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.

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.

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

[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.