Program Integrity Studies

A program's potential for impact can be affected by any diversion of program funds from their intended purpose, and program integrity studies focus on the question of whether such diversions exist. Investigative or enforcement arms of the government often pursue program integrity through detailed investigation of a small number of suspect situations, as in a recent examination of the CACFP carried out by the USDA Office of the Inspector General (USDA, 1999). Such investigations can lead to prosecutions and remedial actions, but they usually do not produce general estimates of the prevalence of problems or the overall percentage of funds diverted. Complementary research using surveys or administrative data from representative samples is therefore often needed.

Impact Evaluation of Demonstrations

The preceding sections considered potential strategies for evaluating ongoing food assistance and nutrition programs, with an emphasis on entitlement or saturation programs that have been operating at substantial volume for over two decades.

We turn now to evaluations of "demonstration" or "pilot" programs. These demonstrations typically represent policy initiatives that are to be tested and examined on a limited scale before full-scale implementation. The intervention may be an entirely new program, but it is more commonly a significant modification to an existing program. Past examples include demonstrations of cashing out food stamps, requiring education and training for food stamp recipients, and delivering food stamp or WIC benefits through electronic benefit transfer.

Perhaps the largest set of examples of demonstration impact evaluations consists of the waivers obtained by more than 40 States from the U.S. Department of Health and Human Services to demonstrate the feasibility and effectiveness of State-proposed changes in rules for the Aid for Families with Dependent Children (AFDC) program. The changes ranged from imposing time limits on AFDC benefits to capping benefits upon the birth of additional children. Many of the waiver changes involved requiring preparation for employment and mandating job searches. Most of the waivers were evaluated for impact by randomized experiments in which the experimental groups proceeded under the changed rules and the controls continued under existing AFDC regulations.

Three distinguishing features of a demonstration lead to evaluation strategies that differ from those for ongoing programs:

- The intervention is new. In principle, evaluation activities can begin at the same time as implementation of the demonstration, or even before.

- The intervention has not been mandated by law for the entire program or service population.

- The intervention is applied to a restricted number of participants. During the relevant periods, some potential targets will be subject to the intervention and some will not.

These features generally make it much easier to identify a Counterfactual in a demonstration than in ongoing programs. In particular, the absence of a legal entitlement and saturation volume remove the main obstacles to randomized experimentation, which make this the preferred impact evaluation design. Nevertheless, some circumstances require quasi-experiments, as discussed below.

Randomized Experiments

In evaluating a demonstration intervention that modifies an existing program, the intervention's impact is normally defined as the difference between outcomes with the new intervention and outcomes with the preexisting version of the program. The Counterfactual is the status quo; the control subjects experience the usual program services but are not offered the new services incorporated in the intervention. For example, the several demonstrations of cashing out food stamps estimated the effects on food purchases of receiving benefits in the form of checks rather than in the form of food stamps. They did not estimate the overall impact of subsidizing food purchases.

Strengths and Limitations of Randomized Experimentation in a Demonstration

The randomized experiment is the strongest design available for evaluating demonstration interventions. The findings of such an evaluation are considered substantially more reliable than findings from even the strongest of the quasi-experiments. If a randomized experiment can be devised that addresses the policy question and is operationally feasible, this is the preferred choice.

Nonetheless, the design does have some conceptual limitations that need to be considered in approaching any demonstration effort. One limitation stems from the rather obvious point that the experiment measures only impacts that occur after the point of random assignment. In the welfare waiver demonstrations mentioned above, families were randomly assigned after they were already receiving welfare benefits or as they were first approved for benefits. If the intervention caused recipients to leave welfare earlier, the experiment would measure that impact. But if the intervention caused fewer people to apply for welfare benefits, or caused different kinds of people to apply, that effect would occur logically prior to the point of random assignment and would not be measured.

In order to capture the intervention's effect on application behavior, the experiment would have to randomly assign families not currently receiving welfare and make sure that the experimental group was told about the intervention and the control group was not. This approach is feasible and has been used in demonstrations of new programs. It is less applicable for modifications of existing programs, where the community of potential participants already has substantial program information and active communication channels.

Another limitation of randomized experimentation occurs when part of the impact may be determined by people or situations other than the randomized subjects responding to the intervention. For example, imagine an intervention in which FSP recipients are given vouchers for particular food items, redeemable at the shelf price of the item. In communities with large FSP populations, the higher demand might lead to a general price increase for the specified items. The control group, facing the higher price, might reduce their consumption of those items, leading to an inflated difference between the experimental and control groups and an overestimate of the intervention's actual impact.

The Debate About Whether To Randomize in a Demonstration

Despite the obvious (to a researcher) appeal of the randomized experiment, this design is not commonly used to evaluate food assistance and nutrition program demonstrations. Major examples are limited to evaluations of food stamp cashout (Fraker et al., 1992; Ohls et al., 1992), food stamp employment and training (Puma et al., 1990), and, recently, innovations in WIC nutrition education (Randall et al., 1999) and the SBP (Abt Associates Inc., 2000). There are also a few examples of small-scale randomized experiments carried out in the early years of the WIC program (e.g., Metcoff et al., 1985), when many eligible pregnant women could not be served because of limited WIC funding.

Some of the most common objections to random assignment are noted below.

- **Administrative burden.** Because random assignment is usually implemented within the caseload of the local service delivery organization, it asks more of program administrators than most other types of evaluation. This is essentially a matter of costs, so the problem can be addressed by supporting additional staff time to carry out the evaluation's requirements.

- **Ethical concerns.** Program operators often argue that, since the demonstration benefits or services are in limited supply, they should be allocated on the basis of the potential participant's need or potential for benefit (as judged by the program operator). Of course, this assumes that the service is beneficial—exactly the proposition that the evaluation is supposed to test. Another argument favors first-come, first-served allocation as most "fair." In both cases, random assignment can be argued to be at least as equitable and reasonable a way of rationing services.

- **Evaluation cost.** Randomized experiments are often more costly than other forms of evaluation. Most of the higher cost, however, results from a greater commitment to long-term followup and careful measurement of multiple outcomes. The cost associated with randomization itself is usually minor.

- **Caseloadwide interventions.** Some interventions can be implemented only at higher administrative levels than the individual participant. Although random assignment can theoretically occur at levels such as the office or county, a large number of units must be randomly assigned. This can become infeasible, especially if implementing the intervention in each unit is costly.

The debate about random assignment is only sometimes won on its merits. Not many people, especially those outside the research community, have an intuitive understanding of how much reliability is gained by randomized experiments or how much cost or hassle they actually entail. Popular understanding does seem to be increasing somewhat, perhaps because of the well-publicized use of experimentation in pharmaceutical trials. Until that understanding becomes more widespread, however, random assignment will be used less often than it should.

Unit of Randomization

Most demonstration interventions are implemented at the level of individuals, families, or households. Sometimes, however, the target of an intervention is a higher level organized unit. Within the NSLP or SBP, for example, one could imagine randomization at the level of the student, the school, or the school district, depending on the nature of the intervention. The statistical models appropriate to randomized experiments using large organized units are ably discussed in Murray (1998).

A prime example of a randomized demonstration relevant to food assistance and nutrition programs is the Child and Adolescent Trial for Cardiovascular Health, or CATCH study (Luepker et al., 1996). The demonstration involved 96 elementary schools located in California, Louisiana, Minnesota, and Texas. Fifty-six schools were randomly chosen to be intervention sites and 40 to be controls. Over 5,000 children who were in third grade at the start of the demonstration participated over a 3-year period, 1991-93, or until the fifth grade.

The intervention included training sessions for food service staff and teachers, changes in the curriculum for students, and efforts to reach parents of participating students with information about the importance of nutrition and physical activity.

Of particular interest is the training given to foodservice personnel, which consisted of 1-day sessions at the beginning of each school year and monthly visits and additional "booster" sessions as needed. The training sessions focused on ways in which menus and recipes could be changed to decrease levels of fat and saturated fat and to increase fruits, vegetables, and grains. Baseline measures of both the nutrient content of school meals and students' actual food intake were taken in 1991 and were used, in conjunction with followup measures taken in 1994, to gauge change over time.

The analysis showed that by 1993 the total energy provided in lunch meals declined in the intervention schools, whereas there was a slight increase in the control schools, leading to a statistically significant difference between the two at the end of the trial. Similar statistically significant differences favoring intervention schools were found with respect to the percent of food energy obtained from total fat and saturated fat.[13]

Another example of randomization at the level of the school is an evaluation of providing universal free breakfasts in the SBP, which is in its early stages at the time of this writing (Abt Associates, 2000). In each of 6 school districts around the country, 12 matched pairs of schools were identified and randomly assigned to the treatment or control group. Treatment group schools will offer free breakfast to all students without means-testing, and control group schools will operate the SBP as it currently exists, with means-testing for free and reduced-price meals.

Demonstrations with organized units as targets tend to be more costly than those randomly assigning individuals or families. Such demonstrations cannot be accomplished with just a handful of targets.[14] Furthermore, obtaining the willing cooperation of organizations is often difficult. Of course, when the objective of a demonstration is to change the ways in which organizations operate, there is no alternative to such designs. When alternatives exist, however, it is usually cost-effective to choose the most disaggregated, yet feasible, unit.

Participation vs. Intention To Treat

Randomization ensures that experimental subjects and control subjects are comparable at the outset of the demonstration, but selection processes often come into play thereafter. In most demonstrations, some experi-

[13] CATCH's primary objective was to improve the cholesterol levels of children in the experimental schools, but no improvement was found at the end of the experimental period.

[14] The numbers needed in the experimental and control groups can be determined by power calculations (Lipsey, 1990), which take into account the size of the expected difference in outcomes between the two groups. Although there are examples of large unit demonstrations with as few as 5 to 10 units in each group, ordinarily the required numbers are much greater. In those examples, the effects expected were quite large (relative to the variance of the outcome measure). Small expected effects require larger sample sizes to produce reliable findings.

mental targets never actually receive the demonstration treatment. Sometimes this occurs because people leave the program immediately after random assignment. Additionally, some interventions require action by the program participant, such as attending a diagnostic or service session, and some people never take the required action.

At the end of the observation period, then, the average outcome for experimental group subjects is not the same as the average outcome for those who actually received the intervention treatment, usually regarded as the "participants." Nonetheless, it is the full experimental group that must be compared with the control group. Limiting the comparison to participants introduces an opportunity for selection bias, which random assignment is designed to avoid.

Comparable selection processes affect control group members. They may enroll in some alternative programs with intended outcomes similar to those of the demonstration program. This means that demonstration experiments are, strictly speaking, tests of the effects of "intention to treat." This will depart from the effect of receiving treatment to the extent that the intended treatment is not received by experimental group members and is received from competing programs by control group members.[15]

Testing the intention to treat, rather than actual receipt of services, is more often an advantage than a disadvantage. The relevant policy question is how much difference the program can make for the people it is intended to serve. If a program's ability to affect the target population is limited because people do not enroll or drop out, or because they would have gotten the same services without the program, the policymaker needs to know this. The policymaker also needs to know about subgroups of the target population—in terms of both differential participation rates and differential program impacts.[16]

For voluntary interventions, the researcher must decide at what point to conduct random assignment. If all targets are randomly assigned, the impact is measured for the full target population. If randomization occurs as people volunteer, the impact is measured for volunteers, and people who never actually receive the treatment may still be included. Either choice can be appropriate, depending on the nature of the intervention and the policy questions of greatest interest.

Complex Random-Assignment Designs

Demonstrations often consist of a "bundle" of conceptually separable interventions. The WIC program, for example, can be viewed as a combination of supplemental foods, nutrition education, and health and social service referrals. A simple experiment would test the impact of the WIC treatment "bundle"—i.e., it would measure the effects of the program overall, but would not separately estimate the effects of each component.

Complex forms of randomized experiments attempt to unbundle the treatment by forming more than one experimental group, with different experimental groups receiving different interventions or combinations of interventions. A complex WIC demonstration designed to test the separate effects of supplemental foods and nutrition education might have two experimental groups. One group would receive both supplemental foods and nutrition education, and the other would receive only supplemental foods. A control group would receive no WIC benefits.[17] Comparing average outcomes in the two experimental groups could show whether WIC nutrition education had effects over and above the effects of the supplemental foods.

Complex experiments are sometimes designed to measure the effects of varying the treatment "dosage." These experiments provide useful information such as whether the outcome response function is linear or has some curvilinear form. A complex WIC dosage experiment might vary the amounts of food provided in WIC food packages or the amount of WIC nutrition education provided, in an effort to determine how outcomes are affected by dosage.

[15] When the amount of nonparticipation becomes worrisome, statistical analysis can often attempt to compensate by constructing instrumental variables, as in quasi-experiments (e.g., Ludwig et al., 1998.)

[16] As policies change, the definition of the intended target population often changes as well. This argues for the research design to incorporate the broadest relevant definition of the target.

[17] This design assumes that there is little or no policy interest in whether nutrition education alone is effective. If there were such interest, one would add another experimental group that would receive only nutrition education. This expanded experiment would permit the estimation of the net effects of nutrition education.

Maintaining the Integrity of the Experiment

Numerous events can undermine the integrity of the experiment and therefore the reliability of the impact estimates. The most important ones to bear in mind during the design process are:

- **Nonrandom assignment.** If local program staff perceive the demonstration benefits to be sufficiently important, they may try to influence the assignments. For this reason, the actual assignment is usually performed under the control of the researcher or a central operating agency. Formal, in-process reviews of randomization are usually needed.

- **Contamination.** Over time, control group members may be erroneously given program services that should be limited to the experimental group, and vice versa. Periodic reviews of local program operating procedures and sample case folders are needed to monitor contamination (sometimes called "cross-over").

- **Attrition.** Loss of subjects from the original experimental and control groups may result from causes unrelated to the research (e.g., moving out of State, institutionalization) or from research problems such as survey nonresponse and unlinkable administrative data. Although there are no hard-and-fast rules, a rough rule of thumb is that attrition from all sources must remain below 30 percent for experimental results to be credible, and any level above 10 percent calls for an analysis of nonresponse bias. Because outcomes are often measured in surveys conducted long after the subjects leave the program, strong survey designs are essential.

- **Policy changes.** Multiyear experiments often encounter policy changes that alter the experience of either the experimental group or the control group. These must be examined to determine whether they require some modification to the design or analysis.

The integrity of the evaluation can also depend on how well the experimental and control subjects understand the policies that apply to them, which is not always easy to control or even to know. In some of the welfare waiver experiments, the overwhelming majority of welfare recipients were subject to the new rules, and only control group targets were subject to the old AFDC rules. This meant that only a small percentage of State welfare recipients were to be treated in special ways, a condition difficult to maintain over the several years that the experiments were run.

In the experimental evaluation of the New Jersey Family Cap Demonstration (Camasso et al., 1998), for example, it was discovered 2 years into the experiment that the majority of members in the control group wrongly believed that their benefits would not increase if they gave birth to additional children. It is not clear whether this resulted from a failure of the welfare agency to inform control group members adequately, or whether intense media attention to the family cap provision effectively drowned out the welfare agency message. In any event, the failure of the control group members to understand that their incentives to avoid additional births were different from those in the experimental group diluted seriously the contrast between the experimentals and controls.

Quasi-Experiments

As in impact evaluations of ongoing programs, it is often necessary to use quasi-experimental designs to assess the impact of program demonstrations. In this section, we describe three quasi-experimental designs that are commonly used in this context. Several other less common designs are also mentioned. All of the quasi-experimental designs are similar in structure to those described in the preceding section on evaluating impacts of ongoing programs.

Quasi-Experiment 5: Comparing Demonstration and Comparison Sites Before and After an Intervention

In this design, the demonstration intervention is implemented in designated "sites," where a site is typically a local operating entity or jurisdiction such as a food stamp office, a WIC clinic, or a school district (see box). A companion set of sites, which will continue to operate the program under the nondemonstration rules, is chosen to provide the comparison group.

Outcome measures are taken in both demonstration and comparison sites at two or more points in time, with at least one measurement occurring before the intervention is implemented and one after. Measurements are taken for separate samples in each site in each time period. The evaluation compares successive cross-sections, rather than using a panel design, in order to have representative samples of the participant population at both points in time. Although the sample

Quasi-Experiment 5 Demonstration vs. Comparison Sites, Before and After

Features:

Impact estimate: Difference between outcomes for program participants in demonstration locations and outcomes for program participants in nondemonstration locations, subtracting out pre-demonstration differences in participant outcomes.

Key requirement: Multiple demonstration and nondemonstration sites.

Advantage: Strongest of the quasi-experiments.

Disadvantage: Most costly of the quasi-experiments.

The Three Questions:

Alike before exposure? Similar, and the difference is subtracted out.

Difference solely from intervention? Mostly. The two kinds of locations might have different events between pre- and post-demonstration measures.

Full force of intervention represented? Yes, if demonstration fully implements planned intervention.

is drawn from the pool of active program participants at the selected time points, outcome measurement may occur either immediately or at some later time, when the impact is expected to have occurred fully.

Impacts are estimated through multivariate modeling. The outcome measure is modeled as a function of whether the intervention was in place, location (site), time period, and individual or household characteristics potentially related to the outcome.

The Nondemonstration Program as the Counterfactual

In evaluating modifications to existing programs, the Counterfactual is the preexisting version of the program, which is also the version of the program that exists in locations where the demonstration is not being implemented. The quasi-experiment therefore represents the Counterfactual not with nonparticipants, but with participants in other locations or pre-demonstration periods.

This distinction generally means that quasi-experiments are stronger for evaluating modifications to ongoing programs than for evaluating the ongoing programs themselves. Consider a demonstration such as food stamp cashout. It is easy to believe that the people who participate in the food stamp program in a county where the demonstration exists would closely resemble participants in the neighboring county. Even if the demonstration has some influence on participation, most of the same people would be food stamp recipients with or without the demonstration.

In contrast, it is more difficult to believe that people who are income-eligible for food stamps, but choose not to participate, closely resemble the actual food stamp recipients. But those nonparticipants are used to represent the Counterfactual in Quasi-Experiment 1 and, to a lesser degree, in Quasi-Experiment 3. Other things being equal, then, quasi-experimental designs in which all groups consist of program participants probably yield more reliable results than those in which program participants must be compared to people who could be participants but are not.

Selecting Sites. The greatest vulnerability of this research design lies in the possibility that the comparison sites do not adequately represent the Counterfactual—that is, outcomes in the comparison sites differ from the outcomes that would have been observed in the demonstration sites if there had been no demonstration. The pre-demonstration measurements help limit this vulnerability, allowing the researcher to account for between-site differences that existed even before the demonstration began. But sites can also differ in the trajectory they follow between the pre- and post-demonstration periods. For example, if the comparison site enjoys an economic growth spurt while the demonstration site suffers a sharp downturn, participant outcomes in means-tested programs may not be comparable.

Minimizing this vulnerability requires multiple demonstration and comparison sites. There is no fixed prescription for the number of sites, and the actual number usually reflects a tradeoff between cost and reliability. To achieve statistical generalizability to the

U.S. population of program participants would require a very large number of sites, probably in the range of 20-40 demonstration sites and a roughly equal number of comparison sites. Because the cost of implementing an intervention is typically a direct multiple of the number of sites, most demonstrations adopt less lofty ambitions. They attempt to choose just enough sites so that a single "bad" site will not severely distort the findings. A design with 5 to 10 demonstration sites and an equal number of comparison sites is generally considered to meet this criterion.

Minimizing vulnerability also requires that the demonstration and comparison sites be as well-matched as possible. Four dimensions are generally important for food assistance and nutrition programs: the administrative regime, the economy, population demographics, and cultural or geographic factors associated with dietary patterns. With respect to the administrative regime and the economy, it is desirable to select demonstration-comparison pairs that are likely to be affected equally by any policy changes and economic shifts that may occur during the study. This usually argues that pairs be matched within the same State and, if possible, within the same regional economy.

Site randomization is sometimes used within comparison site designs. Matched pairs of sites are selected, and one site in each pair is randomly assigned to implement the demonstration intervention. This procedure protects against the possibility that program administrators will choose only "good" sites for the demonstration. It does not, however, ensure comparability of the demonstration and comparison groups in the way that a randomized experiment does. Randomization ensures comparability only when quite large numbers of units are randomly assigned. Thus, even though the design prevents administrators from assigning the good sites to the demonstration, random assignment with a small number of sites can yield the same result by chance.[18]

Quasi-Experiment 6: Simple Comparison of Demonstration and Comparison Sites

This design is essentially the same as Quasi-Experiment 5, omitting the predemonstration measurement of outcomes. The demonstration is implemented in selected sites, and each site has a matched comparison site. Program participants in the comparison sites represent the Counterfactual for participants in the demonstration. Outcomes are measured for participants in demonstration and comparison sites at the same time, which may be while they are actively participating or after they have left the program. Impacts are estimated in multivariate models that include presence of the demonstration, site, and participant characteristics (see box).

Quasi-Experiment 6
Demonstration vs. Comparison Sites

Features:

Impact estimate: Difference between outcomes for program participants in demonstration locations and outcomes for program participants in nondemonstration locations.

Key requirement: Multiple demonstration and nondemonstration sites and strong modeling.

Advantage: Administratively easy.

Disadvantage: Limited reliability.

The Three Questions:

Alike before exposure? Similar, at best.

Difference solely from intervention? No. Site differences may be important.

Full force of intervention represented? Yes, if demonstration fully implements planned intervention.

This design is much weaker than Quasi-Experiment 5 because it is highly vulnerable to preexisting site differences. Program participants in one site may have different nutrition and health outcomes than participants in another site for reasons that existed long before the demonstration began. The multivariate model adjusts for differences associated with those individual characteristics for which data are available. Any differences stemming from site-level forces (such as the differing effectiveness of local program staff) may be confounded with the effect of the program.

[18] If four pairs are assigned by chance, the probability is around 6 percent that the "good" site in all four pairs will be assigned to the same status.

The only way to limit this vulnerability is to include numerous demonstration and comparison sites in the design. In general, more demonstration and comparison sites are needed when the design omits the pre/post dimension included in Quasi-Experiment 5. Thus, if 5 to 10 demonstration sites would be used for Quasi-Experiment 5, 10 to 15 would be recommended for this design (Quasi-Experiment 6).

Quasi-Experiment 7: Demonstration Targets vs. Comparison Targets, Before and After

Quasi-Experiments 5 and 6 respond to situations where the intervention being tested is a modification of an existing program. The same general research structure is applicable when a new program concept is being tested in a limited number of locations, but program sites cannot be used to represent the Counterfactual because the only program sites are those of the demonstration itself (see box).

Quasi-Experiment 7 Demonstration Targets vs. Comparison Targets, Before and After

Features:

Impact estimate: Difference between outcomes for a demonstration target group and a comparably defined nondemonstration group, subtracting out predemonstration differences.

Key requirement: An identifiable target population that incorporates all demonstration participants and not too many nonparticipants.

Advantage: Strongest quasi-experiment for a new program demonstration.

Disadvantage: Sometimes impossible to find an efficient target population (one with few nonparticipants).

The Three Questions:

Alike before exposure? Similar, at best.

Difference solely from intervention? Mostly, but site differences may be important.

Full force of intervention represented? No, diluted to the extent that not all members of demonstration target population are reached by the demonstration.

Under Quasi-Experiment 7, the researcher begins by defining a demonstration target population. The target population is normally defined in a way that reflects program eligibility criteria. Four target populations must be identified using the same definition: the target population in the demonstration sites during the demonstration period; an equivalent population in the demonstration sites before the demonstration begins; and equivalent populations in nondemonstration sites during the same two time periods. Demonstration participants constitute a subset of the target population in the demonstration site and the demonstration time period.

Outcomes are measured for all four populations. Impact on the target population is estimated in a model that includes presence of the demonstration, time period, location, and individual characteristics.

Defining and Using the Target Population. The greatest design challenge in a demonstration of a new program is finding an appropriate group to represent the Counterfactual. The researcher cannot normally assume that participants in any existing program closely resemble the people who will participate in the new program. Therefore, it is necessary to find some nonprogram population that constitutes an adequate comparison group.

Although the researcher's first choice would be to define a target population that is the same as program participants, this is rarely possible. It occurs only when the new program will be applied universally to a category of people who can be clearly identified in the absence of the program. School-based programs provide the most ready examples. Imagine a demonstration testing a new nutrition education program, where the full program will ultimately be implemented on a mandatory basis in all seventh grade classrooms. During the demonstration, selected classrooms implement the new program. Students in those classrooms constitute both the target population and the participant population for the demonstration site and time period. Students in other seventh grade classrooms make up the comparison group target population. The prior year's students in those same classrooms become the two pre-demonstration target populations. In all four situations, it is assumed that all students would be participating in the program if it were offered.

More commonly, the researcher must work with a target population that is defined more broadly than the participant population. Suppose that the example above concerned a nonuniversal program, in which certain seventh grade students volunteer or are selected to receive special nutrition education. Some students in the demonstration classrooms participate in the program and some do not. The researcher does not know which seventh graders in the nondemonstration classrooms would be comparable to those who actually participate in the demonstration. The design must therefore compare target populations rather than participant populations: all students in the demonstration classrooms must be compared to all students in the comparable seventh grade classrooms.

An evaluation based on a target population will necessarily find a smaller average impact than one based on demonstration program participants, assuming that the program does not affect nonparticipants. The measured impact for the target population is the weighted average of the impact for participants and the (zero) impact for nonparticipants. It is convenient for the researcher if the demonstration's target population is defined narrowly, which will reduce the proportion of nonparticipants and yield a clearer estimate of the demonstration program's effect.

It is important to note one unacceptable design that is sometimes suggested: comparing demonstration participants, rather than the target group that includes demonstration participants, to a target population in a nondemonstration area. The target population includes some people who would not participate (unless all members of the target population are required to participate, as in the example above). Comparing the participant and target populations introduces selection bias. The direction and magnitude of the bias are unknown, and the design provides no opportunity to correct for the bias.

Estimating Effects for Participants. Because the impact for target populations understates the impact for participants, and because the magnitude of the understatement can vary from one study to the next, it is desirable to attempt an estimate of the effect for demonstration participants. The attempt must be cautious, and the result must be accompanied by caveats, however.

A simple but sometimes risky approach is to inflate the estimated impact according to the ratio of participants to targets. If the demonstration has zero impact on nonparticipants, and if nonparticipants make up half of the target population, the impact for participants must be double the impact estimated for the whole target population.

This approach assumes that the demonstration has zero effect on nonparticipants. The assumption may not hold if, for example, information about the demonstration is provided to other members of the target population. In the earlier example, if some students in the classroom are selected for special nutrition education, others may become interested in the topic and alter their behaviors.

In such an instance, estimating the demonstration effect on participants requires modeling participation. The instrumental variables approach described earlier is appropriate for this situation. Other modeling approaches are sometimes used to define "probable participant" subgroups within each of the target populations, and then estimate impacts separately for probable participants and probable nonparticipants.[19]

Other Quasi-Experimental Designs for Evaluating Demonstrations of New Programs

Three other quasi-experimental strategies, all representing minor variations on designs discussed previously, are worth mentioning as candidates for evaluating demonstrations of new programs. Two of the designs—participant vs. nonparticipant before and after, and time series analysis—are reasonably strong designs, but cannot often be applied to new program demonstrations. The third design, demonstration vs. comparison target populations, is a weak design that would rarely be recommended.

Comparing Participants to Nonparticipants, Before and After Program Participation (Quasi-Experiment 3). In this design, a new program demonstration is applied to a defined target population. Some members of the target population participate, and some

[19] It is important to estimate the impact for both the probable participants and the probable nonparticipants. Imperfections in the participation model can lead to a situation in which, for example, a substantial positive impact is estimated for probable participants, but a significant negative impact is estimated for probable nonparticipants. If the program cannot logically have a negative impact on nonparticipants, the implication is that the estimate for probable participants overstates the real impact for participants.

do not. The researcher obtains outcome measures for both the participants and the nonparticipants at a time before the demonstration begins and at a time when the impacts should be visible. Controls for selection bias are required in impact estimation.

One example of this design is a study of the SBP carried out by Myers and colleagues (1989). In 1986, the Massachusetts legislature required the introduction of the SBP into schools in which 40 percent or more of the school lunches were served free or at a reduced price. Myers took advantage of the fact that six of the elementary schools in the Lawrence, MA, school district were affected and that this district routinely gave standardized achievement tests.

The researchers compared scores on the Comprehensive Tests of Basic Skills administered in April or May 1986 with scores for the same students in 1987 (after the School Breakfast Program had been in place for about 3-4 months). The students consisted of all children in six elementary schools in grades 3-6 who were eligible for free or reduced-price meals and who were in the schools for the second semesters of 1986 and 1987. Scores for those who participated in the program were compared with those of eligible nonparticipants. Participants were defined as those who ate a school breakfast at least 3 days during the same week that the tests were administered. Using multivariate analyses that adjusted for children's characteristics, significant positive effects of SBP participation were found for total test battery scores, absences, and tardiness, but not for language, math, or reading.[20]

This design is rarely applied because of the requirement for measuring participant and nonparticipant outcomes before the demonstration is implemented. Most new demonstration programs do not offer such a readily located target population, and most do not offer preexisting measures of relevant outcomes for the full target population.

Time Series Analyses (Quasi-Experiment 4). A time series design for evaluating a new program demonstration differs only in scale from the design for evaluating an ongoing national program. The approach uses aggregate data from multiple time periods before and after implementation of the demonstration. The difference is that the aggregation unit cannot be the whole country, but must be a unit that closely tracks the demonstration's target population.

One interesting example of using parallel time series in multiple sites is a study now in progress at the Manpower Demonstration Research Corporation. This demonstration evaluation concerns JOBS+, a program sponsored by the U.S. Department of Housing and Urban Development (Bloom, 1996). The intervention consists of intensive job training and employment search assistance in 10 housing projects across the country.[21] The outcomes of interest are employment and earned income. Measures are to be obtained by constructing a time series of employment and earned income from existing employment security quarterly earnings records[22] of residents in the public housing units for each project affected by the program. The availability of the 10 time series will provide insight into the consistency of JOBS+ effects across locations.

Comparing Demonstration and Comparison Target Populations (Quasi-Experiment 6). In this design, program outcomes are measured for the demonstration target population and for a comparably defined population elsewhere. Impacts are estimated in a model that includes presence of the demonstration, location, and individual characteristics.

This design is highly vulnerable to the possibility that outcome differences are related to the subjects' location rather than to the effect of the demonstration program. Impact estimates are therefore not very reliable.

"Theories of Change" Evaluations. In recent years some evaluators have advocated an approach to evaluation most often referred to as "theories of change" (Weiss, 1995; Chen, 1990). Proponents of this approach do not claim that it can yield quantitative estimates of program impact. Rather, it assembles information that, in the absence of solid impact estimates, provides some perspective on the possibility

[20] Unfortunately, Myers' study had serious technical failings, including a high rate of missing observations, that undermine its credibility.

[21] The evaluation is also a small randomized experiment with organized units (housing projects) as targets. Ten public housing authorities each identified three housing projects, one of which was randomly selected for the intervention and the others to serve as controls. Ten demonstration units and 20 control units constitute the evaluation sample.

[22] Each State employment security agency maintains files of quarterly earnings from covered employment by individual earners. Because the names and social security identifiers of public housing residents can be obtained from administrative records, it is planned to obtain quarterly records for residents for several years before and after the public housing program is in place.

that a program could be having an impact. (It is also offered as a useful tool for program development and for developing hypotheses that may be tested in more formal evaluations.)

The approach is considered especially applicable to demonstration programs that are not only new, but are developing even as they are being implemented. The approach recommends that close attention be paid, during this developmental stage, to explicitly describing whatever theory underlies a program. This entails detailed specification of the steps or "pathways" through which program activities lead to outputs, intermediate outcomes, and ultimate outcomes. Data are then collected on the volume of activities, outputs, and intermediate outcomes.

The theories-of-change approach does not include an explicit representation of the Counterfactual, and hence cannot refute the hypothesis that observed outcomes would have occurred without the program. The underlying proposition is that if the program generates the planned volume of activities and outputs, and if intermediate and ultimate outcomes occur as theorized, one cannot reject the hypothesis that the program has some impact. Alternatively, finding minimal levels of program outputs and intermediate outcomes would make it quite difficult to believe that important impacts are occurring.

The primary application of this approach has been to interventions that aim for institutional or community-level effects such as enhanced community development. Some applications have involved initiatives in which objectively measurable outcomes are not clearly identified and program operations are not fixed, but evolve in response to local conditions. The approach itself is fluid and typically involves the participation of major program stakeholders in eliciting underlying theories.

Within the context of food assistance and nutrition programs, which aim to enhance the nutrition and health status of reasonably well-defined populations, this approach may be useful in designing and developing programs that then need to be tested for effectiveness. It would not be recommended for impact evaluation of food assistance and nutrition programs that are beyond the design phase.

Research Activities That Complement Demonstration Impact Evaluations

Monitoring and participation studies were described previously as providing important information for assessing ongoing programs. Such studies can also play an important part in evaluating demonstrations. Participation studies are particularly important. If a demonstration intervention proves to be poorly targeted, or unable to reach its intended target population, corrective changes may be needed before the intervention is implemented on a large scale.

Because demonstrations involve interventions that have never been tried before, how well the intervention can be implemented in the field is an important question that should be answered before full-scale implementation. For this reason, it is usually recommended that program process studies be conducted to complement impact evaluation (Werner, 2001 (forthcoming)). Program process studies employ a variety of research methods, including ethnography, focus groups or indepth interviews held with demonstration participants and agency staff, the analysis of program administrative data, and surveys of participants and staff.

Process studies typically seek to describe the program from several perspectives. Operating statistics are used to describe flows of participants into and through the program and to identify bottlenecks or unintended attrition. Interviews with program staff and observation of program activities yield detail on the services provided and the procedures through which participants are handled at each stage of their involvement with the program. The participant perspective includes descriptions of how individuals gain information and access to the program, barriers to participation, knowledge and attitudes about the program, possible stigma or burdens attached to participation, and satisfaction with services and benefits offered by the program.

The policymaker ultimately wants to know whether the program or intervention is worth its cost, which is the question addressed in the cost-benefit or cost-effectiveness study. Such economic efficiency studies juxtapose the results of the impact evaluation with information on the costs and burdens the program imposes on taxpayers, program participants, and sometimes other stakeholders.

Economic efficiency studies typically have two quite distinct components. The first is a program cost study, which typically involves both primary data collection and the assembly of data from program accounting records. The most important costs are typically the direct cost of the service or benefit and the various administrative costs of delivering the service. Service delivery costs usually occur at State and local levels of program operation as well as at the Federal level. Even in programs in which the Federal Government makes a payment for administration, State or local operators often incur costs beyond those reimbursed. In addition to these costs, many evaluations must consider costs to participants, most commonly time expended in complying with the requirements for program participation, but sometimes also tax payments or work expenses associated with income received. Some participant costs can be difficult to express in monetary terms, such as the potential for job loss associated with taking off time from work or negative psychological consequences of receiving assistance. When other stakeholders are involved in service delivery, as food retailers are in redeeming food stamps and WIC vouchers, costs to these groups may have to be measured as well.

The second major component of economic efficiency studies consists of transforming the impact estimates and program costs into comparable time periods and perspectives. Often the program costs for a particular participant are incurred quickly, during a brief period of program participation, while impacts develop slowly and endure for some years. Efficiency studies are therefore typically framed in terms of the “participation lifetime” (i.e., all of the costs and impacts that are incurred between the time the participant comes in contact with the program and the time when impacts cease to be counted). The studies usually recognize explicitly that one party’s cost may be another party’s benefit. Thus, cost and effect data are typically presented from at least three perspectives: that of the taxpayer, that of the participant, and that of society as a whole (usually conceived as the net of all parties’ perspectives).

When costs and benefits are naturally measured and expressed in dollar terms, it is easy and meaningful to calculate a benefit/cost ratio or net benefit per participant. When translating effects into monetary units requires heroic or tenuous assumptions, however, it is seldom useful to make the translation. This is most often the case with food assistance and nutrition programs, whose nutrition and health impacts are not usually measured in dollar terms.[23] Even when some effects or costs cannot reasonably be monetized, however, the efficiency study is a critical requirement for policymaking. Only when program costs and effects are presented together can the policymaker understand what the program returns for a dollar spent.

Other Program Evaluation Situations

Most evaluations of USDA’s food assistance and nutrition programs will probably be overall evaluations of the ongoing programs or demonstration interventions. Two other evaluation situations, which arise less frequently, are discussed in this section. In one situation, the evaluation concerns a change to an ongoing program that is implemented at the same time in all program locations rather than being introduced as a pilot or demonstration initiative. This situation is distinguished by a very limited set of options for evaluation design. In the second situation, the evaluation focuses on a single component of an ongoing program, attempting to distinguish its impact within the overall program package.

Impact Evaluation of Programwide Modifications to Ongoing Programs

Major national programs sometimes undergo important general changes, such as in eligibility criteria or the nature of program benefits or services. Such changes often result from legislation requiring nationwide implementation of the change on a particular date. Unlike the demonstration trial of a program modification, this situation offers no opportunity to observe the old rules and new rules operating in parallel for different individuals or areas.

A current example is the PRWORA, which radically changed the way participating family child care homes are to be qualified for eligibility for cash subsidies in the CACFP. Prior to PRWORA, a fixed per meal subsidy was paid to all participating family child care homes for all children who were served meals in the

[23] There are exceptions, such as the study in which Devaney and colleagues (1991) calculated that the savings in Medicaid expenditures achieved by raising the average birthweight of newborns more than offset the costs of running the Medicaid program. (Devaney's is not a full cost-benefit study, however, because only some costs and some benefits were considered.)