United States Department of Agriculture

Economic Research Service

Food Assistance and Nutrition Research Report Number 19-1

USDA

Effects of Food Assistance and Nutrition Programs on Nutrition and Health

Volume 1, Research Design

William L. Hamilton
Peter H. Rossi

Food Assistance & Nutrition Research Program

Effects of Food Assistance and Nutrition Programs on Nutrition and Health: Volume I, Research Design. By William L. Hamilton of Abt Associates Inc., and Peter H. Rossi of the University of Massachusetts at Amherst, for the Food and Rural Economics Division, Economic Research Service, U.S. Department of Agriculture. Food Assistance and Nutrition Research Report No. 19-1.

Abstract

This is the first of four reports in the "Nutrition and Health Outcome Study," which assesses the effects of USDA's food assistance and nutrition programs on nutrition and health outcomes. This report reviews the research designs available to evaluators for assessing the effect of USDA's food assistance and nutrition programs. The random assignment experiment is the "gold standard" design for such an evaluation. Where random assignment is impossible, quasi-experimental designs are used to infer what would have happened to program participants if the program had not existed. Eight types of quasi-experimental design are identified as having been used in evaluations of food assistance and nutrition programs, although none can guarantee unbiased estimates of program impacts.

Keywords: USDA's food assistance and nutrition programs, nutrition and health outcomes, research design.

Acknowledgments

The authors wish to recognize the contributions of Nancy Burstein, David Hoaglin, Stephen Kennedy, and Larry Orr of Abt Associates in preparing various parts of this report. The report also benefited greatly from insightful review, comments, and suggestions from many reviewers, including members of the project's expert panel: Johanna Dwyer, Rick Trowbridge, Robert Pruzek, Janet Currie, Barry Popkin, Shiriki Kumanyika, and Suzanne Murphy; staff at the Economic Research Service (ERS), most notably Mark Prell and the Food and Nutrition Service; the ERS Project Officer (Biing-Hwan Lin); and the Abt Associates Project Director (Mary Kay Fox).

This is the first of four reports completed by Abt Associates Inc., under the "Nutrition and Health Outcome Study," which is a research contract from the Economic Research Service. The views expressed are those of Abt Associates and not necessarily those of ERS or USDA.

Washington, DC 20036 February 2002

Contents

Summary

This is the first of four reports in the "Nutrition and Health Outcome Study," which assesses the effect of USDA's food assistance and nutrition programs on nutrition and health outcomes. This report reviews the research designs available to evaluators for assessing the effect of USDA's food assistance and nutrition programs. The random assignment experiment is the "gold standard" design for such an evaluation. Where random assignment is impossible, quasi-experimental designs are used to infer what would have happened to program participants if the program had not existed. Eight types of quasi-experimental design are identified as having been used in evaluations of food assistance and nutrition programs, although none can guarantee unbiased estimates of program impacts.

Since the mid-1940s, the U.S. Government has committed to ensuring that its citizens neither go hungry nor suffer the consequences of inadequate dietary intake. Today, the U.S. Department of Agriculture (USDA) implements 15 programs as a "food safety net," to provide low-income citizens with food or the means to purchase food. These food assistance and nutrition programs (FANPs) were funded at a level of $33.5 billion in fiscal year 1998.

Under contract with the Economic Research Service of USDA, Abt Associates Inc. has completed a review of knowledge about FANP effects on nutrition- and health-related outcomes. A thorough literature review was conducted to evaluate the strengths and weaknesses of the research designs, analytical methods, and data sources employed to analyze FANP outcomes. A series of four reports has been produced to document what we know and do not know about these outcomes and to identify future research needs.

This report reviews the research designs and analytic approaches that have been used to assess FANP outcomes. The discussion focuses on the five main food assistance and nutrition programs: the Food Stamp Program (FSP); the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC); the National School Lunch Program (NSLP); the School Breakfast Program (SBP); and the Child and Adult Care Food Program (CACFP). The research designs found in this evaluation literature are applicable not only to other Federal food assistance programs but to practically all social programs that directly serve individuals and families.

In the same vein, although the discussion refers most often to nutrition- and health-related outcomes for participants, the research designs are generally applicable to any outcomes measured for individuals. However, the data available (or not available) on nutrition and health outcomes in existing data sets, along with the procedures required to collect these data, sometimes constrain design choices for evaluating food assistance and nutrition programs.

The random assignment experiment is the "gold standard" research design for evaluating food and nutrition assistance programs, as for many other social programs. It is particularly well suited to evaluating demonstration programs or proposed modifications of existing programs, such as raising the age limit for children in WIC or adding a nutrition education component to the Food Stamp Program.

Numerous quasi-experimental designs have been applied in evaluating food and nutrition assistance programs. The eight quasi-experimental designs discussed here offer varying ways to estimate program impact, where impact is defined as the difference between outcomes for program participants (or for a target population that includes participants) and the outcomes that would have been expected in the absence of the program. Quasi-experimental designs represent the outcomes expected in the absence of the program, called the "Counterfactual," by outcomes in nonprogram time periods (pre-post and time-series designs) and/or by contemporaneous outcomes for nonprogram populations (comparison group designs).

All quasi-experimental designs are potentially vulnerable to selection bias, a situation in which an observed difference between participant and Counterfactual outcomes is caused by some force other than the intervention being evaluated. Researchers have used various statistical approaches that attempt to correct for selection bias, but none of these techniques provides certainty that selection bias has been eliminated. Because the properly implemented random assignment experiment is not vulnerable to selection bias, it is the best available approach to estimating program impacts.

Effects of Food Assistance and Nutrition Programs on Nutrition and Health

Volume I, Research Design

William L. Hamilton and Peter H. Rossi

Introduction

This report reviews the research designs available to evaluators as they contemplate assessing the impact of USDA's food assistance and nutrition programs on nutrition and health outcomes. The focus of the discussion and the examples presented concern the five main food assistance and nutrition programs: the Food Stamp Program (FSP); the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC); the National School Lunch Program (NSLP);, the School Breakfast Program (SBP); and the Child and Adult Care Food Program (CACFP). The research designs discussed cover the gamut of those commonly mentioned in the evaluation literature, and are applicable not only to other Federal food assistance and nutrition programs but to practically all social programs that directly serve individuals and families.

In the same vein, although the discussion refers most often to impact on participants' nutrition and health outcomes, the designs are generally applicable with any outcomes measured for individuals. However, the availability or nonavailability of nutrition and health outcome data in existing data sets and the procedures required to collect these data sometimes constrain the set of design choices for evaluating food assistance and nutrition programs.

The impact of a program or other intervention is defined as the difference between what happens in the presence of the intervention and what would have happened in its absence. What would have happened in the program's absence is generally called the "Counterfactual."

All of the evaluation designs discussed in this report involve measuring outcomes that occur in the presence of the intervention and comparing them to some representation of outcomes in the Counterfactual.[1] The main difference in the designs lies in the ways they represent the Counterfactual.

Establishing the Counterfactual—estimating what would have happened without the program—is usually accomplished by examining a population that has not been subjected to the intervention being evaluated. What makes the task difficult is the fact that people[2] who become participants in a social program are often quite different from those who do not, because they either have been selected for participation or have selected themselves (Campbell and Stanley, 1963). These selective processes may make participants different in important ways from those who do not participate. These differences include not only people's permanent characteristics, such as their gender or race, but also transitory characteristics such as their current income or employment, the opportunities they face, and the experiences they have had. Many of the transitory characteristics result from the time and place in which people live, which means that similar people in a different time or place may not appropriately represent the Counterfactual. All these influences may contribute to selection bias, which distorts the evaluation of a program's impact.

[1]This excludes prospective simulation approaches to evaluation. In these approaches, outcomes in the presence of the intervention are not measured directly, but projected on the basis of prior information or assumptions about the effects of the intervention or its elements.

[2]Evaluation designs often focus on units other than people, either aggregations of people (e.g., families, students in a school, the population of a county) or operating entities (program offices, schools, businesses). For simplicity of presentation, the discussion generally refers to individuals rather than aggregations or other entities.

The general strength of an evaluation design in a particular situation can be assessed through the following three questions:

- Is the population representing the Counterfactual equivalent in all pertinent respects to the program population before that population is exposed to the intervention?
- Is the intervention the only force that could cause systematic differences between the two populations once exposure begins?
- Is the full force of the intervention applied to the program population, and is none applied to the Counterfactual population?

The technically preferable evaluation design in any situation is one that provides strong affirmatives to all three questions. In the sections that follow, these three questions will be used to characterize the conceptual strengths and weaknesses of each design.

The report has three main sections. The first section deals with designs for evaluating ongoing national programs, such as the five major food assistance and nutrition programs. Because these programs are available to practically all potentially eligible people nationwide, and because they have been operating for a long time, they pose particularly difficult challenges for evaluation.

The second section focuses on designs applicable to evaluations of demonstration initiatives that would modify existing programs or create new ones. Many food assistance and nutrition program evaluations are likely to fall into this category, which fortunately tends to be more tractable. The third and final section of the report considers two less common evaluation situations: evaluation of a mandated programwide reform and natural- and planned-variation evaluations of program components.

Impact Evaluation of Ongoing Programs

The question of whether and how much the major food assistance and nutrition programs affect the nutrition and health outcomes of participants has obvious policy importance. These programs account for very sizable Federal expenditures—$33.5 billion in fiscal year 1998—but little scientifically sound evidence exists on the programs' impacts, particularly their effect on nutrition and health outcomes.

The ongoing food assistance and nutrition programs have two characteristics that make it extremely difficult to assess their overall impact on participants' nutrition and health outcomes. First, they are essentially universally available throughout the United States. For practical purposes, there exists no current population that has not been exposed to the programs, where people are considered "exposed" if they have reasonable access to information about the program and would be able to participate if they applied and were found eligible. Second, the programs have operated nationally at a substantial scale for a minimum of two decades. This means that, even if one could find measures of the relevant outcomes for a period before the programs began, no identifiable population in the preprogram period is likely to have permanent and transitory characteristics equivalent to those of today's participants.

Of the several possible research designs described in this report, only randomized experimentation is actually capable of providing reliable estimates of the programs' impacts. However, randomized experiments have not been applied to measure the overall impact of these programs to date (although they have been used to measure the impact of program modifications), and we recognize the likelihood that such experiments may not happen in the near future. For this reason, we discuss several possible quasi-experimental designs. The quasi-experimental designs, which are second-best

choices in any circumstances, are made especially weak by the long-term universal availability of the food assistance and nutrition programs. Nonetheless, their operational feasibility makes them more likely to be applied than randomized experiments. If a quasi-experimental design is applied, even with best efforts to adjust for possible biases, it is important to remember that the estimate of program impact has a substantial probability of being far from the true value.

The Randomized Experiment: The Gold Standard

The randomized experiment is the "gold standard" of program evaluation. The scientific community is not completely unanimous on this point, but the consensus is strong enough that, for example, pharmaceutical companies must conduct randomized trials of new drugs in order for the products to be approved for marketing in the United States. Better than any other design, the randomized experiment answers affirmatively the three central questions posed earlier (see box).

The Randomized Experiment

Features:

Impact estimate: Difference in post-program outcomes between one group randomly assigned to intervention and one group randomly assigned to control status.

Key requirement: Ability to randomly assign subjects before exposure.

Advantage: Most credible estimates.

Disadvantage: Legal or ethical prohibition of withholding program services from controls.

The Three Questions:

Alike before exposure? Yes, within the range of chance variation.

Difference solely from intervention? Yes.

Full force of intervention represented? Usually yes, if designed and implemented carefully. Contamination and attrition can be issues.

In the simplest form[3] of a randomized design, program targets are randomly assigned either to an "experimental" (or "treatment") group that will be subject to the program being assessed, or to a "control" group from which the program will be withheld. The program's impact is then estimated by comparing the average outcomes in the experimental group, after sufficient exposure to the program, with control group outcomes measured at the same time.

Because the experimental and control groups differ at the outset only by chance, they are considered fully "alike" at that point—equivalent, in the statistical aggregate, on all permanent and transitory characteristics. Subsequently, the only systematic difference between the groups is exposure to the program. Accordingly, it is credible to infer that any post-program differences between the two groups are caused by the program, provided that the differences are greater than what might occur by chance.

When feasible, it is advantageous to enrich these inferences by designing the experiment so that the randomization takes place separately within each of two or more relevant subgroups of subjects (which might, for example, be defined by income, nutritional status, or age). This strategy, known as "blocking" or "stratification," ensures that each of the subgroups is adequately represented in the experimental group and the control group.

The analyst can then examine how the effect of the program differs across subgroups. When the program produces similar effects in the subgroups, it is straightforward to reaggregate the subgroups and thus simplify the analysis. If the characteristics that define the subgroups are known only after the data have been collected during the experiment, it may still be possible to gain information by forming analytic strata (though the numbers of subjects in each stratum will be a chance outcome of the randomization). Differences in effects are much more difficult to deal with, however, if the subgroups were not set up initially.

[3]Complex experiments that involve comparing alternative programs or varying components of a program are common. These complex designs are discussed in a later section, which deals with impact evaluation of demonstrations or program changes, because that is the context in which these designs are most frequently used.

Another important point is that, although the experimental and control groups differ at the outset "only by chance," randomization gives only the expectation of sound inferences. That is, not all randomizations yield groups that are closely similar. Relatively large sample sizes will generally minimize the chances of erroneous inferences. In addition, replication studies are highly desirable to strengthen the base for policy decisions.

With regard to the third question—whether the group differences in an experiment reflect the full force of the intervention—the randomized experiment does not have an inherent advantage over other designs. In fact, special effort may be required in the research design, in implementing the experiment, or in implementing the intervention itself, to make sure that the experimental group experiences the intervention fully and that the control group experiences none of it. This requires attention not only to the subjects' exposure to the treatment, but also to the potential for a "placebo" effect, in which surveys or other research activities bring about behavioral changes that can be confounded with the treatment effect.[4] Where other factors permit randomization, however, an experiment can usually be designed and implemented to meet these criteria.

The Obstacle to Randomized Experimentation in Assessing Ongoing Programs

The fundamental requirement of randomized experimentation is that the program service be deliberately withheld from some people who are otherwise like the people who receive the service. This generally cannot be done in entitlement programs and is difficult in saturation programs.

In entitlement programs—including the FSP, the NSLP, the SBP, and the CACFP—law and regulation require that program benefits or services be provided to everyone who meets program eligibility requirements and takes the necessary steps to qualify. Benefits cannot legally be withheld.

Saturation programs, such as WIC, pose quite similar problems even though they are not entitlement programs. Whether a potentially eligible person can receive program benefits from a nonentitlement program depends on the local availability of program funding and infrastructure. A saturation program is one with sufficient funding and infrastructure to serve essentially all eligible persons. For many nonentitlement programs that approach full saturation, then, it can be virtually impossible to find a reasonably representative set of targets to whom the program could be considered unavailable. If program services would normally be provided to everyone who applies and is eligible, it may be considered unethical to withhold services from people who might apply.

Potential for Randomized Experimentation

The financial and human stakes involved in the major food assistance and nutrition programs make it extremely important to use the most reliable methods to evaluate their effectiveness. Given the general unreliability of nonexperimental methods, especially for entitlement and saturation programs, this means using random assignment wherever it is legally and ethically possible.

As noted in the previous section, current law probably prohibits denial of service to eligible applicants in any of the five major programs except WIC, thereby ruling out random assignment to a no-service control group for these programs (FSP, NSLP, SBP, and CACFP).[5] An argument could be made, however, for asking Congress to exempt program evaluations from this prohibition, in order to obtain reliable measures of the programs' effectiveness. Both taxpayers and program participants have a strong interest in knowing whether these programs are working as intended. An ineffective program can waste billions of tax dollars year after year. Moreover, an ineffective program imposes costs on its intended beneficiaries as well, by consuming government and personal resources that might be used more effectively to address their problems. Faced with a choice, Congress might well decide that these risks outweigh the costs that a random assignment evaluation would impose on a small number of program eligibles.[6]

The same legal barriers do not apply to nonentitlement programs, and, in fact, several ongoing national pro-

[4]Because the placebo effect is not typically a concern in evaluating food assistance and nutrition programs, we do not treat here the ways in which the research design can be modified to deal with the problem. Most responses involve adding an additional group to the design. Thus, in addition to a group representing the treatment condition and one representing the Counterfactual, a further group represents the Counterfactual in the absence of those activities expected to cause the placebo effect.

[5]A legal opinion would be needed to determine whether a particular random assignment evaluation strategy for a particular program would be legally permissible.

[6]For a discussion of the broader ethical issues involved in the evaluation of ongoing programs, see Orr (1999), pp. 19-22.

grams have been evaluated with random assignment. The U.S. Department of Labor, for example, has launched random-assignment evaluations of each of its major ongoing employment and training programs—the Job Training Partnership Act (JTPA), the Job Corps, and the Economic Dislocation and Worker Adjustment Assistance (EDWAA) Program.[7]

If a nonentitlement program has many more applicants than can be accommodated, randomization can be justified as an even-handed method of selecting program participants. Those selected randomly for participation can be regarded as an experimental group, and those who are not selected become members of a control group. For example, in the first years of WIC, when appropriations were adequate to cover only a fraction of persons applying for benefits, it would have been possible to design and carry out randomized experiments.

A somewhat more complicated version of this opportunity may exist even when a program does not have substantial excess applications but is not reaching all of its intended population. For example, a feasibility test examined the possibility of evaluating WIC's effect on children through a randomized experiment (Puma et al., 1991). The design took advantage of the fact that program funding was limited and that children 1 to 5 years of age were considered at low priority for receiving WIC benefits. Few children could be served, and outreach and referral networks for children were very limited in some areas.

In conducting the feasibility test, referral outposts were established in underserved areas to identify potentially eligible children whose mothers were unaware of WIC or unaware that they might qualify for WIC benefits. These mothers were randomly assigned to experimental or control status. Those in the experimental group were referred to WIC, and funding was made available to ensure that they would be enrolled in the program. Control group members were not referred to WIC, but benefits were not withheld from any who learned of WIC through normal channels and applied (if eligible, they would be served or placed on a waiting list, depending on funding availability at the time and the clinic's normal procedures). Although this design was difficult to implement and did not result in perfect separation of experimental and control groups, it illustrates the point that randomization can sometimes be accomplished where it initially seems infeasible.

Another possibility might be to offer program benefits and services to a population that would not otherwise be eligible for program benefits. For example, WIC benefits might be offered to a random sample of families with incomes between 185 percent and 250 percent of the poverty line, or to 5-year-old children (currently, the program serves children up to the age of 5). Theoretically, these groups should have less need for WIC, and the program should therefore have less impact. If positive impacts were found in a randomized experiment, it would be quite reasonable to infer that impacts also exist for the actual program population. Conversely, if no impacts were found, it might then be deemed acceptable to conduct a randomized experiment within the eligible population, perhaps limiting it to those closest to the eligibility cutoff.

The startup phase of a new entitlement or saturation program may provide other opportunities for the employment of randomized designs. Sometimes such programs are put in place in a staggered sequence, starting up earlier in some jurisdictions than in others or starting with some categories of eligible targets first and later adding others. We postpone discussion of these opportunities until the final section of the report, dealing first with quasi-experimental approaches to evaluating the ongoing programs.

Quasi-Experiments

For most ongoing programs, it is necessary to identify Counterfactual conditions without random selection into control and experimental conditions. The class of such impact evaluation designs is known as quasi-experiments. That is, they resemble experiments in providing a specific representation of the Counterfactual, but the Counterfactual is identified through some means other than random selection. In the sections that follow, we review four quasi-experimental designs that may be used when a randomized experiment is not feasible. A key theme running through the discussion is selection bias—the ways in which it arises in a particular design and the ways in which it can be reduced.

[7]See Orr et al. (1996) for a description of the National JTPA Study and its results and Burghardt et al. (1997) for a description of the Job Corps Evaluation.

Quasi-Experiment 1
Participants vs. Nonparticipants

Features:

Impact estimate: Difference between post-program outcomes for program participants and outcomes for nonparticipants.

Key requirement: Strategy for minimizing selection bias.

Advantage: Operational feasibility.

Disadvantage: No certainty that estimate is unbiased.

The Three Questions:

Alike before exposure? No. Different kinds of people choose (or are chosen) to participate.

Difference solely from intervention? Probably no. Different characteristics can affect outcomes during program period.

Full force of intervention represented? Usually yes. Participants are exposed, nonparticipants aren't.

Quasi-Experiment 1: Comparing Participants to Nonparticipants

This design involves identifying comparable groups of participants and nonparticipants and interpreting the average post-program outcome differences between the groups as effects of the program (see box).

Several factors make this an operationally feasible approach to evaluating ongoing entitlement or saturation programs (but one with serious technical risks, as discussed subsequently).

Positive feasibility factors include:

- **Availability of subjects.** As long as any substantial portion of potentially eligible targets does not participate, which is the case with the USDA food assistance and nutrition programs, sufficient nonparticipants are likely to be available for research. Unlike random assignment, no special administrative operations are required to build a sample.

- **Broad-scale analysis.** Routinely collected national surveys—such as the National Health and Nutrition Examination Survey (NHANES), the Continuing Survey of Food Intakes of Individuals (CSFII), the Current Population Survey (CPS), and the Survey of Income and Program Participation (SIPP)—have potentially useful outcome measures for participants and nonparticipants alike, as well as measures of participation status. This makes it possible to consider the whole national program (whereas random assignment can normally be conducted in only a small number of locations).

- **Applicable after intervention.** This design is often chosen when timing or funding limitations preclude collecting data on the key outcome dimensions before people are exposed to the program (i.e., before the participants become participants).

An important constraint on the operational feasibility of this design is that the nonparticipants must be potentially eligible—i.e., people who apparently could have applied and qualified for the program, but did not—to be a credible representation of the Counterfactual. For the food assistance and nutrition programs, the researcher normally attempts to apply an approximation of the means test, choosing nonparticipants with incomes below the eligibility cutoff for the program in question.

The practical consequence of this requirement is that most researchers applying this design use data from broad population surveys that were conducted for other purposes. A special-purpose survey can generate a representative sample of eligible nonparticipants, but it is very costly because eligible persons usually make up a tiny fraction of the general population. Hardly any administrative data sets include both participants and nonparticipants, identify which is which, and provide the information needed to judge potential eligibility (although we describe below one study that did use administrative data in this way). Thus, the participant vs. nonparticipant design is most feasible with large national surveys, especially surveys that oversample the low-income population, and large national programs like the food assistance and nutrition programs.

Selection Bias in Participant/Nonparticipant Comparisons. The major problem with this quasi-experimental design is that identified nonparticipants may not be sufficiently comparable to participants. This problem, known as selection bias, is a difficult issue in all quasi-experimental designs and is especially troublesome when comparing people who have taken the actions necessary to participate in a program with people who have not.

Selection bias often occurs because participants are more highly motivated to achieve the program-relevant outcomes than are nonparticipants. Suppose, for example, that the women who seek WIC benefits for themselves or their children tend to be very concerned about the effect of diet on their children's health. Such women may well take other actions with the same objective, such as following dietary guidelines in brochures they pick up in the doctor's office—or getting to a doctor's office at all. If this were true, one would expect the children of mothers who seek WIC benefits to have better nutrition and health outcomes—even in the absence of the program—than children of mothers who are less motivated and do not seek WIC benefits. A simple comparison of WIC and non-WIC children would therefore reveal that the WIC children had more positive outcomes even if the program had no effect at all.

Sometimes selection bias operates in the opposite direction. Mothers of children experiencing nutrition-related problems might be especially motivated to seek WIC benefits, for example, whereas mothers of healthy children might be less inclined to participate. WIC might improve the participating children's condition, but the participating children might not catch up to their nonparticipating, healthier counterparts. In this example, the simple comparison would find WIC children to have less positive outcomes even though the program had a positive effect.

Motivation toward the program outcome is one of the most common sources of potential bias, and one of the most difficult to counteract. Other common sources of self-selection bias include need (often proxied by income), potential for gain (often proxied by the dollar value of the benefit), and the individual's desire not to depend on public assistance.

Selection bias may also result from program rules or procedures. In nonentitlement programs, local staff often decide which applicants will be approved for participation based on a combination of program policies and individual judgment. In all programs, outreach practices, referral networks, office locations and hours, and community customs may make some people more likely to participate than others.

Finally, some selection bias occurs when program participation is based on transitory characteristics. For example, some people who qualify for means-tested programs are permanently poor, or nearly so, with incomes below the programs' limits in most or all time periods for many years. Other people who qualify for those programs are not permanently poor, but are at a temporary low point in a fluctuating income pattern. In an earlier period, their income was sufficiently high that they did not qualify for the program, and their income will at some point regain its previous level. These two types of people might have similar income at the time they enter the program, but their subsequent outcomes, in the absence of the program, might not be at all similar.

Approaches To Dealing With Selection Bias. Researchers have used a variety of approaches to attempt to counteract selection bias, the most common of which are described below. All have the basic objective of making the participant and nonparticipant groups "alike" on certain specified dimensions. However, all leave open the possibility that bias remains.

Regression adjustment. A prime example of this approach is the impact evaluation of the WIC program for pregnant women, conducted by Devaney (1992). Taking advantage of the fact that all Medicaid recipients were automatically eligible for WIC benefits, Devaney contrasted birth outcomes of recipients who had participated in WIC during pregnancy with those who had not participated in WIC. The relevant data set was assembled by linking Medicaid records to WIC participation records and birth registration records. Birth registration records provided information on the critical outcome of birthweight, WIC records identified WIC participants, and Medicaid records identified those who gave birth during the period of study. Devaney's research included 112,000 births to Medicaid mothers during a 2-year period in 5 States.

To minimize selection bias, Devaney used regression adjustments. Her equations included variables that were likely to capture ways in which participants and nonparticipants might differ, including educational attainment, prenatal medical care, gestational age, race, mother's age, and birth parity. As happens typically, Devaney was limited to the variables captured in existing data sets, which seldom measure all the factors that might create different outcomes for participants and nonparticipants. Alternative attempts by Devaney and her colleagues to counter selection biases led to quite drastic changes in estimates of the effects, without any clear indications of which attempt was more sensible.

Matched pairs. Sometimes researchers construct a comparison group by matching participants and nonparticipants on characteristics that are thought to be related to selection tendencies. For each participant in the research sample, the researcher identifies a nonparticipant with identical or closely similar key characteristics on variables. Because the matching procedure can normally consider only a few variables, regression adjustment is still needed to estimate impacts.

The matched-pair approach is advantageous mainly when there is a substantial marginal cost for including subjects in the evaluation, typically when significant new data collection is to be carried out. If the analysis is based on existing administrative or survey data sets, the matched-pairs approach excludes otherwise usable observations and thus reduces the sample size available for analysis.

More general matching procedures may identify more than one nonparticipant (perhaps even many) who is similar enough to each participant. When combined with regression adjustment, matched sampling is one of the most effective methods for reducing bias from imbalances in observed covariates (Rubin, 1979).

Dose-response. If program rules prescribe different amounts of the program benefit or service for different participants, a dose-response analytic model may be applicable. The underlying hypothesis is that greater benefits will lead to greater effects on outcomes. The dose-response relationship may be estimated with a sample that consists only of participants, which eliminates the issue of whether participants differ from nonparticipants in unmeasurable ways. If this relationship can be estimated, then the program's impact may be described as the difference between the effect at any given level of benefits (typically the average benefit) and the projected effect at the zero benefit level (what participants would receive if they did not participate).

The Food Stamp Program, with benefits measured in dollars and a very large number of actual benefit amounts, is the main candidate for dose-response analysis among the food assistance and nutrition programs. A number of researchers have used this approach, although with considerable variation in the way the approach is applied. In particular, some researchers have estimated models that exclude nonparticipants (Neenan and Davis, 1978; Levedahl, 1991; Kramer-LeBlanc et al., 1997), while others include nonparticipants and specify the model to include both a term representing the benefit amount and a term representing participation per se (Fraker, 1990; Devaney and Fraker, 1989).

The dose-response model requires that benefits must vary across households that are similar in terms of the factors expected to affect their health and nutrition outcomes. The food stamp situation does appear to meet that condition. Households of a given size with a given amount of cash income receive differing benefit amounts depending on, for example, how much of the income is earned and their allowable deductions. Because the underlying logic driving benefit rules is that the benefit amount should be responsive to need, it would be desirable to see more extensive analysis of the extent to which food stamp benefit variation actually meets the requirements of dose-response analysis. Nonetheless, with careful application, this appears to be a promising approach.

Two-stage models. Some researchers use a two-stage approach in which they first model the likelihood that an individual will be a participant in the program. The model yields a predicted probability of participation for each participant and nonparticipant. The second stage of analysis models the outcome as a function of some measure of participation.

One class of solutions simply uses the predicted probability of participation in place of actual observed participation as an explanatory variable in the second-stage model. Another includes observed participation along with an inverse Mills ratio, which is a function of the predicted probability of participation (Heckman, 1979).

In order for these approaches to offer a material gain over simple regression adjustment, the participation model must include one or more "instruments"—variables that predict participation but are not correlated with the outcomes of interest. Finding an appropriate instrument is often impossible, however, especially when the researcher is working with existing data sets. Participation is typically related to demographic characteristics, need or potential benefit, motivation, and pre-program measures of relevant outcomes such as nutrition or health status. These same factors usually influence post-program outcomes. And many factors that initially seem like good instruments turn out on closer examination to be related to outcomes. For example, living close to a program office might be expected to make an individual more likely to participate and initially seems unrelated to health and nutrition outcomes, but the program's location may have been selected to give easy access to a high-risk community.

In addition to the instrumental variable, some two-stage approaches use functional form to achieve identification in the models. In a procedure known as the two-step Heckman method, the participation model uses a nonlinear functional form (Heckman, 1979; Heckman and Hotz, 1989). Alternatively, the participation and outcome equations can be estimated simultaneously using a maximum likelihood approach. In both cases, the effectiveness of the method depends on the validity of assumptions made about the error terms in the model, assumptions that cannot be verified empirically.

All of these approaches have been used in evaluating food assistance and nutrition programs, but with no clear consensus that any of them can be considered generally reliable. For example, Gordon and Nelson (1995) used three approaches and a rich data set to estimate WIC effects on birthweight (instrumental variables, Heckman two-step, and simultaneous equations). They found that the approaches to selection bias correction yielded "unstable and implausible results, [possibly] because the factors affecting WIC participation and birthweight are very nearly identical, since WIC targets low-income women at risk for poor pregnancy outcomes." Ponza et al. (1996) similarly used multiple approaches to selection bias adjustment in evaluating the Nutrition Program for the Elderly. The authors rejected all of the two-stage approaches and based their conclusions on the results of the simple, one-stage regression adjustment.

Use of propensity scores. In principle, regression adjustments can be used to take account of any observed differences in the characteristics of the treatment and comparison groups. In practice, regression adjustments must often be limited to a relatively small number of covariates and, in the case of continuous covariates, to simple adjustments for differences in averages. Propensity scoring allows a more comprehensive and complex treatment of covariates that is particularly useful when the number of potential covariates is quite large (Rosenbaum and Rubin, 1983). The approach starts by reducing observed characteristics to a single index, the propensity score, which estimates the probability that a sample observation is in the treatment group, given its observed characteristics.

The propensity score can then be used in several ways. Rosenbaum and Rubin (1985) describe techniques for matching that use the propensity score as a distinct matching variable. In many applications, the propensity score serves as the basis for stratification (often into five strata) before comparing the treatment and control groups. Within the strata, the subjects in the treatment and control groups should be comparable. This benefit is a consequence of a theoretical result on propensity scores: if the propensity scores are relatively constant within each stratum, then (within each stratum) the distributions of all the covariates should be approximately the same in the treatment and control groups (Rosenbaum and Rubin, 1983).

Also, the strata based on the propensity score provide a natural setting for examining the relative numbers of participant treatment and nonparticipant control subjects and checking the overlap of their covariate distributions. Strata with higher values of the propensity score will generally have larger sample sizes from the treatment group than from the control group (and conversely). If the sample sizes are too imbalanced, or if the covariate distributions have too little overlap, it becomes clear that the data cannot support the intended comparison.

Thus, propensity score methods, supported by numerous theoretical and applied studies, should offer much promise for dealing with selection bias. They have been used extensively in the public health domain, but very little to date in evaluations of food assistance and nutrition programs.

The caveat. The most troubling aspect of statistical approaches to adjusting for selection bias is that one cannot be certain that the procedure, once applied, has in fact eliminated selection bias. Well-conceived applications of selection bias adjustment models have yielded some plausible and some implausible results in evaluating food assistance and nutrition programs. The situations that produce implausible results cannot be identified a priori, and none of the approaches has consistently yielded plausible results. A plausible adjustment has not necessarily accomplished its purpose just because it is plausible.

Also, when researchers have contrasted the effects estimated in randomized experimental evaluations with those derived from comparing participants with nonparticipants, the two sets of findings have often been divergent. For example, La Londe and Maynard (1987) compared the findings from a randomized experiment to those obtained by using comparable nonparticipants as the Counterfactual and found that none of several methods to identify comparable nonparticipants produced results that were consistent with the experimental findings. Subsequent work argued that specification tests could have led to a result approaching the estimate from the experiment (Heckman and Hotz, 1989). Nonetheless, after decades of research and debate, the statistical community has not yet reached a consensus that any particular approach will consistently remove selection bias.

In addition, data limitations hamper nearly all attempts to counter selection bias. Careful theorizing about the determinants of participation usually suggests many factors that are not measured in existing data sets. Even with special data collection, many of the factors pertain to the time period before the individual began participating (or not participating) and usually cannot be measured reliably on a retrospective basis. (When the situation permits prospective measurement, stronger designs can be employed—see Quasi-Experiment 7, which deals with impact evaluation of program demonstrations.)

Although the extent of any remaining bias cannot be known for sure, testing the robustness of the results is usually informative. A program impact estimate that remains stable under various alternative specifications is somewhat more credible than one that varies dramatically. Of course, if several specifications fail equally to remove the bias, their results will be consistent with one another but inaccurate.

Quasi-Experiment 2: Comparing Participants Before and After Program Participation

Comparing program participants before and after participation is a simple design that eliminates some dimensions of selection bias but has other major vulnerabilities (see box). In this design, subjects are selected into the study before they have been meaningfully exposed to the program. For example, people may be selected as they apply for program services. They are clearly aware of the program at this point and have already taken some action to respond to its requirements, but they have not normally been "exposed" to any of the program's benefits in ways that would affect their status on the outcome dimensions of interest.[8] The subjects' status on the outcome dimensions is measured upon their selection for the study and again after program exposure (long enough after exposure that effects are expected to be visible).

This design is particularly appealing when preprogram data collection can occur as a part of the program's normal administrative process. This can allow collection of a great deal of data—potentially including all participants nationwide for an extended time period—at a low incremental cost. It is not uncommon for social service programs to conduct benchmark or diagnostic measurement as participants enter the program, but unfortunately, none of the major food assistance and nutrition programs applies measures that would support serious outcome evaluation. WIC programs, which collect some measures of nutritional status as a means of assessing nutrition risk, might offer the best opportunity for this approach.

Although this design is usually applied prospectively, it can be applied retrospectively if panel data sets provide appropriate information. The researcher must be able to identify people who participated in the program, determine when they began participating, and have comparable measures of the key outcome dimensions for both the pre- and post-program periods. Note, however, that a data set meeting these requirements would probably contain information on nonparticipants as well. In this case, the researcher would probably incorporate data on nonparticipants, and would actually be using Quasi-Experiment 3.

[8]This may not be true if the program requires some action before enrollment that may itself affect the person's status on outcome variables of interest. Examples would be pre-enrollment requirements such as looking for a job or visiting a doctor.

Quasi-Experiment 2
Participants Before vs. After

Features:

Impact estimate: Difference between post-program measures and pre-program measures for the same program participants.

Key requirement: Impacts must be expected within a short time period.

Advantage: No self-selection bias.

Disadvantage: Temporal bias. Change over time may occur without program.

The Three Questions:

Alike before exposure? Partly. The same people, but different points in their life cycle.

Difference solely from intervention? Probably no. Status would change during program period.

Full force of intervention represented? Usually yes. All are exposed, but only after pre-program measurement.

The Vulnerability: Nonprogram Sources of Change Over Time. Comparing the participant's status before and after participation places the pre-participation situation in the role of the Counterfactual. The design assumes that, in the absence of the program, the individual's pre-program status would not change. If this assumption is valid, the before-vs.-after difference represents the effect of the program. Often, however, this underlying assumption cannot be considered valid.

A prime example of the use of before-vs.-after designs in food assistance and nutrition program research can be found in Yip et al. (1987). They studied infants and preschool children participating in WIC and contrasted hematocrit levels at the time of admission into the program with levels found at the next followup visit a few months later. The data showed a marked decrease in iron deficiency anemia over the few intervening months. Because the time frame was so short, it is unlikely that the effects found by Yip et al. could be attributed to natural developmental processes or to long-term secular declines in iron deficiency anemia among American children.

When program effects are not expected to occur quickly, the assumptions of the before-vs.-after design become more tenuous because forces other than program participation might cause changes in participants' status. For example, normal patterns of child development involve substantial changes in many variables over relatively short periods of time. A related issue is that some conditions improve naturally over time without intervention, a phenomenon known in medical treatment as "spontaneous remission" and in some statistical circumstances as "regression toward the mean."[9] Many people become eligible for program participation in means-tested programs because they have experienced a temporary drop in income. With the passage of time, many such people experience an improvement in income, even if they do not enroll in a program. Accordingly, it would be a mistake to assume that the program causes such post-participation gains in income—or in any conditions affected by income, such as many dimensions of nutrition and health status.

General societal trends may also improve conditions of a target population. These include not only long-term trends, like the general reduction in nutrient deficiencies in the United States, but such short-term phenomena as swings in the unemployment rate or changes in Medicaid coverage. Any before-vs- after period that lasts more than a few months is potentially vulnerable to such temporal effects, and seasonal effects can sometimes occur even within a few months.

Given this vulnerability, the participant before-vs.-after design is useful mainly for evaluating impacts that are expected to be fully visible within a brief period. If temporal effects might be argued to occur, the design can neither refute the possibility nor control for it statistically.

[9] A related issue is measurement error. If a measure is not fully reliable (i.e., capable of producing the same result in repeated applications), a before-vs.-after design may indicate negative results for an individual simply because of measurement error. Special measurement efforts may therefore have to be made with this design. For example, infant development studies often require two independent measures of infant length at each time point because infant length is difficult to measure accurately.

Quasi-Experiment 3 Participants vs. Nonparticipants, Before and After

Features:

Impact estimate: Difference between post-program outcomes for program participants and outcomes for nonparticipants, subtracting out preprogram difference in outcomes.

Key requirement: Obtaining before-vs.-after data on participants and nonparticipants.

Advantage: Strongest of the quasi-experiments.

Disadvantage: Cost of obtaining data.

The Three Questions:

Alike before exposure? No, but the difference is subtracted out.

Difference solely from intervention? Mostly, but the two groups might naturally change at different rates.

Full force of intervention represented? Yes. Participants are exposed, and nonparticipants are not.

Quasi-Experiment 3: Comparing Participants to Nonparticipants Before and After Program Participation

This design combines the strengths of the two previous quasi-experiments. It has less vulnerability to selection bias than the simple comparison of participants to nonparticipants (Quasi-Experiment 1) and less vulnerability to temporal sources of bias than the before-vs- after examination of participants (Quasi-Experiment 2).

In Quasi-Experiment 3, outcomes for participants and nonparticipants must be measured once before participation occurs and again after the effects of participation are expected to be visible. Conceptually, the program's impact is estimated as the post-program difference in outcomes, subtracting out the difference that already existed before participation. This design is therefore commonly called a "difference in differences" or "double difference" design (see box).

In practice, this design is usually applied with multivariate modeling. The dependent variable in the model is often the post-program outcome, with the pre-program outcome measure as a predictor variable, along with participation status. As in the regression adjustment model discussed earlier (Quasi-Experiment 1), the model adjusts for the differing composition of the participant and nonparticipant populations by incorporating covariates that are expected to be related to the outcome measure or to the likelihood of participation.

Practical Requirements. Although this is the strongest of the quasi-experimental designs, it is rarely used to evaluate ongoing entitlement or saturation programs. Because the design calls for pre-participation and post-participation measures on both participants and nonparticipants, data collection can be complicated and very costly.

Imagine, for example, what would be required to evaluate the short-term impact of the FSP on dietary intake, applying this design and relying on primary data collection. The researcher would identify and measure dietary intake for a sample of households that do not currently receive food stamps but might do so in the near future; a few months later, the same households' dietary intake would be measured again. The problem is that people who begin participating in the FSP within a month represent a small fraction of the U.S. population, less than 1 percent.[10] Those households cannot be identified with high reliability in advance. Nor can their counterparts, the households that will be eligible but will not participate. To capture enough actual participants and potentially eligible nonparticipants, data must be collected for a considerably larger pool of households than the required evaluation sample (i.e., there may be several "wasted" interviews for each useful one). Moreover, the larger pool cannot be drawn from a list, but must be screened from a general population sample by obtaining income information. For every household selected for the pool, income information must be collected on several who are not selected. In short, the cost of collecting dietary intake data for the analysis subjects—in itself a costly under-

[10]Around 9 percent of U.S. households currently participate. Historical turnover rates have been in the range of 7-8 percent per month. This implies that the expected number of new households each month would be about 0.7 percent of U.S. households.

taking—may represent only a small fraction of the total data collection cost.

The alternative to primary data collection is to use existing national surveys or administrative data sets. Unfortunately, few data sets containing nutrition and health outcome measures meet the key requirements: permitting identification of participants and eligible nonparticipants and measuring outcomes for both groups before and after the participation period. The major national surveys that collect substantial amounts of nutrition and health outcome data are cross-sectional rather than longitudinal in design.

Quasi-Experiment 4: Aggregate Time Series Analyses

Time series analyses are an important extension of before-and-after studies that can be employed when many observations of outcomes exist for periods before and after program implementation. Unlike simple before-and-after designs, time series analyses take trends into account. Observations that occur before the program is put in place are used to model outcome trends in the absence of the program. The predicted trend represents the Counterfactual, and is contrasted with the trend actually observed after the program is in place. The difference between the two trends is attributed to the program.

In contrast to all the designs discussed previously, time series analysis normally relies on aggregate rather than individual-level data (see box). For example, one might examine annual national statistics on the percentage of low-birthweight births, an outcome that WIC is hypothesized to affect. The low-birthweight rate in any given year might be modeled as a function of previous rates, key demographic variables, economic conditions, and the presence or absence of WIC. (A more complicated version of this analysis, using cross-sectional time series analysis, is described below.) Because time series analysis is conducted at the aggregate level, it can be used with data series that do not offer individual- or household-level data, such as vital health statistics or summary data from administrative or survey series.

Essential to the employment of this design is the existence of a consistent data series extending from before the beginning of a program to a time period after the program is in place. This requirement usually restricts this design to programs on which extensive time series of outcomes can be constructed from administrative data.[11] The ability to distinguish between pre- and post-program time trends increases with the number of observations. More than 20 time points are usually recommended.

Estimating program effect on health or nutrition outcomes through a single time series would be very difficult, and we know of no instance in which it has been done. No data sets with extensive nutrition and health data are collected frequently enough to create a useful series. Moreover, the very large number of factors

Quasi-Experiment 4 Aggregate Time Series Analyses

Features:

Impact estimate: Difference between target population outcomes after program implementation and outcomes predicted by pre-program trends.

Key requirement: Many measures of outcomes before program implementation. Measures of factors potentially affecting outcome.

Advantage: Easy when the data exist.

Disadvantage: Data unavailability: Potential confounding with other factors causing change over time.

The Three Questions:

Alike before exposure? Yes.

Difference solely from intervention? Limited by predictive accuracy of model.

Full force of intervention represented? Limited by program penetration of target population.

[11]The series used need not be confined to administrative data from only one source. Time series analyses that rely on several sources are quite common, using, for example, data on wages obtained from unemployment insurance files, food assistance files, and welfare files. Of course, what is needed in all cases of linking data sets is a set of individual or aggregate identifiers common to all the data sets to be linked.

potentially affecting the nutrition and health status of the population—not only economic and demographic factors, but also changes in knowledge, consumer information, and professional and household practice in the health and nutrition fields—would make estimating a model difficult, even with a fairly substantial number of annual data points in the time series.

Cross-Section Time Series. A potentially more powerful variant of the time series approach is the cross-section time series. This approach uses time series on multiple units, such as series for individual States or counties, rather than for the Nation as a whole.

A good example of cross-section time series analyses of a food assistance and nutrition program can be found in the study undertaken by Rush and colleagues (1988) of the effects of the WIC program on pregnant women. Taking advantage of the rapid growth of the WIC program in the 1970s, Rush and his colleagues conducted a time series analysis of the effect of WIC program growth on birth outcomes. They related the growth of WIC programs in a large number of counties over the period 1972-80 to county aggregate birth outcomes. The research strategy was based on the expectation that, if WIC is effective in improving birth outcomes, improvements ought to be proportional over time to the growth of the WIC program. Using birth registration records and State WIC records, Rush found that the growth of WIC over this period led to increased average birthweight, longer average duration of gestation, and decreased fetal mortality. These effects were over and above the secular trends for this period and were especially pronounced for births to less-well-educated and minority women. The analysis covered 19 States and almost 1,400 counties.

Focus on the "Target" Population. Unlike the analyses discussed previously, time series analyses do not focus on outcomes for program participants. Rather, they focus on some more broadly defined population that can be examined both before and after the program is introduced. Because the unit of aggregation in most data series is some geographic unit, the analysis estimates the program's impact on the overall population of that area. Where a data series is available for a programmatically relevant subpopulation, such as low-income households or pregnant women, the analysis can speak to the impact on that more specific target population.

Estimating impacts for the target population has both advantages and disadvantages. An impact estimate for the target population combines the program's effectiveness in reaching people (its penetration or participation rate) with its effectiveness in helping those it does reach (the impact on participants). Because food assistance and nutrition programs are designed to ameliorate problems in specified target populations, this kind of analysis addresses the question of how well the program is achieving its ultimate objective. However, it risks the possibility that a positive impact on program participants may be so diluted by nonparticipants that it is invisible in the analysis. If the data represent the entire population of an area, including those outside the program's target population, the dilution problem is exacerbated.

Key Limitations. Although the aggregate time series design can be powerful in theory, time series analyses have seldom been applied in the evaluation of food assistance and nutrition programs for two reasons:

First, time series data with sufficient observation points for most nutrition and health outcomes are simply not available. As discussed elsewhere in this report, the most relevant data tend to come from national surveys, many of which provide estimates less often than annually and have been established too recently to provide an adequate pre-program series.

Second, it is often difficult to distinguish the effect of a policy intervention from other influences on a time series trend. The introduction of a new program is seldom the only important event occurring during a year. Other major policy actions, changes in the economic cycle, or even short-term demographic shifts may be at work. If, in addition, several years must pass before the new program has its full effect, that effect may not be separately visible in the time series analyses. These considerations make the cross-section time series design preferable, providing that it can take advantage of differences across locations in the timing and pace of program implementation.

Nonimpact Research for Assessing Ongoing Programs

Although the central focus of this report is on estimating program impacts, it is important to recognize that other aspects of program assessment can also be valuable in evaluating ongoing programs. Research that is not specifically focused on program impact typically employs different research methods than those described previously. Especially important are monitoring studies, participation studies, and program integrity studies, described briefly below.

Monitoring Studies

Monitoring studies (also called surveillance studies) periodically measure outcomes for the participant or target population without specifically attempting to attribute the outcomes to the program. For example, there is considerable interest in altering the nutrient content of school meals to promote compliance with the Dietary Guidelines for Americans. Sampling school meals on a periodic basis can establish whether the trends in meal preparation are in conformity with expectations. Finding that trends are not improving might indicate the need for more programmatic activity, such as an expanded effort to train school food service staff. Alternatively, positive trends might indicate that no program changes are needed. Of course, neither finding substantiates any program effect. An improving situation may mean that the policymaker has less to worry about, which is important information, but additional information is needed before the program can be credited with the improvement.

Participation Studies

Participation studies are typically concerned with issues of targeting and target population penetration. Targeting studies measure the extent to which the actual clients of a program are limited to those intended to be served by the program. Penetration studies address the issue of whether the program reaches a sufficient percentage of its intended target population. Programs that are serving unintended clients are wasting resources, relative to the program objective. And programs that do not have high enough participation levels may not be able to achieve the effects they seek.

Targeting Studies

Targeting studies typically involve surveys of program participants to determine whether their characteristics are as intended. For example, Glantz and his colleagues (1997) surveyed the families of children served in the CACFP. A major finding from this study was that the families of children in the family child care portion of the program had average incomes far above the poverty level. This finding indicated that this portion of the program was weakly targeted. (The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA) contained provisions designed to strengthen the targeting.)

Penetration Studies

Penetration studies have to be designed to represent the entire potential target population in order to ascertain what proportion and types of units are participating in the program. Accordingly, penetration studies tend to be more extensive and expensive than targeting studies. An exception to this rule occurs for programs that are included in ongoing periodic national surveys, such as the Current Population Survey (CPS) or the Survey of Income and Program Participation (SIPP), both run by the Bureau of the Census. Trippe (1995) studied trends in participation in the FSP as shown in the SIPP,[12] and showed that participation rates as well as the overall number of eligible households increased during the period 1988-92.

Other examples of penetration studies reveal serious program flaws. A 1987 provision in FSP legislation extended the use of food stamps to pay for meals in soup kitchens serving the homeless. Surveying a national sample of soup kitchen operators, Burt and Cohen (1988) found that few kitchens had applied to be authorized to accept food stamps, and individuals who used soup kitchens were largely unwilling to pay for food, which was otherwise free, by using food stamps. Because soup kitchens relied heavily on surplus food given without charge to kitchens that did not charge their customers, these organizations could have lost an important subsidy if they had accepted food stamps.

[12]Both the CPS and SIPP have consistently underestimated participation in the Food Stamp Program and other assistance programs.

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).

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.

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.

[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 situa-

tions, 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.)

homes. PRWORA allows full subsidies only for homes located in low-income areas or operated by low-income providers. Other providers receive reduced subsidies, except for individual children who meet a means test.

The law mandated that the changes take effect in July 1997. Because of concerns raised about the possible effect of these changes on children and child care providers, the legislation also mandated an evaluation of the impact. The evaluation mandate was not accompanied by permission to phase in the changes or otherwise to operate the new and old systems in parallel.

Such legislatively mandated, programwide reforms have generally been interpreted as precluding randomized experimentation. Quasi-experimental designs are therefore employed, as described below.

Quasi-Experiment 8 Pre-Change Participants vs. Post-Change Participants

Features:

Impact estimate: Difference between outcomes for program participants before the change and program participants after the change.

Key requirement: Obtaining pre-change outcome measures.

Advantage: The only option available in some cases.

Disadvantage: Unreliable because of confounding with contemporaneous factors.

The Three Questions:

Alike before exposure? Reasonably similar.

Difference solely from intervention? No, other events in same time frame may cause differences.

Full force of intervention represented? Yes, if change is implemented fully.

Quasi-Experiment 8: Comparing Pre-Change Participants to Post-Change Participants

This is one of the simplest and weakest of quasi-experimental designs. Outcome data are collected for all or a sample of program participants before the change is implemented. After the change is implemented, the data collection is repeated, again for all or a representative sample of program participants. Regression adjustment is used in estimating impacts to account for any shifts in the measured characteristics of the participant population that may occur between the pre- and post-change periods (unless the intervention itself is expected to cause such shifts) (see box).

The evaluator's main challenge in this situation is usually to obtain appropriate outcome measures during the pre-change period. Legislative changes often must be implemented quickly, and if the changes have not been anticipated, there may be insufficient time to mount a primary data collection effort. This may require the evaluation to rely on administrative data, one of the large periodic national surveys, or a previous study.[24] In one interesting example, the expectation of welfare reform legislation led to what might be called speculative data collection. Research planning to measure pre-welfare reform outcomes for families on AFDC had to begin several years before PRWORA was passed in 1996—with no firm knowledge about the nature of the changes to be enacted and based only on the firm belief that some sort of welfare reform would be legislated within the next few years (Rossi, 1999).

Even when the data collection challenge is met, this design is very weak. Important national events occurring in the same time period as the program changes may influence the outcomes of interest. The CACFP changes provide a good example. Implementation of these changes coincided with welfare reform and an unprecedentedly strong labor market, both of which are expected to have great influence on the demand for and supply of child care. The pre/post evaluation of the CACFP changes, no matter how carefully designed, will not be able to determine how these major changes influenced the measured difference in outcomes.

[24] In the CACFP example discussed above, pre-change measures of the characteristics of meals served will come from a study conducted several years previously (Fox et al.,1997).

Time Series Analysis

Time series analysis (Quasi-Experiment 4) is an alternative to the simple pre/post comparison in evaluating the impact of programwide changes. It is particularly attractive when the outcomes of interest are measured in data series generated by the program itself. Most examples of this approach, therefore, deal with the impact of program changes on participation. Quite sophisticated time series modeling has been used to estimate the impact of eliminating the purchase requirement in the FSP and the impact of the 1981 eligibility changes in AFDC (Moffitt, 1986). When the outcomes of interest concern nutrition and health status, however, adequate data series are rare.

An important limitation of the time series approach, even when appropriate outcomes are routinely measured, is the need for multiple observations in the post-change as well as the pre-change period. Virtually all data series contain period-to-period fluctuations that are not part of the general trend. When a deviation from the trend line occurs in the last one or two periods of the series, time series models cannot readily determine whether it represents a lasting change from the trend line or a temporary fluctuation. Moreover, if multiple events occur in the same period, such as welfare reform and the CACFP changes, time series analysis is no more effective than simple pre/post analysis, as neither can distinguish the effects of the different events.

Impact Evaluation of Ongoing Program Components

Most policy decisions about ongoing programs are not "go/no go" decisions about the program as a whole, but decisions about whether to modify or eliminate particular program components. A program component of interest can be an element of the service package, such as nutrition education in WIC or employment and training in the FSP. Alternatively, the program component may be a portion of the service delivery mechanism, such as food stamp cashout or electronic benefit transfers.

Many program components offer the advantage (from an evaluation perspective) of not being tightly specified in the authorizing legislation or program regulations. In these situations, State or local operating entities can create variations in the program by virtue of the ways they choose to implement the program component (see box). WIC nutrition education, for example, varies considerably in the frequency with which nutrition education sessions are offered, the topics covered, and the format in which the service is delivered.

Natural or Planned Variation Studies

Features:

Impact estimate: Difference between outcomes for program participants with alternative levels or styles of the intervention.

Key requirement: Numerous representations of each version of the intervention.

Advantage: Can identify "good practices" for replication.

Disadvantage: No estimate of overall impact of the intervention.

The Three Questions:

Alike before exposure? No, but some differences can be taken into account.

Difference solely from intervention? No. Other location-related forces may cause differences.

Full force of intervention represented? Only the difference in interventions is represented.

Natural Variation Studies

The natural variation evaluation, which is applicable in the situation described above, can be considered a "dose response" study. The underlying proposition is that more of the intervention (or more of a particular quality of the intervention) leads to more of the outcome. The design does not yield an estimate of the impact of the intervention itself.[25]

[25]The inference is sometimes drawn that, if a higher level of the intervention has a significant impact relative to a lower level, it also has a significant impact relative to no intervention at all. This inference may be incorrect if, in the absence of the intervention, the participant would have done something different that would have resulted in equivalent or better outcomes.

Because differences in program features are normally introduced by different operating units, these units become key elements of the design. In fact, the design does not literally compare differences in the program; rather, it compares different sites that happen to implement the program differently. Thus, the design is highly vulnerable to confounding the effect of program variations with the effect of other factors that differ among sites, which may range from regional economic characteristics to the abilities of site staff. The only way to reduce this vulnerability is to have multiple sites representing each programmatic variant.

Implementing this design requires first defining meaningful variants of the program component and then identifying a number of sites that implement each variant. Outcomes are then measured for all or a sample of participants in each group of sites. Participant outcomes are modeled as a function of the program variant they face, their site, and an array of participant characteristics.

Planned Variation Studies

To the extent that program legislation and regulations allow program operators discretion in shaping program components, they also open the possibility for planned variation. In a planned variation design, the agency sponsoring the evaluation (or sometimes the evaluator) arranges for the use of specified variants of the program component by particular sites or in particular circumstances.

If planned variation is feasible, a randomized experiment is likely to be possible and is the preferred design. Individuals or aggregates of individuals are randomly assigned among the variants being tested. Differences in outcomes can be attributed to the differences in the program component. If randomized experimentation is precluded, the possibilities include the same array of designs described earlier for evaluating demonstration modifications to ongoing programs.

Parting Words

This report has noted, at several points, that randomized experimentation is the preferred design for impact evaluation in practically all situations. However, the bulk of the discussion has been devoted to the many quasi-experimental designs that are often used in place of randomized experimentation.

Lest the word count distort the message, we must reemphasize here the importance of exerting all possible efforts to use randomized experiments. For programs that deliver services and benefits directly to individuals and families, randomized experimentation is the only design that, properly applied, is guaranteed to produce unbiased estimates of program impact. All other designs are vulnerable to some bias. Their sources of bias can sometimes be described, but the direction and magnitude of the bias cannot be measured reliably. Thus, all the nonexperimental designs have some substantial probability of producing answers that are far from the truth—which can lead to inappropriate policy decisions that may affect millions of people and billions of dollars of public expenditure.

References

Abt Associates Inc. (2000). *Evaluation of the School Breakfast Pilot Project: Study Overview*. Cambridge, MA: Abt Associates Inc.

Bloom, H. S. (1996). *Building a Convincing Test of a Public Housing Employment Program Using Non-Experimental Methods: Planning for the Jobs-Plus Demonstration*. New York, NY: Manpower Demonstration Research Corporation.

Burghardt, J., S. McConnell, A. Meckstroth, et al. (1997). *Implementing Random Assignment: Lessons from the National Job Corps Study*. Princeton, NJ: Mathematica Policy Research, Inc.

Burt, M. R., and B. E. Cohen (1988). *Feeding the Homeless: Does the Prepared Meals Provision Help?* Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Camasso, M. J., C. Harvey, R. Jagannathan, et al. (1998). *A Final Report on the Impact of New Jersey's Family Development Program: Experimental-Control Group Analysis.* New Brunswick, NJ: Rutgers University.

Campbell, J. T., and J. C. Stanley (1963). *Experimental and Quasi-Experimental Designs for Research.* Chicago, IL: Rand McNally.

Chen, H. T. (1990). *Theory-Driven Evaluations.* Newbury Park, CA: Sage Publications.

Devaney, B. (1992). *Very Low Birthweight among Medicaid Newborns in Five States: The Effects of Prenatal WIC Participation.* Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Devaney, B., L. Bilheimer, and J. Schore (1991). *The Savings in Medicaid Costs for Newborns and Their Mothers From Prenatal Participation in the WIC Program, Vol. 2*. Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Devaney, B., and T. Fraker (1989). "The Effect of Food Stamps on Food Expenditures: An Assessment of Findings from the Nationwide Food Consumption Survey," *American Journal of Agricultural Economics*, Vol. 71(1), pp. 99-104.

Fox, M. K., F. B. Glantz, L. Geitz, et al. (1997). *Early Childhood and Child Care Study: Nutritional Assessment of the CACFP, Vol. II.* Alexandria, VA: U.S. Department of Agriculture, Food and Consumer Service.

Fraker, T. M. (1990). *Effects of Food Stamps on Food Consumption: A Review of the Literature.* Washington, DC: Mathematica Policy Research, Inc.

Fraker, T. M., A.P. Martini, J.C. Ohls, et al. (1992). *Evaluation of the Alabama Food Stamp Cash-Out Demonstration, Volume 1: Recipient Impacts.* Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Glantz, F., D. Rodda, M. J. Cutler, et al. (1997). *Early Childhood and Child Care Study: Profiles of Participants in the CACFP, Vol. I*. Alexandria, VA: U.S. Department of Agriculture, Food and Consumer Service.

Gordon, A., and L. Nelson (1995). *Characteristics and Outcomes of WIC Participants and Nonparticipants: Analysis of the 1988 National Maternal and Infant Health Survey.* Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Heckman, J. (1979). "Sample Bias as a Specification Error," *Econometrica*, Vol. 47, pp. 153-62.

Heckman, J., and V. J. Hotz (1989). "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, Vol. 84, pp. 862-80.

Kramer-LeBlanc, C., P. Basiotis, and E. Kennedy (1997). "Maintaining Food and Nutrition Security in the United States with Welfare Reform," *American Journal of Agricultural Economics*, Vol. 79(5), pp. 1600-07.

La Londe, R., and R. Maynard (1987). "How Precise are Evaluations of Employment and Training Programs: Evidence from a Field Experiment," *Evaluation Review*, Vol. 11(4), pp. 428-51.

Levedahl, J.W. (1991). *Effect of Food Stamps and Income on Household Food Expenditures.* Technical Bulletin No. 1794. Washington, DC: U.S. Department of Agriculture, Economic Research Service.

Lipsey, M. (1990). *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage Publications.

Ludwig, J., G. J. Duncan, and P. Hirschfield (1998). *Urban Poverty and Juvenile Crime: Evidence from*

a Randomized Housing-Mobility Experiment. Working Paper, Joint Center for Poverty Research. Chicago, IL: Northwestern University/University of Chicago.

Luepker, R.V., C. L. Perry, S. M. McKinlay, et al. (1996). "Outcomes of a Field Trial to Improve Children's Dietary Patterns and Physical Activity: The Child and Adolescent Trial for Cardiovascular Health (CATCH)," *Journal of the American Medical Association.*, Vol. 275, pp. 768-76.

Metcoff, J., P. Costiloe, W. M. Crosby, et al. (1985). "Effect of Food Supplementation (WIC) During Pregnancy on Birthweight,"*American Journal of Clinical Nutrition*, Vol. 42, pp. 933-47.

Moffitt, R. (1986). *Trends in AFDC Participation Over Time: Evidence on Structural Change (Special Report No. 41)*. Madison, WI: University of Wisconsin, Institute for Research on Poverty.

Murray, D. (1998). *Design and Analysis of Group-Randomized Trials.* New York, NY: Oxford University Press.

Myers, A. F., A. E. Simpson, M. Weitzman, et al. (1989). "School Breakfast Program and School Performance," *American Journal of Public Health*, Vol. 143, pp. 1234-39.

Neenan, P. H., and C. G. Davis (1978). *Impact of the Food Stamp Program on Low Income Household Food Consumption in Florida*. Gainesville, FL: University of Florida, Institute of Food and Agricultural Sciences.

Ohls, J. C., T. Fraker, A.P. Martini, et al. (1992). *Effects of Cash-out on Food Use by Food Stamp Program Participants in San Diego.* Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Orr, L. L. (1999). *Social Experiments: Evaluating Public Programs with Experimental Methods.* Thousand Oaks, CA: Sage Publications.

Orr, L. L., H. S. Bloom, S. H. Bell, et al. (1996). *Does Job Training for the Disadvantaged Work? Evidence from the National JTPA Study.* Washington, DC: Urban Institute Press.

Ponza, M., J. C. Ohls, B. E. Miller, et al. (1996). *Serving Elders at Risk: The Older Americans Act Nutrition Programs, National Evaluation of the Elderly Nutrition Program, 1993-1995.* Washington, DC: U.S. Department of Health and Human Services, Administration on Aging.

Puma, M. J., N. R. Burstein, K. Merrell, et al. (1990). *Evaluation of the Food Stamp Employment and Training Program.* Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Puma, M. J., J. DiPietro, J. Rosenthal, et al. (1991). *Study of the Impact of WIC on the Growth and Development of Children. Field Test: Feasibility Assessment, Final Report.* Cambridge, MA: Abt Associates Inc.

Randall, B., K. Sprague, D. B. Connell, et al. (1999). *WIC Nutrition Education Demonstration Study: Final Report for Prenatal Intervention.* Cambridge, MA: Abt Associates Inc.

Rosenbaum, P., and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, Vol. 70, pp. 41-55.

Rosenbaum, P. R., and D. B. Rubin (1985). "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistician*, Vol. 39, pp. 33-38.

Rossi, P. H. (1999). *Research on PRWORA: What Can be Learned from Four Research Projects Currently Underway*. College Park, MD: University of Maryland, Committee on Welfare Reform Research.

Rubin, D. B. (1979). "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, Vol. 74, pp. 318-28.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley and Sons.

Rush, D., J. Leighton, N. L. Sloan, et al. (1988). "Historical Study of Pregnancy Outcomes," *American Journal of Clinical Nutrition*, Vol. 48, pp. 412-28.

Trippe, C. (1995). "Rates Up: Trends in FSP Participation Rates: 1985-1992." In *Nutrition and Food Security in the Food Stamp Program,* D. Hall and M. Stavrianos, eds. Alexandria, VA: U.S. Department of Agriculture, Food and Consumer Services.

U.S. Department of Agriculture, Office of the Inspector General (1999). *FNS Child and Adult Care Food Program. National Report on Program Abuses. Presidential Initiative: Operation Kiddie Care.*

Weiss, C. H. (1995). "Nothing as Practical as a Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives." In *New Approaches to Evaluating Community Initiatives*, J. P. Connell, A. C. Kubisch, L. Schorr, and C. Weiss, eds. Washington, DC: The Aspen Institute.

Werner, A. (2001 (forthcoming)). *Implementing Research for Welfare Reform*. Madison, WI: University of Wisconsin, Institute for Research on Poverty.

Yip, R., N. J. Binkin, L. Heshood, et al. (1987). "Declining Prevalence of Anemia Among Low-Income Children in the United States," *Journal of the American Medical Association*, Vol. 258, pp. 1619-23.