for the results of their policies may seem counterintuitive or surprising…that accountability for what an agency does depends crucially on a hypothetical situation in which *it is not doing that thing*. But it is vital that this point be understood and embraced if one expects to make sense of the rest of the discussion.

Imagine a government policy that can speak for itself, challenged at a public forum to justify its existence to skeptical taxpayers. The taxpayers are doubtful that they need this policy, wondering aloud whether it has value. The policy reacts like many people do when feeling under-appreciated, by challenging others to think about the world without them. A particularly articulate policy might say "Consider what your lives are like with me in place. Now compare that to what you'd have without me. [pause] That's my value to you."

Formalizing the With/Without Comparison

This is impact and benefit-cost analysis in a nutshell: "With me, you get A; without me you get B." Using mathematical notation common in the evaluation literature the situation of "what lives are like" *with* a policy in place is Yp, where Yp represents an aspect of labor market outcomes of importance to workers or businesses that the policy is intended to change. The subscript p stands for the world with the policy or program of interest in place. For example, Yp could be the long-term earnings of trade-dislocated workers with Trade Adjustment Assistance services in place.

We want to contrast this outcome with what would have happened *absent those services*; call that outcome Yh to denote the hypothetical nature of this result. Our talkative policy might then say:

> "Consider Yp—what you have with me—compared to Yh—where you'd be without me. My value to you is the difference between these two situations, Yp – Yh."

This is precisely how impact evaluators define the measure of central importance, Yp – Yh: how actual outcomes with a policy in place differ from their hypothetical level without the policy.

Knowing Yp, "what lives are like" *with* existing programs and policies in place, is easy since this is the real world we live in and observe regularly. Figuring out Yh, "where you'd be without me," requires a leap of imagination or, in keeping with earlier emphases, actual information rather than assumptions about how things would turn out for key constituencies were they operating in a *different* world than the one actually in place. This desire—this inescapable need, if government programs are to be judged by the difference they make—to measure what researchers call a "counterfactual" world that would exist without a particular policy or program in place dominates all methodological thinking on impact and benefit-cost analysis.

Meeting the Central Evaluation Challenge: The "Gold Standard" of an Almost-Perfect Counterfactual

Randomized experiments provide the best counterfactuals. These research studies deliberately exclude some members of the group of people or employers a program is intended to help in order to create and observe a "world without the program." They pick the excluded cases purely by chance, through a lottery-like process that randomly divides the consumer population into two groups:

- A "treatment group," assigned to receive the program or policy of interest; and
- A "control group", excluded from the program or policy for research purposes.

Traits of a Strong Counterfactual

Because it is chosen by chance from the relevant target population and then kept out of the intervention, an experimental control group meets three critical conditions for a successful counterfactual (i.e., for accurately representing the world without the policy/program):

1. They are not subject to the intervention, and thus experience no effects from it.
2. They otherwise operate in entirely the same policy and labor market environment as the program participants in the treatment group.
3. Except by chance, they are (collectively) the same kind of people or firms as the people or firms put in the treatment group.

The first of these conditions assures that the control group differs from the treatment group on the factor of interest—the intervention whose impact we wish to understand—while the last two conditions assure that *nothing else between the two groups differs…nothing except the policy/program of focus.* This is akin to carefully controlling all other factors in a chemistry lab while deliberately varying one factor to see how much difference it makes: if the results turn out differently, and the pattern is repeated over and over in multiple "runnings" of the experiment, scientists can confidently conclude that the variable factor *caused* the difference in outcomes produced.

Some Bellwether Experiments

The use of randomization in this way, following what is often called an "experimental design" for measuring causal impacts, began in the social sciences in the late 1960s and early 1970s with the negative income tax (NIT) experiments.[5] Researchers proceeded in much the same way as scientists test new drugs in a laboratory or clinical setting. They took a group of individuals that normally would be subjected to the same set of policies and split it at random in order to deliver the new, NIT intervention, to just one of the two subsets (the treatment group). The remaining, non-treated group—i.e., the evaluation's control group—differed from the treatment group *at the point of selection* only by random sampling error.

The same design has been used in many employment and training program evaluations, both ETA-sponsored and others. These include the National Job Training Partnership Act (JTPA) Study of the 1990s and the more recent National Job Corps Study, two ETA-funded projects that randomized potential program participants in a large number of localities across the country. Many welfare-to-work employment strategies have been evaluated this way, again involving large numbers of would-be program participants in a de facto lottery that determines which get the "treatment" and which do not.[6]

In large samples, with many different individuals allocated to the treatment or control groups on a purely random basis, any chance differences in preexisting characteristics between the two groups tend to disappear through what is known as the statistical "law of large numbers". In this way, it becomes very unlikely that observed differences in later labor market outcomes between the two groups are caused by anything other than differential exposure to the program or policy of interest. And the analysis of data from an experiment of this sort can be simple and transparent—what happens "with" an intervention, compared to "without", all other things equal—and at the same time conclusive and unambiguous in what it tells us about program or policy effectiveness. For example, experimental findings from the National JTPA Study led Congress to make major changes to the funding of the out-of-school youth component of that program.[7]

[5] See Greenberg et al. (1997) for summaries of the NIT experiments and the large number of other social experiments that have been undertaken since then.
[6] The most recent assessment of this set of studies, including its methodological strengths and weaknesses, appears in Moffitt (2002).
[7] See Greenberg et al. (1997), 391.

The Power of the Experimental Method

This description explains why we should do experiments to decide if programs or policies are effective. For the same reason experiments are ubiquitous and invaluable in advancing our knowledge about chemistry, biology, or medicine: to vary the one factor of paramount interest (in this case, access to ETA-funded employment assistance) while controlling all others in order to isolate the effect of the variable factor. In the chemistry laboratory, the experimentally-varied factor might be a heat setting or the time interval for immersion of a solution in some catalyst. In a controlled psychology "laboratory" it could be ambient noise or time of day and its effects on performance.

In the policy world, it gets messier but the principle is the same—if you know outcomes can only vary from one run of a process or test procedure to another because of some manipulated factor, you can be sure you are looking at the *causal impact* of that factor when you get different results. Otherwise, the question arises—and at root never goes away—of the possibility that other co-mingled influences were really what made the difference.

Experiments as a Response to the Problem of Selection Bias

This raises the ubiquitous problem in *non*-experimental studies of employment and training program impacts, that of "selection bias". Selection bias arises when those who receive an intervention differ from the people in the counterfactual *on factors other than simply the intervention*, for the very reason that *they were selected (or selected themselves) to get the intervention*. For example, many participants in Workforce Investment Act (WIA) training programs probably turn to that source of assistance because they have special needs or distinctive motivations, factors that likely portend differences in later labor market outcomes compared to individuals who do not use WIA services *even if neither group were treated*.

Where outcomes would have differed anyway, based solely on preexisting differences that lead to selection, the comparison of program and counterfactual results down the line are biased—systematically skewed to either overstate or understate the impact of the intervention by adding it indistinguishably to the influence of these prior differences. Systematic skewing of the quantity of interest (here, the impact of intervention) is called "bias", and skewing as a consequence of who selects into or out of the intervention of intervention is called "selection bias."

The largest challenge in measuring policy impacts using a counterfactual comes from selection bias. Many groups of employers or workers can fill requirements 1 and 2 of a good counterfactual listed above—the intervention of interest does not affect them, and they operate in the same economic and policy context otherwise. Consider, for example, adult workers returning to the workforce following an absence (e.g., for parenting or schooling) who choose not to use the Employment Service's labor market exchange resources. They operate in the same economic and policy context as workers who do use ES, but may not meet requirement 3 above: *that they be people with equivalent motives,*

interests, and opportunities as program participants. Instead of meeting this condition, they may be self-selected on factors that affect later economic success in their own right, such as easy access to a job in a family business or anxieties about the computer skills needed to use ES databases.

By looking only at people or firms that follow the same selection path—and who all are poised to participate in the intervention at the same time—but dividing their actual exposure to the intervention artificially and at random, social experiments solve selection bias problems of this sort. They are the only research method known to do so.

The Perhaps "Not Golden" Aspects of Randomized Experiments: Threats to Reliability Other than Selection Bias

Carefully designed evaluations of employment policies using random assignment to measure impacts achieve the "hold all other things equal" standard needed to measure effectiveness (and, ultimately, program benefits and costs in a benefit-cost analysis) without selection bias. They should be used when they can be used, and when they do not suffer from other analytic limitations even larger than the selection bias problem they solve. We look at the second of these issues here—the conceptual or analytical weaknesses of random assignment experiments—before going on to the question of feasibility later in the chapter. But first we explain why eliminating selection bias, the one widely-trumpeted and universally-acknowledged benefit of randomized designs, must take first priority in deciding how to measure the impact of employment and training programs.

Evidence of the Prevalence and Consequences of Selection Bias

The literature on employment and training program impacts spotlights selection bias as the principal challenge, finding it to be ubiquitous and often highly consequential in nature. Glazerman et al. (2002) recently surveyed the substantial literature that checks for selection bias in non-experimental impact estimation techniques by comparing their results to those of a randomized experiment where selection bias cannot arise. They conclude—based on all 16 studies they could find that measured the impacts of employment and training services on disadvantaged populations experimentally—that "our findings cast serious doubt on the ability of quasi-experimental designs to replicate experimental findings (p. 43)."

In study after study, the performance of non-experimental methods in replicating experimental findings has proven spotty at best. This has led many authors of these "replication studies" to conclude along the lines of Wilde and Hollister (2002) that "non-experimental estimates [are] not reliable guides to true impact" (p. 32). The two types of findings differ because non-experimental estimates are subject to selection bias while experimental findings are not. Thus, ridding impact estimates of selection bias has to be the starting point for assessing the value and contribution of ETA's policies.

Basically, we must assume that self-selection and program intake selection will produce misleading results when using comparison group counterfactuals to measure impacts unless countered by some means. The one reliable way for countering selection bias is random assignment. Greenberg et al. (1997) in their definitive *Digest of Social Experiments* make the same two points: "The selection bias associated with non-experimental estimates [is] often larger than the true program impact" (p.13), and "There is essentially one reason for social experimentation: random assignment is the only known means of eliminating selection bias" (p. 12).

Common Complaints about Experiments—"Knocks" Both Deserved and Undeserved

Random assignment designs do have their downsides, which must be acknowledged at the outset. Apart from the issue of feasibility and costs (discussed later), these include many points that apply to all types of research studies—incomplete data collection, limited sample sizes when looking at effects on subgroups, inability to sort out cleanly the causes of impact variations across locations, lack of assured reliability for national policy when study sites are not nationally representative, and better ability to measure the average impact of a program or policy than to sort out the distribution of impacts across a heterogeneous population.[8]

Much has been made about these shortcomings in the literature comparing experimental and non-experimental methods, but often without acknowledgement that they are not intrinsic to the use of random assignment to create treated and untreated analysis samples. Naturally occurring populations, one with and one without program/policy exposure, can be and often are studied in non-representative locations, with incomplete data and little capability to sort out who benefits more and what accounts for differences in apparent impacts across subgroups and locales.

When comparing the pros and cons of different impact analysis approaches, the importance of these limiting factors varies from study to study. But it has no inherent relationship to whether random assignment or some other means is used to obtain a "no-intervention" counterfactual sample. Going with random assignment can make it harder to avoid one or more of these limitations, easier to surmount others, and add to or subtract from the total resources needed to implement and collect data for the research at a given level of study reliability; no general, universal distinctions can be drawn on this basis between randomized and non-randomized impact evaluations.

The factors that are *distinctive* between the two methods are but few. These include experiments' one distinct advantage:

- Random assignment eliminates selection bias as a threat to reliable measurement of policy effects by decoupling who gets a program or policy intervention from everything else that could influence their outcomes. The treatment and control groups differ because the first "won the lottery" and the other did not. This is no difference at all except for the events that come as a consequence of "winning the

[8] The particular challenges to experiments in this latter regard are discussed in Heckman et al. (1997).

lottery"—i.e., exposure to the intervention of interest—that do not take place for lottery "losers." All subsequent differences in outcomes are therefore known to stem from this policy distinction, not from personality or background factors that led one group to be exposed and the other not.

They also include five distinctive disadvantages:

- Measurement of effects on those *assigned* to get the intervention treatment, not necessarily those who get it;

- Failure to compare the intervention's services to no services at all, instead comparing "our services" to "everything else that's out there;"

- Counterfactual experiences in the control group that are distorted by the "eased up" rationing of substitute services from sources other than the one being evaluated;

- Distorted treatment group experiences due to changes in program scale or the population served;

- Elimination of selection bias only for the difference in experience controlled by random assignment, not in other places where analyses of ancillary research questions might encounter it.

Most of these are valid concerns when random assignment is applied to *programs with a limited number of "slots" that can serve only a fixed number of people*. But they do not apply to regulatory policies or other ubiquitous interventions that automatically affect all people or firms in a given category no matter their number. Other issues on the list arise for both ubiquitous policies and limited-scale service delivery programs. All are totally unique to experiments, since they arise entirely as a consequence of the evaluator's manipulation of program/policy access through the device of random assignment.

Each could be effectively countered by *not* using random assignment.[9] This means there is a clear trade-off when picking an impact evaluation approach (assuming both methods are feasible and affordable): run an experiment with all the bulleted virtues and concerns listed above, or get no help on selection bias while avoiding the other listed concerns by not randomizing. The case for favoring random assignment experiments wherever they are feasible follows from the overriding importance of removing selection bias in

[9] For example, a purely "no service" counterfactual could be created reactively after the fact in a way that assures a sharp comparison to the services provided by the studied intervention. This would be achieved by excluding from the comparison group anyone who was exposed to the intervention under study or anything like it from other sources. But doing so with an experimental sample destroys the comparability of the remaining control group cases to the treatment group. Similarly, the risk of comparison group cases getting more services than is natural due to "eased up" rationing could be eliminated by forming the "no service" counterfactual after the fact…but on a non-matching basis with the treatment group. By selecting comparison group cases from populations that occur naturally, there is no risk that the research changes the program's scale or service population.

evaluating labor market interventions as attested by the literature reviewed earlier, and the fact that the five concerns about randomization identified above are all either manageable or off base. We explain why next.

Measuring the Effect of Service Receipt Rather than Service Assignment

The most easily dispensed with criticism of random assignment impact evaluations is the charge that they reveal the impact of the "intention to treat" —called ITT impacts by Heckman et al. (2000)—rather than the impact of actually being treated—what Heckman et al. call the "impact of the treatment on the treated", TOT impact. This charge arises whenever less than 100% of the treatment group participates in the intervention…i.e., when the "treated" group is different (smaller) than the entire experimentally determined treatment group. Less than 100% participation is common in random assignment evaluations, since individuals cannot be compelled to take part in an intervention such as job training simply because they applied and appeared likely to do so at the point of random assignment.

But incomplete participation by the treatment group is also readily correctable when measuring impact, by applying what is called the "no-show adjustment" to the original experimental estimate. First formalized in the evaluation literature by Bloom (1984), this adjustment *assumes that the intervention has no effect on those members of the treatment group who never participate in any intervention activities*—for example, those randomly assigned to a residential Job Corps treatment who never report to their assigned Job Corp center. This assumption is viewed as innocuous by almost all evaluators and policy analysts in most situations.

Where the assumption holds, the initial measure of impact—the intervention's average impact across all treatment groups, including both potentially positive (or negative) effects on participants and 0 effects on non-participants (the "no-shows")—can be rescaled to the average impact *on just those who to participate* (i.e., the effect of the treatment on just "the treated").[10] No assumptions regarding the similarity of participants and "no-shows", or the ability of statistical methods to adjust for differences between them, are needed here; they can be as different as night and day and the result is still valid as long as the intervention has no effect on the "no-shows."

[10] The rescaling divides the original impact estimate by the participation rate in the treatment group (or, equivalently, by 1 minus the "no-show" rate). The result of this calculation gives the average effect of the intervention on *participants* necessary to have produced the initially observed overall average effect when "no-shows" experience 0 effects. If E is the overall average effect and F the average effect on participants, we can depict the "no-show" adjustment as solving the equation $E = pF + (1-p)\,0$, where p is the participation rate. This equation simply says that for the treatment group as a whole, the average effect E is a combination of effects of average size F for the share of the treatment group that actually participates (= p) and effects of average size 0 for the share of the treatment group that does not participate (= 1-p). The second term of the equation drops out, leaving $E = pF$. We can then determine the average effect on participants by solving this simplified version of the equation for F: $F = E/p$, the original experimental estimate divided by the participation rate.

Looking Beyond the "No Services" Counterfactual as the Ideal

In a decentralized, fragmented federalist system, the policies and services of one branch of the national government such as ETA will be supplied in similar if not identical form by other government agencies. Random assignment does not control whether individuals access these "substitute" services; as a result, some control group members inevitably will do so. This is the case, for example, when ETA-sponsored interventions do substantially the same things for members of the same target groups as state job training programs for welfare recipients or local economic development projects sponsored by a number of other federal and state agencies.

These forces *give randomized impact studies the same character as the real-world programs they are intended to evaluate, and hence are strengths rather than weaknesses of the experimental approach.* Just as is true of an experimental control group, some of the people currently served by ETA-sponsored programs would obtain similar assistance from other sources were ETA's interventions not in place and some would not. It is precisely the difference between these two circumstances—an ETA-served treatment group and a partially non-ETA-served control group—which the Department of Labor controls when implementing any of its policy and program interventions. *It should not seek to impose any stronger contrasts between intervention and comparison group samples when studying one of its programs.* Knowing how a given intervention strategy, uniformly imposed on all members of the target group, compares to no intervention at all does not help in social decision-making in a fragmented federalist system of many intervention sponsors and selective consumer participation.

To put it in a nutshell, if some of those steered away from ETA's interventions by assignment to an experimental control get someone else's similar intervention, *we have the right treatment and control group comparison for the policy choice ETA actually controls . . . and for the true difference ETA's involvement in this program/policy area actually makes to the average person admitted to its programs given all else that exists beyond ETA's control and purview*. Were ETA not offering this particular intervention, some would get something similar elsewhere, and for those the value added by ETA's programs is truly lessened by the existence of alternatives. It should be measured that way in an impact study designed to document what that particular program contributes, not to decide if the intervention strategy overall has value compared to a world where none of it is available. Looking at "our services" compared to "everything else that's out there" is exactly what ETA should be doing to justify its program and policy portfolio, since if "everything else that's out there" is enough the money spent on ETA programs could be cut back without consequence.

This is precisely the view adopted in the new OMB Program Assessment Rating Tool (PART) referenced earlier. The question right after whether the program of interest "is designed to have a significant impact in addressing the problem" asks "Is the program designed to *make a unique contribution…not…redundant of any other* Federal, state, local, or private efforts?" [emphasis added].

The Unavoidable Uncertainties of Possible "Queuing Effects"

The argument just presented for "letting happen what will happen" in the control group leads to a difficult but crucial question: Doesn't the *existence* of the ETA-sponsored intervention under study *change* what "will happen" to members of the control group by taking some of the people interested in assistance out of their queue and putting them into ETA-sponsored service "slots"? This is a slippery but essential distinction concerning the total supply and demand of services delivered through a limited number (though possibly a very large number) of program "slots." It does not apply to any ETA interventions that automatically apply to every worker or every employer in a defined target group, such as unemployment insurance benefits. But it is crucial in other areas, such as WIA-sponsored worker or employer assistance involving services for which total supply is limited by institutional capacity or funding limits.

For example, the number of people who receive vocational training each year, from ETA-funded programs and those sponsored by other agencies outside DOL, depends on the total number of people ETA funds. As a "thought experiment", suppose WIA-financed training was completely eliminated in a given year. In this scenario, the total "supply" of services and service slots would fall precipitously for the consumer groups served by that program. Where those people turn, and to what extent they access alternative services to help build employment skills, will strongly determine the importance of having WIA in place as it now exists *compared to no WIA training at all*.

If this is the policy choice Congress or the Department is contemplating—continuing versus eliminating WIA-funded training—one would want to run an experiment in which:

- Treatment group members are given access to WIA; and

- Control group members "compete" for access to training services from non-WIA sources but do so in a "market" in which treatment group members are also vying for those same alternative slots.

Unfortunately, the second condition cannot be met: we want our treatment group to simultaneously do the WIA "treatment" and jostle with our control group for access to the limited number of alternatives to WIA, sometimes squeezing them out of those slots. If we don't get the latter—as conventional experiments do not—we will see *too much use of alternative services in the control group*, and hence too small a difference in average outcomes between the treatment and control groups when measuring impacts. Controls really are not accurately reflecting the world without WIA, since in an actual counterfactual world they would have to share non-WIA training slots with members of the treatment group as well as everyone else with whom they actually *do* share those slots.

The negative judgment just delivered on the reliability of the control group counterfactual hinges on two as yet unstated assumptions:

- Impact evaluation results will guide a decision to either keep WIA at its current scale or eliminate it altogether; and

- Other programs that provide similar services to the same consumer group *would not expand their scale were WIA eliminated.*

If choosing between full-scale WIA and no WIA, and if expecting the "hole" it would leave to not be filled in at all by other employment and training service funders, ETA would indeed want control group members to have to tussle with treatment group members for other, non-WIA training slots to achieve the appropriate counterfactual for the policy choice it faces. But not if the policy choice concerns expansion or contraction of WIA funding *at the margin*, or if ETA expects other funders to expand services in response to the "shortages" created by WIA's disappearance.

In the first of these scenarios, only a small share of all those seeking WIA-type services would be affected by WIA capacity expansions or contractions at the margin, with almost as many served by WIA itself as in the experimental treatment group (marginally fewer, or marginally more). In this circumstance, what happens to control group members should represent well the options and outcomes of individuals who are marginally displaced from WIA—they really would not have to compete with the workers staying in that program as its size changes only fractionally. Alternatively, if WIA were to disappear, other programs might expand their scale to make up most or all of the difference. So the contrast produced by the treatment-control comparison in an experiment—matching groups participating in WIA compared to other options—would again trace the right consequences of ETA's policy choice….if choosing less (or more) WIA means others will choose to fund more (or less) of their alternative services.

With most evaluations of existing programs likely to influence funding and scale at the margin rather than in an "all or nothing" way, and with the potential for partially offsetting adjustments in the scale of alternative services in a fragmented federal system, randomized experiments with *full* access to alternative services among control group members seems a better approximation to the desired evaluation counterfactual than experiments with *no* control group access. Neither is perfect, but in principle the perfect version of control group experiences is unknowable until policies are changed—either marginally or dramatically--and other agencies react—either a little or a lot. Absent that information, a cautious approach featuring marginal changes and more modest treatment-control differences in service access—i.e., the approach actually produced by most social experiments—provides the safer basis for policy assessment.

Limiting Potential Distortions to the Treatment Group

A further intractable, but possibly minor, problem of randomized impact studies arises for interventions with a fixed number of service "slots" when some of the people or firms

that would ordinarily occupy those slots instead join the control group. Removing a portion of the normally-served population necessarily results in one of two changes in the program's operations:

- It serves fewer people, operating below capacity (or, if below capacity anyway, even further below capacity than usual); and/or

- It serves people who ordinarily would not be admitted due to capacity constraints.

You can't turn away some of the usual service population without creating added, artificial vacancies in the program or obtaining replacements from outside the normal service population, or both. While not a problem for policies that automatically reach all members of their target population, for fixed-size service programs there is no way around this consequence of randomization—if you have artificially pulled out some would-be participants, you necessarily leave the program short (or shorter than usual) or bring in others who normally would not be served.

The question is whether either of these results matters to the size of program impacts, the quantity one seeks to measure through random assignment. Likely both situations do, though perhaps not to a very great extent. A program with unnaturally created vacancies may deliver services differently for the consumers it does serve. If budgets remain unchanged, the typical participant in a less fully subscribed program may receive more services and experience a larger impact. Or lower numbers may change the dynamic of any group elements of the intervention, such as job club interactions, either increasing or possibly diminishing impacts on remaining participants.

Alternatively, program scale and operations could remain unchanged if added people or firms are served that normally would be closed out due to capacity limits. These are clients of lower priority in the program's view, or clients with less motivation or ability to ensure that they make the first cut. In a normal year, when random exclusions are not imposed on those "ahead of them in line" for the sake of the research, they would not be served. Unless a lottery of some sort is *ordinarily* used to ration slots among a surplus set of applicants, the usual means of obtaining access *creates distinctions between those who get in and those who do not*—the selection problem discussed above. It may be that the applicants thought most in need of help receive priority or that those expected to benefit most from the program's services (which might or might not be the same people) do. On one factor or another, then, entrants differ from the interested non-entrants, and these differences may correlate with the size of program impacts.

This creates a selection problem of another sort, though one less likely to matter appreciably to the size of measured impacts: the intervention (treatment) and counterfactual (control) samples match one another through random assignment *but they both represent slightly the wrong set of people, a somewhat different set than would ordinarily be served.* Or, as noted previously, they need not do so if the pool considered for participation does not broaden…but then the program's operations change scale because of random assignment.

No good data exist on how much these factors could matter to the size of impacts measured from the experimental data. What we do know is that both these problems—artificial shortfalls in enrollment and different-than-usual participant populations—diminish as the control group shrinks in size relative to the program's capacity. When control group members are spread over many local programs, with only one or two individual control group cases in any community, no program can be pushed much below its regular scale or forced to serve very many new customers by the removal into control status of some it normally would serve. The National Job Corps Study provides an excellent example of steering clear of distortions in the treatment group through minimal control group exclusions in any one local program site, while still achieving a large total control group and evaluation sample through inclusion of many sites.[11] This model should be emulated whenever possible in measuring the impact of ETA programs where scale constraints normally operate.

What's Not Solved by Experiments: One Less Problem than Before

A final point often raised by critics of randomized impact studies is that, while eliminating selection bias in measuring the effect of participation *at the point of random assignment*, they do not provide equally unequivocal information on the consequences of participation at other stages of the intake process. For example, they still leave evaluators with no options but to do non-experimental comparisons to measure the relative impacts of different sequences or "dosages" of services that are only determined following randomization (and hence are never known for the control group). Conversely, an experiment cannot show directly how much difference program interactions *prior to random assignment* might have made to participant outcomes, since these effects occur for both treatment and control group members.

All of this emphasizes the importance of choosing wisely where to position random assignment in the intake flow and early steps of participation. But does it create *liabilities* for the experimental approach as compared to non-experimental alternatives? No, since other types of impact analysis do not remove selection bias *at any point*, and hence are equally hobbled at all points where an experiment does not place random assignment *and are much more hobbled where the experiment does place random assignment.*

Solving one selection bias problem—the one that could distort answers to the most important policy question to be evaluated—is clearly a virtue over solving none…..just as diets that lead to 10-pound weight losses are preferred over diets that do nothing, rather than faulted because the former still leave a good bit of excess weight still hanging around. Especially when universal weight loss was not the goal; when what counts is slimness in the hips—and, yes, a slimmer waistline and less flabby legs would be nice as well—the diet that takes off pounds in the middle gets the best reviews. It is not clear why the same has not always been evident in critiquing the one impact estimation methodology that eliminates selection bias around the point of random assignment—it's a

[11] See Schochet et al. (2001).

better "diet" than the ones that don't even do that, regardless of what other problems it may not solve!

Are Experiments Sufficiently Feasible to Serve as the Customary Standard of Practice in Impact Evaluation?

By beginning with the alleged failings of random assignment impact evaluations, we have built a case for their use wherever possible and begun to see ways to improve their design and reliability. The rest of the way to a plan for regularly measuring and reporting the effectiveness of ETA programs and policies would seem straightforward: "Just do them (experiments)." But it is not. While frequently touted as the ideal, randomized experiments as a methodology often slip from what might seem to follow that accolade—status as the expected standard of practice currently in wide utilization—into the coveted but difficult-to-attain ideal. Why is that? This paradox must be addressed before we describe the ideal model for conducting impact and benefit-cost research on ETA programs in terms of randomized experiments.

Ethical and Cultural Considerations

The biggest holdup of experiments is the perception of the operational and political infeasibility of randomization as a regular feature of government program administration. Experiments have been used in the U.S. to evaluate numerous pilot and demonstration projects, and less often to study existing national policies or programs (Greenberg et al. 1997), so at some level they are feasible and operationally acceptable. Yet in contrast, they have never been used for national policy evaluation in Western Europe,[12] and rarely for social policy research of any kind on that continent, largely out of concern over the ethics of service denial to individuals.[13] While the atmosphere concerning research rigor is changing there, the advent of highly rigorous randomly-designed impact evaluations on the continent may yet be decades away.[14]

It may be useful to begin with fundamentals when approaching the controversial subjects of feasibility and ethicality. Though these points do not diminish the ethical, legal, and

[12] A recent attempt to use random assignment to measure the "additionality" or net impact of employment services on people with disabilities in the United Kingdom—the New Deal for Disabled People National Extension—was squelched in late 2001 by a surprise (but not surprising) preemptive ministerial announcement in Parliament; efforts in that country to use experimental methods are at present again confined to small-scale demonstration projects with limited geographic coverage, most prominently the Job Retention and Rehabilitation Pilot and the Employment Retention and Advancement Demonstration (source = author's personal involvement in the cited research studies).

[13] Conversation with John P. Martin, OECD Director for Education, Employment, Labour, and Social Affairs, March 13, 2002. Martin and Grubb (2001) provide a comprehensive summary of methods and findings from active labour market policy evaluations in the 30-member OECD over the last ten years.

[14] Martin and Grubb (2001) report that "Few European countries have carried out rigorous evaluation until recently. Happily, this is changing as tight fiscal constraints make it imperative to get better value for public spending on active labor market policies. As a result, some European countries and Australia are beginning to undertake rigorous evaluations of their labour market programmes" (p. 21). No movement toward random assignment is evident anywhere except the United Kingdom (source = conversation with John P. Martin, March 13, 2002).

political concerns around random assignment, they stand in some importance in their own right:

- If a program has to limit the total number of people or firms served due to funding or administrative capacity constraints, *it will in some way ration access*. Random assignment, with control group members left out of the program's services, is just one way to ration. Whether a better or worse way is the real question.

- Society benefits from good information on program effectiveness and may be justified in allowing small numbers of individuals or firms to be harmed in order for the research needed to gather that information to take place. The interests of many future program participants and every taxpayer may legitimately outweigh the true costs born by a comparatively small number of control group members.

- If a program has to be evaluated to determine if it benefits participants, *being turned away from participation cannot be presumed to be a worse thing for the individuals involved than being admitted.* If job training on average does not lead to better employment outcomes—the very question an impact study seeks to answer ("question", as in "we don't at present know")—participating in it at best constitutes a neutral situation and may be a disadvantageous use of time or unrealistic heightening of expectations.

No researchers, and perhaps not even policy makers, can appropriately weigh the balance of these considerations in seeking a justification for the broad use of randomized impact studies as a way to improve policy. At a minimum, a government benefit or service established by law as an *entitlement* for all eligible individuals cannot be pushed into the random assignment mode without legal ramifications. Such programs under ETA's purview—most prominently the federal-state system for providing Unemployment Insurance benefits to workers who have lost their jobs—will not under current law be evaluated through random (or any other type of) exclusion of qualified individuals who wish to claim benefits.

Technical Challenges that Sometimes Increase Scale and Costs

Apart from legal limitations and the ethical and political judgments that others must make, what else can researchers say about the feasibility of randomization as a research technique in specific applications? A single axiom governs all other situations, in the eyes of this researcher:

> *With enough resources, any intervention for which it is legal (and ethically and politically acceptable) to provide access to fewer than 100% of those who want access can be successfully evaluated with random assignment.*

This "feasible…with enough resources" maxim encompasses all the situations where scientific and technical limitations on experiments are often cited, except those for which the limitation would be present even without randomization. The best known such

cases—together with the reasons they don't undercut the viability or comparative difficulty of experiments—are listed below:

- Saturation interventions that affect entire local communities, including interventions that influence behavior simply through the knowledge that they exist (e.g., regulatory policies on workplace safety). The U.S. is a very large nation, with thousands of local communities that could be randomly assigned into or out of a particular policy or intervention. (Saturation also makes data collection more difficult and expensive and any impacts that do occur harder to find if diffused across many people in the community, but these drawbacks afflict any impact analysis not just experiments.)

- Programs that struggle to meet enrollment targets—only a few control group cases need to be sampled in any locality, as long as enough localities can be included in the study. And sufficient technical assistance resources can be added to the evaluation budget to raise application counts sufficiently to accommodate a modest-sized control group.[15]

- Interventions whose full effect will not be seen in contrasting treatment and control group behaviors unless control group members believe the intervention will *never* apply to them—control group "embargoes" need not be time-limited unless ethical concerns become too extreme.

- Interventions with low participation following randomization, leading to small average effects for the treatment group as a whole even when successful for those who participate—small average effects can be detected in sufficiently large samples, and readily translated into average effects on just participants through the "no-show" adjustment discussed earlier.[16]

- Programs or policies that pose questions of effectiveness in multiple areas, each needing to be answered without the confounding effects of selection bias—multi-stage random assignment is not impossible to design and implement effectively, nor are evaluation strategies that position random assignment at different points in different sites and selectively examine only those sites relevant to each policy question when bringing together the multi-site evidence. Moreover, as noted previously, one selection bias problem solved through randomization is better than none.

- Interventions that have "general equilibrium" consequences beyond the experimental sample, such as policies that change which workers have the fastest access to a fixed number of job openings and thus affect those "displaced" from

[15] This was done successfully, for example, in the National JTPA Study.

[16] The translation simply divides the average treatment-control difference by the participation rate. This so-called "no-show adjustment" (Bloom 1984) factors up the average impact on all treatment group members to an impact size on participants large enough to account for the full difference in outcomes observed between the treatment and control groups by assuming that nonparticipants in the treatment group—who cannot have gained from the intervention—do not contribute to that difference.

the jobs as much as those placed[17]—general equilibrium analyses of labor market interventions are always difficult, and no more so for having measured the direct effect experimentally.

- Evaluations of national programs that need to be based on a statistically representative set of sites—at least a half-dozen random assignment impact evaluations of social programs have now been conducted in geographically-based probability samples of the nation without substantial attrition of local programs from the research.[18]

- Policy assessments in which results are needed quickly, without a multi-year lag to set up and conduct random assignment and wait for medium- and long-term labor market outcomes to emerge—all research based on long-term outcomes lags by a considerable amount the exact policy intervention under study; *additional* lags while random assignment is set up would not occur in a system of regular, experimental evaluation of ongoing programs and policies.

All of these difficulties—and likely others not catalogued here—are either common to non-experimental studies or can be overcome with enough investment in site recruitment and data collection on the government's part. That plus enough determination on the part of the federal agency involved to wield, if necessary, its influence and authority over local partner organizations that depend on it for funding.

Are experiments sufficiently feasible to serve as the customary standard of practice in impact evaluation? Technically, yes, if the will and the investment resources are there to use them.[19] Should and will they be used? That depends in large measure on their costs compared to those of alternative research strategies—the final feasibility topic to be taken up here.

Are Experiments Too Expensive?

The financial costs of experiments, to those sponsoring the research and, hence, indirectly to taxpayers, have often been put forth as an important obstacle to their use. It is not possible in this essay to investigate the costs of alternative methods in any detail. Better than these numbers, perhaps, is an examination of the prominent opinions on this matter among those most prominent in doing, critiquing, and summarizing the use of random assignment methods to measure impacts from labor market interventions. Current thinking virtually across the spectrum is that, while there may be other reasons to avoid experiments in certain circumstances (such as those discussed above), budgetary

[17] Smith (2000) provides several good examples of these general equilibrium effects of labor market interventions and explores the evaluation challenges they create.

[18] The Food Stamp Employment and Training Evaluation (Puma et al. 1990) is one example. The author is indebted to Thomas Cook of Northwestern University for the overall count cited.

[19] It is also worth noting that the *technical feasibility* of random assignment impact research in many applications will only become a *reality* if the scientists engaged to carry out the research believe it can be done on an experimental basis and at the same time recognize the myriad reasons it will be ceaselessly difficult and challenging to accomplish. Then it will happen.

constraints on federal agencies are not valid reasons to avoid them, especially in an era of heightened fiscal accountability and results-focused policy mandates.

One part of this reappraisal is better recognition that the appropriate basis for social choice among competing research techniques is the *marginal* cost of experiments compared to other equally ambitious research studies that tackle the same set of policy questions. Obtaining broadly representative data on labor market outcomes for thousands and thousands of workers, both with and without a policy in place, is never cheap—a facet of cost invariant with *how* the program participants and non-participants are selected.

The one exception is data from large surveys of households and workers collected for other non-evaluation purposes, such as the Current Population Survey and the Survey of Program Dynamics. There, the social cost of data collection has already been paid, and individual federal agencies can use the information at low cost. Unfortunately, reliance on national surveys of this sort to measure the impacts of discrete employment assistance programs was the first non-experimental approach to impact evaluation to be discredited by careful methodological research.[20]

An even more telling cost assessment concerns the "opportunity costs" of *failing* to do experiments—the money spent on ineffective programs that continue to be funded (and continue to offer false hope) because unbiased information on their inadequate impacts is not available. On this basis, some observers see the balance clearly swinging toward experiments as the comparatively *low cost* investment option compared to other methods once an appropriately broad social viewpoint is adopted. Not surprisingly, these have been among the most outspoken supporters of random assignment studies; see for example Burtless and Orr (1986) and Orr (1999).

Importantly, prominent researchers on the other side of the issue have more recently taken similar stances. Smith (2000) summarizes these views as follows:

> Random assignment does have its costs, as it typically requires substantial staff training, ongoing staff monitoring and information provision to the potential participants… At the same time, as pointed out by Heckman, LaLonde, and Smith [1999], this case can be overstated (p. 21).

Finally, there is the broad barometer of the "marketplace" for experiments. Here, one takes past practice as a guide to smart investment in methodologies for the future. Greenberg et al. (1997), in their *Digest of Social Experiments*, sum this up as follows:

> From either perspective, sponsoring a social experiment requires complex resource allocation decisions. The social experiments conducted to date were authorized by many different [individuals] representing a wide spectrum of political views… It is striking that many very different individuals decided that this type of investigation is worth its costs.

[20] See LaLonde (1986) and Barnow (1987).

It would be difficult for today's national government to back away from this same decision on the grounds of insufficient funds, in circumstances—such as those facing evaluators of labor market intervention—where sound science and reliable policy guidance are known to depend on using random assignment designs.

C. THE SYSTEM AS IT WOULD LOOK IF EXPERIMENTAL IMPACT RESEARCH WERE DONE EVERY TIME

The conclusion that experiments are generally feasible to measure the impacts—and hence the social benefits and costs—of labor market interventions—creates the opportunity to describe how ETA's ongoing program evaluation agenda might be structured to provide the best possible information on what is working and what is not in terms of impacts. This "diorama" of the ideal program evaluation system abstracts from the costs of implementing the best-available system, both financial and in terms of political resolve, on the theory that a scientist should first describe to government decision makers the best science has to offer before looking to potential practical limitations on its application.

However, the discussion does recognize legal constraints on evaluating entitlement programs with experimental methods. The implications of funding constraints, and of the management and public relations burden of taking on many large-scale experiments as an ongoing policy evaluation strategy, are introduced later.

Compatibility of the Major ETA-Sponsored Interventions with Random Assignment Evaluation

To describe the best possible approach to impact accountability in ETA's policies and programs, we begin by reviewing the main types of interventions ETA sponsors. These fall into six broad areas:[21]

- Provision of labor market exchange information;
- Delivery of job search assistance
- Provision of worker training;
- Entrepreneurial training and support;
- Local economic development assistance and incentives; and
- Payment of income support benefits, primarily unemployment insurance benefits.

[21] Special demonstration projects, such as Growing America Through Entrepreneurship (Project GATE) and newly proposed but not enacted interventions such as President Bush's proposed Personal Reemployment Accounts are not included here since they do not at this point require ongoing national-level evaluation. The latter would, of course, if enacted.

Detailed Breakdowns Not Examined

Much of the delineation of individual programs within these areas has to do with target groups rather than the types of assistance provided: Indian and Native American programs; migrant and seasonal farm worker programs; youth-oriented (e.g., Job Corps, apprenticeship) and seniors-oriented (Senior Community Service Employment Program, or SCSEP) interventions; and assistance targeted to those who lose their jobs due to consumer goods imports or job "exports" (e.g., Trade Adjustment Assistance, or TAA), military veterans (Veterans Employment and Training Service, or VETS) and welfare or former welfare recipients (National Welfare-to-Work Grants Program). Another dimension of variability concerns the point-of-entry for the intervention: whether employment tax incentives (e.g., TAA's Health Insurance Tax Credit) go to employers or employees, whether worker training is tailored to individual employers and their work sites or offered to the wide range of workers at some "outside" location.

ETA also has certain regulatory roles involving such things as labor certification for foreign workers, regulatory compliance assistance for recipients of ETA funding grants, and worker certification under the Work Opportunity Tax Credit and the Welfare-to-Work Tax Credit. These are presumed to be necessary legal and administrative functions of the agency not subject to review in terms of net contribution to society—or, if subject to review, part of a separate category beyond the scope of this chapter[22]—and will not be considered further here.

Summary of Random Assignment Issues Posed by Each Program Area

The next step in developing an optimal impact assessment strategy for ETA-sponsored programs is to consider the relationship between these different types of policy intervention and the scientific and technical limitations on the use of random assignment detailed in the previous subsection. As argued previously, this is *not* because those limitations preclude the use of experimental methods to measure intervention impacts, but because they necessarily influence the *scale* on which random assignment will need to take place if experiments are to go forward.

Table 5 provides a checklist of the scientific and technical issues surrounding random assignment discussed earlier that apply to each of the main ETA program types. Because the issues are the same for job search assistance and worker training, the two program areas are dealt with as a single entry at this point. We discuss the basis for each categorization first moving on to its implications.

[22] For example, more efficient execution of these functions would be of interest, but this lies more in the realm of operations research than policy evaluation.

Table 5: Technical Challenges that Raise Issues Concerning Randomized Impact Studies, for Different Areas of ETA Policy Intervention

	TYPE OF POLICY INTERVENTION				
ISSUE:	Public Labor Exchange	JSA & Worker Training	Entrepre-neurial Support	Local Econ. Devel.	Income Support Benefits
Saturation intervention	-	-	-	X	-
Enrollment Shortfalls possible	-	X	X	-	-
Long-run embargo issue	-	X	-	-	X
Low parti-cipation after RA	X	-	-	X	-
Multiple points of impact	X	-	X	-	-
General equilibrium concerns	X	X	-	-	-
National representa-tion needed	X	X	X	-	X
Impacts needed by site	X	-	-	X	X
Timely results vital	X	X	X	X	X

Issues Concerning Participation Levels

Assistance with local economic development is the one area where ETA's activities affect entire communities, rather than deal with people or firms one at a time, making it a saturation intervention where individual-level random assignment is not feasible. On the other hand, there is no concern about insufficient numbers of people or firms to fill enrollment goals in such programs, nor (it is presumed) for interventions that operate on a "take all comers" basis such as public labor exchange services and unemployment insurance benefits. These concerns can arise for programs with a given, planned capacity such as those delivering job search assistance, worker training, and entrepreneurial development supports.

Concerns over long-run embargoes from services—or even any embargo, in the case of entitlement programs—might arise in any of these policy areas. However, they seem likely to be most acute for interventions focused on individual workers and that address highly debilitating needs for which permanent exclusion from ETA-sponsored assistance might be viewed as most harsh. These include worker skill training and income support during times of prolonged unemployment.

Low participation of the "treatment" group following random assignment can also happen for any experimentally evaluated intervention but is usually avoidable with well-designed randomization approaches. But not always, particularly when dealing with populations that include people with fleeting or underdeveloped desires for the services involved such as those with passing interest in the Employment Service as but one source of job or worker "leads" and those targeted by—but perhaps uninterested in—local economic development efforts.

Measuring the Appropriate Domain of Program Effects

Without examining specific intervention models, it is difficult to anticipate which may generate interest in impact measurement at multiple points in a "building block" kind of process. Examples that come first to mind are labor market exchange services—where policymakers might wonder about program effects from accessing job information, from obtaining specific job referrals, and from going through a job interview on the basis of these connections—and entrepreneurial support programs where the components of independent interest might be business planning assistance, capitalization (e.g., financial stipends), licensing support, etc. One could imagine studies that do random assignment just prior to application of each of these program components to answer multiple impact questions. In contrast, job training and income support programs tend to deliver a single, unitary "package" to everyone who encounters them, posing a single impact question—does the approach work, taken holistically?

The general equilibrium effects of labor market interventions that can confound conventional impact analyses involve the displacement of non-assisted workers by assisted workers when competing for the same job. Labor exchange services and job search assistance and training all have the potential to do this. Entrepreneurial support

and local development assistance are more about creating jobs than filling a fixed number of available openings, whereas income support benefits involve neither. All these interventions, with the possible exception of local economic development programs, need to be studied if possible on a nationally representative basis since they are generally available everywhere and receive support either universally or not at all based on a single national policy decision.

Timing is Everything

The final challenge to experimental impact evaluation—the need for timely results—is ubiquitous…it applies to all of ETA's policy and program areas. *How* timely this information needs to be is more difficult to gauge; i.e., how long can one set of time-bound evaluation findings serve to guide policy before another becomes available based on more recent program experience? The OMB PART guidelines referenced previously suggest that "an evaluation may be scheduled on a periodic basis, such as every two to five years or whatever time schedule is reasonable based on the specific program, its mission, and goals" (instructions for question II.5).

Rather than make specific recommendations in this regard for each policy area—which would necessarily depend on the pace of legislative and "best practice" change, factors that vary from year to year—the discussion here presumes that updated impact information at least in some sites, if not on a representative sample of the nation, is wanted every two years.

Addressing Challenges to Compatibility while Holding to the Experimental Ideal

With enough effort and fiscal capacity, each of the circumstances posing a challenge to randomized impact evaluations of ETA programs can be overcome except the imposition of entitlement status on programs such as unemployment insurance. Legislative authority to experiment in this area through waivers of existing program rules—under carefully regulated and justified circumstances, of course—would be a major benefit to effective policy accountability and improvement at ETA, making randomized impact studies possible in all its major program areas. In most other instances, multiple reactions to the challenges raised are already possible within the experimental paradigm, any one of which is sufficient to ensure success. Table 6 lists the responses available to cope with each generic issue.

Importantly, none of the responses shown drops the random assignment approach needed to assure reliable impact results free of selection bias. The main purpose of Table 5 is to drive home the point that, apart from the problem of legal entitlement to benefits (which precludes random assignment of any sort by prohibiting imposition of the "counterfactual" state, a problem that will remain unsolvable until waiver authority is created), *experiments are always possible, if the government is "willing to pay the price" of conducting them* in one form or another. A subsequent section of the chapter considers the scope for less rigorous evaluation of policy impacts, benefits, and costs

should a decision be made against the needed investment in "doing it right" with random assignment.

Table 6: Design Responses to Technical Challenges in Random Assignment Impact Evaluations

TECHNICAL CHALLENGE	DESIGN RESPONSE(S)
Saturation intervention	Randomly assign sites, not individuals or firms, and pay for wider data collection (or settle for available community-wide indicators)
Enrollment shortfalls possible	Provide technical assistance with recruiting Spread study over many sites, with fewer control cases at each Live with "the heat" of underutilized programs
Long-run embargo issue	Do not acknowledge that later admission will be granted Live with "the heat" of permanent denial of the few for the sake of the many
Low participation after RA	Include very large samples, with attendant data collection costs Accept that the impact of early steps will go unmeasured
Multiple points of impact	Include large samples and multiple points of random assignment Add sites and sample size to permit different single points of random assignment in different sets of sites
General equilibrium concerns	Randomly assign sites, not individuals or firms, and pay for wider data collection (or settle for available community-wide indicators)
National representation needed	Spread study over many sites, with fewer control cases at each Live with "the heat" of imposing random assignment on reluctant agencies as *quid pro quo* for receiving federal funds
Impacts needed by site	Add sites and sample size, with attendant data collection costs
Timely results vital	Institute perpetual random assignment in a rotating set of sites, focusing analysis on outcomes available from low-cost administrative data (e.g., earnings, welfare receipt)

Responses in Scope

As can be seen in the table, for several key challenges—saturation interventions, low participation after random assignment, multiple points of impact, general equilibrium concerns, and the need for site-level impacts—one possible response to preserve the experiment involves expanding the study's scale and hence its cost. This could take the form of more sites, more individuals or employers randomized, or both.

In only one instance does the determination to stay with experimental methods leave any alternative to greater investment in study scale, and there only by relinquishing some of the information the study might have provided: where participation following random assignment would be low, move to a later point of randomization but give up information on the impact of the initial elements of the intervention—the ones that would now precede randomization.

Skirting Public Relations Pressure over Random Assignment

Three other circumstances would exact a different kind of cost if ETA holds firm to a commitment to used randomized designs to measure effectiveness as reliably as possible: the political and public relations "heat" that accompanies potentially adverse media, policymaker, and community reaction to random assignment when programs are under-subscribed, embargos from program services need to be long-term, or the need for nationally representative findings requires the compulsory involvement of reluctant local agencies in randomization. Fortunately, in all these cases there are alternatives that do not pose such risks yet preserve a rigorous random assignment design.

For example, both enrollment shortfalls and acute resistance from local agencies chosen for a nationally representative sample (two situations that often go together) can be met by spreading a study over more total sites, so that shortfalls are only trivially increased by random assignment in any one. Enrollment shortfalls can also be met by funding technical assistance to raise flagging outreach and recruiting and raise the applicant flow to provide sufficient numbers for a control group. A particularly bold, but perhaps not unethical, strategy for assuring that excluded control group members behave on the expectation of long-term nonparticipation is to in fact *not* impose a permanent embargo but at the same time neither acknowledge nor announce at the point of initial embargo that admission will later be granted.

Increasing the Timeliness of Research Findings

The final situation in Table 6, shown at the bottom, may also constitute the most universal: the need for timely results from policy impact studies. This challenge is always difficult to meet, and at some level cannot be met: evidence on what works, gleaned from the outcomes of those affected by a policy, necessarily applies to *interventions that have already happened, not interventions that might take place in the future* (i.e., the ones that can be influenced by *today's* policy decisions).

In some sense, all outcome-based research evidence is too old from the time it becomes available; the question is "how much too old?" Lags for experiments tend to be longer, but only because the randomization that sets them up has not yet happened at the point they are proposed as the best way to answer some impact question.

But if recognized as the best methodology from the outset, *the added lag for experimentation can be avoided*. Agencies such as ETA know the impact questions for which they will be held accountable in running their programs, and could certainly plan now to answer them reliably in 2005 or 2006. Knowing this need for information lies ahead, randomization should be undertaken in advance and every two years on a new sample based on the time-lag standard adopted here, so that the most recent information on impacts always becomes available on a timely basis. This would get data into the policy process just as quickly as non-experimental impact analyses of the same people and the same outcomes undertaken at two-year intervals…and better data in the bargain.

Recommendations for Using Random Assignment to Measure ETA Program Effectiveness in Different Policy Areas

Relating the whole set of design response options to the challenges confronting particular ETA programs brings us to the pay-off point in this analysis: a depiction of how full-spectrum, rigorous impact evaluation in all possible policy areas could be carried out as the top research priority at the agency. This constitutes the ideal strategy for establishing accountability and measurement of social contributions from each of ETA's intervention areas, and the main recommendation in this chapter as to the agency's course.

To get there, for each policy area we will address each of the challenges to random assignment impact analysis flagged in Table 5 with one of the response strategies described in Table 6—the one thought best for tackling the particular circumstance involved. For the five policy areas, this leads to the following recommendations. We begin with interventions where the preferred course is clearest and has the greatest commonality to random assignment research projects undertaken in the past. Policy areas that call for more innovative or difficult research plans come last. The specific random assignment challenges the proposal intends to address, from Table 5, are listed at the end of each description in brackets.

- *Job search assistance and worker training.* Do random assignment in a large, nationally representative set of sites, cycling through a different set of sites each year to generate new national impact information each year without burdening any local program with more than a handful of control group assignments per decade.[23] Let denial of access be permanent for those individuals, since training is not an entitlement, has at times been found nonproductive, and can be obtained by most disadvantaged workers from other non-ETA sources. Let any general

[23] This strategy, adopted for just one round in ETA's National Job Corps Study, was first proposed by Larry Orr in a memorandum to Raymond Uhalde in the late 1980s (internal correspondence, Abt Associates, Bethesda, MD). A more complete statement appears in Orr's book on social experiments (1999), from which a number of the suggestions in the current chapter are drawn.

equilibrium effects—i.e., the possibility that the workers served take jobs away from other workers not in the research sample—go unmeasured on the grounds that policies that increase aggregate labor supply should lead to full absorption of all workers into a growing economy with the right macroeconomic policy.[24] [*Challenges addressed: enrollment shortfalls possible, long-run embargo issue, general equilibrium concerns, national representation needed, timely results vital*]

- *Entrepreneurial support.* Conduct experimental evaluations of the largest such programs on a rotating basis over several years. Use multiple points of random assignment to determine who gets assistance with (a) business plan development only, (b) business plan development and—if the plan is completed successfully—financing, and (c) neither. Address enrollment shortfall issues through expanded outreach assistance but allow under-enrollment if necessary. Assure that experiments are conducted under a range of local economic conditions, to provide a basis for determining whether ETA's help makes the most difference in slack or strong labor market circumstances.[25] Use variations in impact from different experimental sites and years, in relation to local economic conditions, to simulate up-to-date estimates of national impact every two years based on current state-level economic conditions. [*Challenges addressed: enrollment shortfalls possible, multiple points of impact, national representation needed, timely results vital*]

- *Local economic development.* In addition to formula-based adjustments, vary the amount of local development aid by site on a random basis. This allows all localities receiving assistance to contribute to experimentally-based evidence on program effectiveness without having to create and justify a large number—probably in the hundreds—of pure control locations as would be needed if the standard random assignment model for individuals were taken to the site level. Use available community-wide indicators of labor market and economic health as key outcome measures, to avoid massive primary data collection costs needed when collecting primary data on hundreds or thousands of communities and citizen and employer participation rates in program-funded activities follow community-level random assignment is low. Keep each site at the same relative funding level year after year and check indicators annually, to have a consistent and up-to-date trend on whether communities do better when funding is higher, all other things equal. Do not publish site-by-site results, since there are not pure control sites with which to compare each local intervention and doing so might call attention to disparities in funding levels maintained for research purposes. [*Challenges addressed: saturation intervention, low participation after random assignment, impacts needed by site, timely results vital*]

[24] See Bell and Orr (1994) for a more in-depth development of this argument.

[25] The first such experiments have already been run, though may be too old to contribute to understanding the effectiveness of these services today. See Benus et al. (1994) for results from early state-run demonstrations in self-employment assistance in Massachusetts and Washington based on a simple two-way random assignment design.

- *Income support benefits.* Random assignment impact analysis is not possible for entitlement programs like unemployment insurance benefits, since by law no individual can be given less than all the benefits for which she or he is eligible—for research purposes or any other. Other methods will need to be adopted to measure the net contribution of providing income replacement through the UI system (see next section of chapter). One component of the UI system that could be evaluated experimentally is the use of worker profiling to refer claimants and require their participation in reemployment services.[26] While these provisions are currently in force statewide, removing referral and mandatory participation following profiling for a small random subset would give an extremely reliable measure of the impact of this component on a state by state basis and nationally. [*Challenges addressed: basic benefits = none; worker profiling = national representation needed, impacts needed by site, timely results vital*]

- *Public labor exchange services.* While not legally proscribed, the challenges to random assignment here—low participation following random assignment, multiple points of impact, general equilibrium concerns, national representation needed, impacts needed by site, timely results vital—are simply too numerous to overcome.[27] Most vexing are general equilibrium effects: an intervention intended to improve the flow of labor market information potentially "rearranges" the worker-employer match-ups that occur throughout the economy, not just for workers or employers who use the service. When some job vacancies and some job candidates "clear the market" faster or in different combinations, options left for those actors *not* using the Employment Service (ES) also change. There is little chance that the randomly split set of would-be ES users, generated by imposing random assignment at intake into the system, will capture all the gains and losses involved…or even necessarily a major share. This suggests that market-wide analyses are needed, based on site-level random assignment. Yet, assigning some localities to a "no ES" counterfactual is not a practical option, given the systems long-established role in local communities and its large group of customers. Nor is it possible to randomly vary the *scale or degree* of system availability across communities as suggested above for local community development assistance—any system that exists will potentially "reshuffle the deck" for job matches in ways that reach a long way into the labor market, irrespective of how many actors it has the capacity to serve (short of a trivially small system, tantamount to creating no-ES control group communities). No recommendations are possible for impact evaluations involving random assignment under these circumstances; we return to what can be done non-experimentally in the next section. [*Challenges addressed: none*]

[26] Jacobson (1999), 12, previously made this suggestion.
[27] See Jacobson (1999), 12, for a more extended discussion of these problems.

D. WHAT TO DO WHEN LEFT SHORT OF THE IDEAL—SUGGESTIONS FOR "NEXT BEST" EVALUATION STRATEGIES AT ETA

The endorsement of experimental techniques as the mode of choice for measuring program impacts, benefits, and costs does not assure it will always be used. For legal or technical reasons, it appears infeasible in ETA's most wide-reaching intervention domains, unemployment insurance benefits and public labor exchange services. It could and should become the basis for routine, repeated program effectiveness research in other policy areas, including worker training and both entrepreneurial and community economic development support, but will it?

Leadership at DOL and elsewhere will decide this, conceivably in the negative. Prior to having that decision, the development of research design options and recommendations cannot safely stop with what has already been said about experiments, even for policy areas where useable random assignment techniques have been identified. ETA may need a place to turn for "next-best" designs in all these realms, we do not know.

The "Second Best" Question

This final section of the chapter considers what should be done where experimentation fails—fails to be viable, technically or legally, or fails in support and commitment as the method of choice. It begins by suggesting ways to analyze the impacts of different ETA programs—and hence their benefits and costs—without random assignment if forced into that circumstance. This opens up new issues concerning the reliability of other methods of inferring the consequences of labor market interventions, particularly the so-called "quasi-experimental" methods that attempt to replicate the all-other-things-equal treatment and control group comparison of an experiment without the benefit of random assignment.

Working through these issues, and choosing wisely a course that minimizes the risk of selection bias in such circumstances, leads us to the recommended "best non-experimental" means of impact assessment for each of the policy areas analyzed in the previous section. For some of these (specifically, unemployment insurance benefits and labor market exchange services), there are in fact the overall "best recommendations" available from impact evaluation science.

The section also discusses some "third-best" strategies for gaining limited insight into the net effects of policies that are not systematically evaluated *in any way*, or at least not so examined very often. These consider the use of far inferior techniques based on old research and performance management system data and recommend strongly against this if avoidable. Their mention constitutes simply a prudent precaution for the worst-case scenario; they are consciously designated as "fall-back" strategies for this reason, and might really be called "fall-way-back" options given how much impact information and scientific reliability they sacrifice relative to the preferred and recommended options.

The chapter closes with suggestions to ETA for supporting the development of more diverse and, potentially, more reliable research tools evaluators might use to measure policy effects when random assignment, for whatever reasons, does not occur. These options might also be looked at as "way below ideal" strategies were it not for their potential to create *the preferred methodologies of the future* in situations where randomization has not thus far provided a solution, such as with labor exchange services and unemployment insurance benefits. The exploration of alternative techniques heretofore untried or untested has the further virtue of going hand in hand with promoting more experiments where they are possible, since the best way to find a trustworthy non-experimental approach is to test the contenders over and over in studies with random assignment to see how they measure up to the selection-bias-free results provided by experiments.

Smart Refinements that Minimize Movement Away from Randomized Designs

It was suggested at the beginning of this chapter that the goal of random assignment impact assessment provides a foundation for *designing and picking all forms of impact evaluation*, experimental and otherwise. We have reached the point where this assertion takes over, the point where something other than a pure experiment has to be considered for ETA's evaluation needs due to legal or technical constraints or because of inadequate commitment to experiments as the most rigorous approach possible to measuring program effectiveness.

Two principles that should guide—and historically have guided—the very best non-experimental impact analyses of labor market policies are derived from the experimental model they replace:

- *Non-experimental design principle #1 (the "stay close" principle).* Move the least distance possible away from purely random selection when other mechanisms, either naturally-occurring or induced, must be found that separate participants in ETA programs from non-participants and, thus, provide the basis for a with/without program comparison.

- *Non-experimental design principle #2 (the "stay grounded" principle).* If at all possible, imbed within the chosen non-experimental impact analysis technique at least some portion of a true random-assignment experiment to act as a check on the reliability of the main, non-experimental method.

The nearer one stays to random selection of participants and non-participants, looking at a fall-back non-experimental sample in terms of how it differs from what the ideal experiment would create, the more clearly one can understand the limitations and possible biases of the method adopted. The thought process involved in checking the conceptual or theoretic appeal of an alternative non-experimental method becomes transparent, and much more manageable, when the "Stay Close" Principle is followed. It often flows like this:

- If this were truly a randomized study, selection bias would not exist. We know it differs from a truly randomized study by allowing (besides chance) only factor P to influence who is in the evaluation's intervention-exposed "program group" and who is in the evaluation's intervention-free "comparison group." If P makes participants different from non-participants on factors related to outcomes, it will do so by causing outcomes of participants to differ in this direction and to a degree, in terms of the magnitude of the resulting selection bias, determined by behavioral reaction Q (or labor market elasticity R, or outside policy interaction S, or ….)"

By this mechanism, risks of selection bias are known and any user of the study's policy findings acquainted with how and with respect to what factors they might miss the mark. And not only are the potential threats to unbiased estimation better comprehended, they are minimized by the closest possible fealty to the experimental design ideal.

To this *conceptual* assessment of the risk of bias, the second non-experimental principle—if usable—adds an *empirical* assessment. The "Stay Grounded" Principle looks for ways to acquire a portion of the information on policy impact experimentally when not all can be obtained in that manner. Whatever other means is used (the mechanism at the heart of Principle #1) can then be applied *in parallel to the experiment* for this small component to actually *see* the degree to which non-random factor P pushes at least one measured impact away from its experimental counterpart. While such checks often fall well short of producing definitive results—if sample sizes are not large, they lack in discernment to reveal what bias may be present, and in every case their conclusions may not carry over to the portion of impact *not* derived in parallel experiments—they give added leverage in understanding the extent and potential direction of *all* the selection bias in the findings, both "testable" and "untestable."

While perhaps difficult to appreciate as an approach to finding the best-available non-experimental estimation technique when described in such generic terms, the ability of this framework to choose well a way to back off the experimental ideal will become apparent as we now apply it to the five different major policy types within ETA's purview. We begin with the policy areas where Section 3 identified a fully experimental option and established it as our primary recommendation. If that recommendation proves untenable, what will have to be changed? What is it about each recommendation that ETA might not be able to countenance and how will backing down from random assignment address this? What other strategy for measuring net impacts might be substituted for the removal of randomization?

With this experience in hand, we next apply the same two principles to the policy domains where experimentation was never possible—to the income support (i.e., unemployment insurance) and public labor exchange functions of ETA for which we have yet to make a recommendation on impact analysis strategies. Again, the search for an alternative begins by highlighting precisely what goes wrong with randomization, in keeping with the "stay close" principal: don't give ground on the experimental approach except where a specific problem cannot be solved by any other means.

The essence of our strategy in each case is summarized in the heading of the subsection, as we move from one ETA policy area to another.

Job Search Assistance and Worker Training—Allow Flexible Interruptions of Random Assignment in a Stable Set of Sites

Random assignment in a different set of nationally representative sites every two years would push DOL's pioneering use of experimental methods to a new level of accomplishment and value, at least in terms of maintained attention to large-scale impact research year after year. From the scientific point of view, there is absolutely no reason to shirk from such a comprehensive effort; a number of top-line contract research organizations in the U.S., all with experience with random assignment, could carry off the assignment successfully if ETA put forth the right authority and expectation for cooperation to all its local Workforce Investment Boards and adequately funded the technical work.

But suppose the "grind" of getting the next set of sites on board for randomization every two years (as earlier sites discontinue their one to two years of random assignment operations) proves unsustainable, given that each set is large in number—to spread around the control group—and spread throughout the nation (to be nationally representative)? What is likely to give way first? Or, put more proactively, what could be changed to relieve the pressure to discontinue the effort while sacrificing as few as possible of the virtues of the overall system?

The simplest and most helpful simplification of this system might come from *not changing the study sites each cycle*, staying with the same set of nationally representative sites for several biennium's running. The experience of most experiments—and certainly of the most similar prior evaluation, the National JTPA Study[28]—is that randomization and exclusion of selected applicants for research reasons sounds much worse in advance than it is in the doing. So sustaining random assignment for four or six years with a given local agency might not prove difficult, once the custom is in place and agency staff and referral sources grow accustomed to it. If feasible in all sites, this strategy for ongoing impact evaluation becomes logistically much more manageable.

A necessary concession to make this possible—and the one place the design would lose scientific rigor compared to the original, optimal recommendation—might be to create a *"safety valve" system for turning off random assignment—i.e., admitting all qualified applicants rather than turning away a small percentage—for brief intervals in individual sites when recruitment shortfalls become particularly difficult.* With up to two years in each cycle to build up the experimental sample, this should be possible from the standpoint of overall sample size. However, it would make the experimental sample, and hence the overall impact results, somewhat less representative of the nation's providers in all types of circumstances. Fortunately, a "by special petition only" respite system from random assignment would likely be used little in practice based on past experience, and

[28] See Orr et al. (1996) for a complete description of this path-breaking ETA-sponsored experiment. The main impact and benefit-cost findings of the project are summarized in Bloom et al. (1997).

hence detract from the comprehensive nature of experimental results only modestly. Another lesson of the National JTPA Study was that the mere *option* of relieving pressures on local agencies around random exclusions when problems become acute could greatly increase comfort and commitment to the approach in individual localities, *whether ever needed there or not.* Moreover, where interruptions in randomization did prove necessary for brief intervals, it might be possible to adjust the weights applied to the experimental data from other times and places where randomization was able to continue through periods of slow intake to partially offset this loss (i.e., by increasing the weight applied to those data).

Another concession to local concerns about random assignment might also be considered if necessary to avoid having to cycle in new study sites so frequently: *elimination of the permanent embargo* requirement on control group members recommended in the preferred design. Experience has shown that temporary exclusion from job training and job search assistance programs translates into permanent nonparticipation in almost all cases—for example, almost none of the many thousands of control group members from the National JTPA Study returned to enroll in the program at the end of their 18-month embargo period. Given the transitory nature of many workers' desire for labor market assistance, a 12-month embargo might work nearly as well.

With these adjustments, and proper recruiting and support of local agencies selected to take part in the research, a long-run nationally-representative set of sites involved in locally-small-sample randomized impact analyses is definitely within ETA's reach to establish regular, up-to-date accountability for net impacts in its job search assistance and worker training interventions.

Entrepreneurial Support—Combine Experiments Wherever Possible with Regression-Adjusted (and Experimentally Tested) Non-Experimental Designs in Additional Sites

The preferred system for evaluating the impacts of entrepreneurial support programs described in Section 3 faces issues of scale of a different sort—the possibility that few localities will generate enough interest in these services to support site-specific impact evaluations once half (or a large share) of eligible applicants are turned aside into a control group. Prior research has established that the desire and energy needed to pursue self-employment when between jobs in the salaried sector are fairly rare, leading to low participation rates in the self-employment experiments run in the 1990s.[29]

An experimental design with multiple points of randomization, as recommended in Section 3 to look separately at the value of business planning assistance and financial support will make this challenge even greater. Also, as explained there, an essential component of our strategy for keeping estimates of impact fairly up-to-date and comprehensive of the nation was to create "building blocks" of free-standing site-level impact studies whose findings could be used to simulate the range of local economic conditions confronting the nation at any time. So looking to gain a large total sample

[29] See Benus et al. (1994).

from random assignment by spreading the study over many locations, as was done in worker training recommendations, will not work here.

One way out of this dilemma would be to strengthen the appeal and visibility of entrepreneurial assistance in local communities so that sufficient numbers of applicants come forward to support multiple research groups in more places. Another response—and one exclusively in the realm of research, as opposed to program enhancement—involves living with less impact information concerning entrepreneurial programs and, if necessary, less rigorous impact information from some sites. Less information means confining the random assignment design to just a two-way split between the full package of services and none at all, rather than attempting to break out the impact of individual components experimentally. This might not be necessary in too many sites, however, given recent unpublished research showing the potential to maintain statistical precision in a three-way design with little more total sample than required by two-way randomization under certain circumstances.[30]

Less rigorous information—a sacrifice made last, and hopefully not at all—comes from abandoning random assignment in sites where entrepreneurial programs are large enough to do free-standing analysis of participants but where applicant flows cannot be brought up to the level needed for a sufficient control group as well, even with the kinds of recruiting assistance recommended in Section 3. If programs in this circumstance would appreciably add to the number of "building block" impact studies obtainable—and especially if they represent the only way to bring in certain parts of the country or certain specialized local economic conditions—combining non-experimental impact analyses there with the larger set of site-level experiments would enhance ETA's ability to understand national impacts more completely. *But only if designed according to the two principals of non-experimental impact analysis established above: "stay close" and "stay grounded".*

To do impact analysis without diverting individuals away from the program, the research must go somewhere other than the pool of approved program applicants for people that can represent the no-service "counterfactual." Statistical profiling of potential entrepreneurship candidates, when viewed from the perspective of the "stay close" principal of non-experimental design, provides a perfect opportunity to do this. It builds

[30] Unpublished Urban Institute manuscript (February 2003) shows the statistical precision gains of *delaying the division of the sample between random subsets receiving the partial as opposed to the full intervention until the partial intervention succeeds in its goals.* While developed in the context of job placement and retention efforts for welfare recipients, where those from the original treatment group who find jobs are randomized into separate partial and full-intervention groups (with only the latter getting the program's additional job retention services), the strategy translates directly to delivery of financial assistance and other later self-employment supports to only a random subset of the original treatment group members who successfully complete their business plans. Where the first step in accomplishing the intervention—here, business plan development and approval—is often not achieved or has little independent impact absent the remaining steps in the intervention, the *common* usage of treatment group cases that do not succeed at the first step in *both* intervention groups, plus the substantially larger *average* impact of later, more powerful components of the intervention (e.g., capitalization of business start-ups) can give nearly equivalent statistical power for detecting either initial or full-intervention impacts as conventional experimental designs provide for any one of the two.

on the "regression discontinuity" approach to impact estimation pioneered in the education field in the 1960s and introduced to employment and training research by Bell et al. (1995) three decades later. As the latter authors argue, the distinction between individuals screened into and screened out of voluntary employment assistance programs can be small…and by definition *is small* for marginal cases—those just barely screened out. Moreover, it by definition depends on *externally observed* traits of the individuals involved: any screening system necessarily excludes or includes based on what it measures, or on a completely arbitrary basis. The latter is the equivalent to random assignment and wholly welcome when dealing with potential selection bias on impact estimates, and the former is the consequence of exclusion mechanisms that are *fully measurable and can be modeled to perceive their consequences.* And because it depends not at all on the *self-selection* of individuals involved on difficult-to-measure motivational and personal factors, it comes closest (Bell et al. argue) to the conceptual equivalent to truly random exclusion of any systematic selection procedure can.

These considerations lead Bell et al. to recommend program "screen-outs" as a promising non-experimental comparison group for measuring program impacts. These are people who, on a measured one-dimensional scale, fall below the threshold level of suitability for inclusion that allows them to participate in the program. That suitability score becomes a key control variable in the comparison of participants to nonparticipants when estimating impacts, with the jump in outcome levels at the cut-point for program admission (the discontinuity in the regression of outcome on score) the estimate of the intervention's impact. "Profiling" of potential candidates for entrepreneurial assistance, in which an appropriateness "score" for each individual is computed and used as the basis for offering or not offering the service, fits this model exactly. It could be used or expanded in sites selected for study of entrepreneurial assistance impacts that cannot do random assignment, thereby "staying as close" to randomization as possible.

The second non-experimental design principal is also critical here, "stay grounded." This means making sure that some portion of the overall impact assessment remains experimental—in this case, likely the majority—and that *the experimental component become the foundation for testing the non-experimental portion*. All this requires is that *all* the local programs studied in examining entrepreneurial program impacts include a regression-discontinuity, profiling-based component, even (especially!) those successful in using random assignment as well. This gives a direct read-out on the success of the regression-discontinuity procedure in many settings—everywhere an experiment is actually run—as the basis for assessing its reliability in settings where nothing else is available (i.e., where the applicant flow is inadequate to conduct random assignment). Such a "reality test" is particularly important in light of Bell et al.'s finding (in a quite different welfare-to-work training program context) that, empirically, the screen-out-based approach did not approximate experimental results as well as had been hoped based on its conceptual appeal.

With appropriate testing in the experimental sites, little risk attaches to including this highly-specialized—and highly appropriate to the entrepreneurial assistance context—non-experimental method in a recurring national assessment of program impacts. And

the heart of the assessment will remain experimental even if circumstances force at least some retrenchment from that ideal.

Local Economic Development—Look for Natural "Instruments" That Create Variation in Funding Utilization from Site to Site

The greatest difficulty in the recommended approach to measuring the impacts of local economic development assistance experimentally concerns equity and political aspects of experimentally varying funding levels among otherwise similar communities. That these variations—the sole basis for inferring impacts of developmental assistance with the experimental design—need to be maintained year after year in the same cross-site pattern only heightens these concerns. Given both formula-based and more capricious factors that already lead to funding differences across communities, perhaps this is sustainable and can be kept "below radar" or justified as essential to the larger long-run national interest. But if not, where should ETA turn for other, less rigorous but otherwise "next best" impact and benefit-cost analyses?

Varying funding amounts *within sites*, offsetting a randomly-determined "down year" with an "up year" offset in the next funding cycle, is not a good solution given the incubation period for job creation initiatives and the importance of sustaining the scale of the intervention over many years to get an accurate reading of its achievements. Which leaves no choice but to not vary funding levels on a randomized basis in the first place, if found unsustainable on political or equity grounds. The experiment is gone once that happens, and the "stay grounded" principal for finding an alternative (which requires maintaining some experimental component) with it.

A number of "natural" sources of variation in the utilization of economic development assistance from one community to another then have to be considered as a means of gauging how much difference that assistance makes. One of the most promising—if things actually work this way—is *utilization variation within existing funding levels*. If some communities consistently underspend their DOL allotments, while others use them fully and perhaps a third set supplements them substantially from other sources, the equivalent of random variation is achieved—but not at random as concerns other local characteristics that may accompany these proclivities. Something that *looks like* an experiment on the surface, but is *not at all like an experiment in its mechanisms*, is not what is meant in Principal #1 by "get close."

But given its value should the determinants of funding utilization vary capriciously and idiosyncratically from place to place—or with factors unlikely to influence local employment opportunities in their own right (e.g., especially effective local Congressional lobbying for added funds, consistent administrative breakdowns in accessing and expending funds)—the first non-experimental recommendation should experimentation prove impossible in studying community assistance would be for ETA to undertake a thorough examination of expenditure-to-funding patterns among its economic development programs to see (a) how much they vary and (b) *why* they vary. Unfortunately, just as was true of deliberately engineered variation under the

experimental approach, these swings would need to go consistently in the same direction at a given locality for them to create a sustained indication of whether more development support matters; up one year, down the next, will contribute only confusion to attempts to sort out which communities got the best results from economic development inputs over time.

Should that not surface any promising "instruments" for the treatment—instruments being labor economists' term for factors affecting the nature of treatment consistently over time but otherwise unrelated to the outcomes of interest—what remains? Not much. The challenge of measuring returns to financial support of local economies has been around for centuries, and arises in many federal, state, and local government agencies besides ETA. Moreover, it has evinced a major amount of research literature pointing to many different tactics for establishing causality and credit for job growth and other neighborhood recovery benchmarks when program funds are inserted. Whether anyone will ever find a satisfactory answer to this challenge is definitely an open question absent the political resolve to randomly vary how much developmental assistance is applied from community to community. An initial step in this direction by ETA would represent a major breakthrough toward bringing accountability to an area of (across all agencies) major social investment on the basis of sound, scientific tools.

Apart from finding an effective natural "instrument" for funding variation—more of a hope at this point than a realistic expectation—nothing else ETA could undertake in this policy area is likely to help sharpen our knowledge base.

Income Support Benefits—Encourage and Synthesize Basic Research, and Consider Possible "Bump Up" Experiments

The unemployment insurance system provides two primary services—temporary income replacement for those who lose their jobs and referral/mandatory participation in reemployment services for those meeting the statistical profile of likely benefit "exhaustees". Devising ways to measure the impact of the income replacement component is quite difficult, even non-experimentally, and is ruled out entirely on a random assignment basis by legal entitlement to benefits on behalf of all eligible claimants. In contrast, both experimental and quasi-experimental impact evaluations of worker profiling may be feasible—though if this feature of the system is now universal experiments in this area would require waiver authority through new legislation.

As concerns the focus of this section—fall-back options when experiments are not possible—an extremely strong non-experimental methodology has already been applied to worker profiling by ETA in its 1997 evaluation of that system. Similar to the "regression discontinuity" strategy described above examining impacts of entrepreneurial assistance, Dickenson et al. (1997) used profiling scores to control for systematic differences between those referred for reemployment services and those not so referred because they were not quite high enough on the predicted probability of benefit exhaustion used in the scoring. This comes closest to mirroring random assignment of any naturally occurring program participation mechanism, particularly given that capacity

constraints in state reemployment services offices led to arbitrary variations in the cut-off level for referral from month to month, creating virtually random variation over time in who was referred and who was not for the group closest to the margin of exhaustion. If ETA does not use truly random variation in referral practices for future impact evaluations of worker profiling, this design is definitely worth repeating on a periodic basis to keep information current on the contribution of the profiling component of the UI system.

Measuring impacts of basic UI benefits poses much greater challenges. Most persons who are eligible for benefits claim them and, by law, must get them. Those who do not are hard to identify and track over time, and no doubt differ from claimants in fundamental ways that will affect their outcomes independently of the effects of not receiving UI payments, including differences in family circumstances, immediate employment alternatives, and longer-term career aspirations. Those ineligible for benefits due to lack of an adequate work history in covered employment have, for that same reason, different labor market hurdles to overcome and likely will experience different outcomes independently of the receipt or non-receipt of UI benefits.

These problems are well recognized in the literature on the incentive and economic well-being impacts of UI benefits and other income support programs. And they have not been solved by a host of econometric attempts to "make equal" recipients and the self-selected or program-selected non-recipients to whom they might be compared to generate impact estimates. The latest of these techniques, propensity score matching, has had spotty success in other applications where it has been tested against experimental findings (Glazerman 2001). And, absent the ability to experiment, it and the other quasi-experimental methods can never be tested directly in the UI policy realm.

More helpful, perhaps, are variations in covered employment from state to state and over time that result in UI benefit recipients and non-recipients that look very similar but who happen to be (or not be) fortunate enough to live in a time or place where they qualify for and receive benefits. Unfortunately, coverage differences among states are diminishing, and historical variations grow less and less relevant as they move further into the past—and certainly can't be counted on to create future variations in coverage. Moreover, none of these variations affect the central core of salaried and blue-collar workers in the mainstream, private-sector economy who receive the bulk of UI benefits.

In light of these difficulties, it helps to step back and consider the impacts that policy makers might wonder about in connection with basic UI benefits. Surely it is known with certainty that these benefits raise family incomes during the transitional time from one job to another, and no one needs research to decide if this aspect of the policy is worth supporting from year to year—not to presume it is or it isn't, only to say that we can be sure this is one important consequence of having an unemployment insurance system. What impact evaluation most needs to address are the potential *indirect* effects of UI benefits on duration of unemployment, quality and stability of subsequent job starts, family stress during the transition, and possible effects on worker and family health (particularly mental health). The economic "cushion" and assurance provided by UI

benefits could certainly affect all of these durations by making them longer, and could influence all other outcomes in a favorable way. But how much, and by equal amounts in all economies and states?

Two recommendations come to mind for gaining better information on the contribution of UI benefits to overall worker and family well-being:

- Continue to sponsor basic research on differential outcomes for recipients and non-recipients wherever the two groups can be found to draw contrasts between reasonably similar workers, across states, over time, by covered and non-covered employment, those just short of as opposed to meeting the work history requirements of eligibility, and even by circumstances just prior to and following benefit exhaustion. Encourage examination of family and medical outcomes as well as labor market results. Be deliberate and regular (in two-year cycles) in drawing together this research, relating it to emerging findings concerning income support programs of other types, and drawing out indicators that things are changing in important ways from one review and synthesis cycle to the next.

- Consider running some "bump-up" experiments that *randomly increase the amount or duration of UI benefits for certain recipients* in order to better study effects going in the *opposite direction* from the one of greatest interest. If legal entitlements preclude rigorous tests of whether current benefits are better then lesser or no benefits because matching experimental samples cannot be created on the "lesser or no benefits" side, check on the *other side*—where more generous benefits are not precluded by law—to see that results in improved family stability, mental and physical health, or "replacement job" quality. If it does, and this is shown with highly conclusive experimental research, then almost certainly so do benefits at their current level compared to smaller or no benefits, the policy issue of perhaps more pressing salience.

This last suggestion—possibly unorthodox and unexpected—merits at least some airing for technical merit with other evaluation methodology experts and unemployment insurance researchers…as well as a careful examination of the required scale and costs of a statistically conclusive trial intervention of this sort (it would not be cheap, given that the "treatment" is to give out additional money to probably thousands of workers). A look back at some of the earliest randomized social policy evaluations, such as the Negative Income Tax Experiments, which also focused on understanding responses over a continuum of more generous income support policy parameters to see how worthwhile they proved to be would also be sensible. All worthwhile, of course, only if policy makers really want to know what the availability of public protection against the upheavals of job loss is doing for American families…and even then likely worth testing, given its cost and indirect policy ramifications, but once.

Public Labor Exchange Services—Misleading Micro-Comparisons, Unrevealing Macro-Comparisons…and Nothing Else?

Measuring the contribution of the Employment Service, ETA's public forum for information exchange on job openings and potential workers, to the functioning of local labor markets can be approached in one of two ways. One option, developed in detail by Jacobson (1999), traces the involvement of job seekers or employers through the different layers of ES services and forms comparisons between those that get as opposed to don't get certain ones. Jacobson identifies a number of circumstances in which "treated" and "untreated" individuals are divided from one another on factors arguably unrelated to their subsequent labor market success—a close approach to making these distinctions based on random assignment.

For example, some out-of-work customers may be home when ES staff call to give them a computer-identified job referral while others may not, the latter sometimes never receiving word (or receiving it too late for the referral to still constitute a "live lead"). By reflecting the "stay close to random assignment" principle of non-experimental impact analysis, these ideas offer good promise for revealing the individual-level effects of particular ES services on selected subgroups of labor exchange users with minimal selection bias.

This approach has two drawbacks: first, similar to the strategies for looking at unemployment insurance benefits discussed above, it is spotty; it will not illuminate a comprehensive spectrum of the service types or customer groups that typify ES operations. So even if successful in producing measures of impact free from selection bias, the analyses cannot be generalized to the bulk of labor exchange activities and contributions. Still, some reliable, if limited-scope information on impacts is better than none, right?

Possibly not, if *the only depiction of impacts obtained concerns individual firms or workers*. Given the strong expectation of potentially far-ranging general equilibrium effects from labor exchange services—the possibility that a gain for one job seeker represents at least a partial loss for someone else who might have found the same job, and similarly for employers—micro-level analysis could be doomed to miss a lot of what counts in improving the overall efficiency of a highly-intertwined "macro-level" market clearing process involving untold numbers of mutually affected workers and firms. Getting a grip on these "macro" effects represents a separate, quite different strategy for judging the value-added of labor exchange services.

Market-wide analysis may be needed to have any hope of capturing comprehensively the impact of ES on local workers and employers. That can only proceed by comparing indicators of overall labor market functioning from one community to another, where the entirety of the "treated unit"—all worker-employer matches in a given community for a given job classification or skill level—is examined and compared. While one could think of doing this, using measures of efficiency such as duration of job search or turnover rates (both voluntary and, especially, involuntary terminations) in the first post-ES job,

can markets be found that *operate with different levels of labor exchange support from the public sector?* ES is to assist job matching everywhere in the U.S., and with today's Internet technology and worker mobility, likely does. Does it do so in measurably greater amounts in some communities than others? Only by finding variations in "treatment dosage" at the community level can a macro strategy for impact estimation gain any leverage on whether what's done to support labor exchange makes any difference. If the labor exchange service input is the same everywhere, regardless of variations in outcomes, we have no opportunity to associate higher or lower inputs with better outcomes.

Jacobson points out one possible reason for differential inputs at the market level: variation in distance to a job service center for different segments of the local community defined by the geography of worker residences and worksites (p. 14). Though a conceptually valid source of uneven inputs from ES, geography has become increasingly unimportant in accessing information-based interventions of all sorts in the Internet era. It would not seem wise by 2003 or beyond to invest in sorting out the physical boundaries of labor "sub-markets" in a major metropolitan areas and finding meaningful measures of labor market functioning in each when physical boundaries on labor exchange information have virtually vanished. A risky but potentially more fruitful strategy would use *measures of aggregate ES service utilization, relative to the total number of jobs in a given local economy, as the indicator of "treatment" variation* and look for patterns of better or worse labor market functioning on that basis. This is risky in three senses:

- It may surface perverse and spurious relationships, as would happen if "naturally" better-functioning labor markets result in less need for and, hence, less utilization of ES services;

- It may find contrasts in labor market results that are truly due to different local inputs, but trace the variations to the *wrong* local inputs—e.g., it may attribute shorter average job vacancy durations to higher use of ES services when the real explanation is differences in occupational mix or industrial shares in the two cities; and

- It may *still* fail to generate appreciable variation in measured ES utilization—i.e., in "inputs—to correlate with outcome levels simply because, as Jacobson points out, "job-seekers using labor exchange services do not have to register to receive services, and often do not have to identify themselves in any way to obtain referrals and other key services…[This] problem is not confined to referrals and placement, but exists for…participation in workshops, counseling, and referral to supportive services" (p. 9).

But there seems little else to work with, unless communities exist where access to job-vacancy and available-worker information has been substantially enhanced by initiatives other than ETA's Employment Service, so that some clear outliers "on the high end" of intensive labor exchange services can be identified and compared to other, more ordinary communities on indicators of labor market functioning. This tactic, if feasible, parallels that of the "bump-up" experiments proposed for UI benefits in the preceding subsection:

if the question is how things would turn out with *less* assistance, but assistance below the current ETA-funded norm can't be created in the real world, look at the effects of *more* assistance and interpret the results "in a mirror" as it were. Unfortunately, even if "bump-up" communities in terms of enhanced labor exchange information systems exist, they almost certainly don't exist in random locations matched on other factors to the conventional "ES-only" labor exchange communities to which they would be compared.

Fall-Back Options – Using What's Presently at Hand

If none of the methods for gaining good information on program impacts discussed to this point is in place for a given program or policy, what should ETA do? The question of the policy's net contribution to social objectives remains important, and the risk of misinterpreting pure outcome data as program impact information—i.e., of assuming every good outcome of a program owes its existence to the intervention—remains strong. OMB or other government agencies will still expect accountability.

Which brings us to "using what's at hand" as an evaluation strategy…clearly a "worst cases" fall-back strategy, but one to be considered before closing the chapter if only to emphasize the importance of *not winding up in this position*. Typically, an agency has four types of information at hand to document its programs and possibly give a notion of their impacts:

- Fiscal information on the expenditure of funds in relation to agency and grantee missions;

- Activity summaries of the number of persons served, quantity of services delivered in each category, and other tallies of the scale of operations per month or quarter or year;

- Process- and outcome-oriented performance indicators used for management oversight and incentive creation—items the agency officially considers to determine if grantees and other local partners are delivering good results; and

- Previously commissioned, and possible quite old, impact evaluations.

The most important thing to be said about any of these options is that *none of them are good substitutes for scientifically planned and rigorously conducted current impact studies of the type discussed to this point*. They are poor fall-backs to even the second-best alternatives discussed earlier in this section, and should never be represented as valid replacements for true evaluation in forward planning or budget requests.

But what if at times, this is it—something on the effectiveness of Trade Adjustment Assistance or extended unemployment insurance benefits must be put forward to inform a policy decision that won't wait and none of what one would like to have to consult—a truly scientific evaluation—is in place and of reasonably recent vintage? Are there any of the above sources that can play a partial role, and any particular ways of looking at them

or realigning or adjusting the information they provide to get something closer to a measure of net impacts?

The safest course is to go with old but reexamined impact studies. These *will not give today's answers with today's (or even last year's) numbers*, but they often address questions that remain salient about programs that change less over years or decades than one might suppose from amending legislation. What the National JTPA Study and unemployment insurance job search/reemployment experiments still tell us about training impacts and work incentives has strong meaning for policy choices today, though they are not in any sense a "report card" on today's WIA system or the efficiency of the income replacement mechanisms under current UI rules. The same will likely be true of the experimentally-designed National Job Corps Study for some years to come.

Though we cannot reasonably "update" the conclusions of those studies, or others like them, to today's circumstances through statistical adjustments to their impact estimates, we can remind ourselves of their lessons balanced against a careful detailing of what is different in the programmatic terms, in labor market conditions, and in the worker populations affected than was true when that research was conducted. This is certainly useful, valid, and (perhaps most importantly) *transparently limited in important ways* as a guide to today's policy choices. For this last reason alone, it creates little risk of over-interpretation or a rush to action based on over-reading the amount of knowledge truly available to justify policy action.

Precisely this risk puts an unsatistfactory cast on the other information types already at hand that might be seen as alternatives to formal evaluation. None of the conventional measures of program expenditures, operational activities, or performance (typically measured by customer outcomes) purport to consider *what the results would have been visa vie the program's policy goals absent the intervention.* When there is no counterfactual, and no standard of "you have to do better than nothing for the customers you serve", the temptation is inevitable to over-read the numbers regarding program accomplishments as essential to the positive outcomes observed. Even WIA performance measurement systems that require information on earnings for individuals prior to program entry (e.g., for dislocated workers prior to dislocation) do not provide reasonable counterfactual information, given what is known about shifts in lifetime earnings trajectories around the time of program entry independently of the intervention's effects.[31]

[31] See, for example, Heckman and Smith (1999). The other common means of using data on program participants to remove other influences on success besides the intervention under consideration—regression adjustments to take account of workers' background demographic and labor market characteristics (e.g., the performance standards system used in administering JTPA in the 1980s and 1990s)—do not attempt to simulate a "no intervention" counterfactual. Rather, they attempt to equalize non-program factors in a set of people all of whom receive the intervention. Their success in this regard has been quite questionable in any case. The large literature on the economic incentives and econometrics of the JTPA performance standards has been summarized most recently by Barnow and Smith (2002).

In this sense, formal performance measurement systems that specify and monitor key consumer outcomes[32] may be worse than older, more accounting and process-focused management tools in seeming to say which programs are working well and which are not relative to the ultimate social goal. Yes, they track results in relation to that goal—a critical step toward establishing accountability for effectiveness—but they don't measure the *contribution of the intervention* independent of other factors that enable good results to emerge[33] like individual persistence, changes in labor market circumstances over time, and the role of other non-ETA (and even nongovernmental) forms of assistance.

Developing Better Tools for Future Impact Studies

A final area of recommendation—and one clearly needed for such vexing evaluation challenges as finding ways to measure impacts of community-wide interventions with strong general equilibrium effects—concerns the development of better non-experimental and experimental tools for the future. ETA has long been a leader among government agencies, both domestically and internationally, in encouraging academicians and professional evaluators to find the best possible techniques for accurately attributing social outcomes—in this case, labor market and worker performance--to policy interventions and other causes. Its support for random assignment experiments has contributed substantially to the ascendancy of that technique over the last 20 years, and many of the best recommendations in this chapter owe something to the foresight and commitment to scientific rigor that characterized that leadership.

Further pioneering patronage of methodology development and methods testing will under-gird the next generation of evaluation recommendations if that commitment remains in place. This means not merely (merely!) doing the best evaluations scientists know how to do with today's research technologies when policies need to be examined, but continuing to "piggy-back" on its major evaluations methodological work to identify other techniques capable of giving reliable results on program impacts, costs, and benefits.

Repeatedly, this treatise has emphasized the importance of *checking the success* of quasi-experimental impact estimation techniques against experimental data, to see if other ways can be found to eliminate selection bias besides random assignment. This remains necessary since, as we have seen, important portions of ETA's policy portfolio remains unassailable through experimental methods, such as the value of entitlement benefits in the unemployment insurance system and the general equilibrium effects of public labor exchange services. And the goals in improving evaluation methods go beyond removing selection bias to finding better ways to approximate nationally representative results

[32] Social Policy Research Associates (2002) provides a summary of the performance measurement components of greatest current importance at ETA, the reporting and monitoring requirements of WIA.

[33] This is seen as a problem even in the relatively underdeveloped policy evaluation milieu of Europe. In their OECD summary, Martin and Grubb (2001) note that "The most common method of 'evaluation' [in Europe] still consists of simply monitoring the labour market status and earnings of participants for a brief period following their spell on a programme. While this sort of exercise provides useful information, it cannot answer the vital question of whether the programme in question 'worked' or not for participants" (p. 21).

("external validity", as opposed to "internal validity"), to shorten lag times in getting impact and benefit-cost results to policy makers, and to track down the causes of site-to-site variation in program effectiveness…all without losing ground in the battle against selection bias.

With an increasing number of nationally-representative social experiments underway or in the planning stages within the federal government, the methodological "gold standards" for achieving external validity and tracing the causes of cross-site variations in impact may form over the next decade—and with them the opportunity to test less state-of-the-art designs against them. ETA would be well served long-term by including in any study that meets the "gold standard" of internal validity, external validity, or cross-site attribution a test of the lesser methods that may continue to be relied upon in other policy areas.

A final role to be pushed as part of the strategic planning process is ETA's options for encouraging new creative thinking on impact assessment paradigms. For all the progress made and the breadth of applications and innovations achieved in formulating evaluation strategies from existing tools, the impact analysis field remains bound up in "with/without" comparisons of one form or another. This leaves one comparatively hamstrung when the world turns out to be entirely "with" for certain key government policies, as in the case of the Employment Service and unemployment insurance benefits. Other means of getting at impacts—perhaps empirically weak, at least at first—that conceptually can break free of these bounds need to be conceived and pushed toward a "doable" prototypes. Jacobson (1999) mentions one of these in his paper on evaluating labor exchange services, suggesting that "it is not far-fetched to believe employers can provide accurate information about the value-added [to them] of labor exchanges" (p. 15). Bell (2002) has suggested others at a recent welfare reform evaluation conference, including locality-specific micro-simulation models to create counterfactuals and "thought experiments" among stakeholders in a policy to map out possible consequences, costs, and benefits in a laboratory setting.[34]

The point of these "blue sky" notions is not that any of them will necessarily supply viable solutions to today's intractable evaluation problems. Rather it is that they are, simply put, within the context of today's impact evaluation technology indeed "far-fetched". But so too were social experiments when they first entered the discussion decades ago. Anything DOL and ETA can do to encourage that discussion at the extensive margin of new methodologies over the next decade will better equip it for successful five-year planning on its program impact information needs ten years from now.

[34] Stephen H. Bell, presentation to the U.S. Department of Health and Human Services' Fifth Annual Welfare Reform Evaluation Conference, Arlington, VA, June 13, 2002.