

Research Methods

This chapter provides an overview of the research methods used in the studies summarized in this report. There are two main sections:

- (1) Evaluation design—Much of this discussion is adapted from Volume 1 of this series, *Research Design* (Hamilton and Rossi, 2002).
- (2) Outcome measures.

Readers with limited knowledge of research design or measurement issues in nutrition and health-related research are encouraged to read this chapter before any of the program-specific chapters that follow and to use it as a technical resource, as needed.

Evaluation Design

The studies reviewed in this report attempted to measure the impact of specific food and nutrition assistance programs (FANPs) on nutrition- and health-related outcomes. 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, generally called the “counterfactual.”

Establishing the counterfactual—that is, estimating what would have happened without a given program—is usually accomplished by examining a population that has not been subjected to the program. What makes the task difficult is the fact that people 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).⁸ 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 ones like 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 of these influences may contribute to selection bias, which distorts the evaluation of a program’s impact.

The sections that follow describe key research designs encountered in FANP research and their various strengths and limitations. In the program-specific chapters that constitute the remainder of this report, the research design used in each study is clearly identified. The text generally includes some discussion of design limitations; however, the present chapter serves as the primary source of information on research methodology.

The Randomized Experiment

There is a strong consensus in the scientific community that only randomized experiments are fully capable of providing reliable estimates of a program’s impacts. The randomized experiment is the “gold standard” of program evaluation.

In the simplest randomized design, potential participants are randomly assigned to either an experimental (or treatment) group, which 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 to be fully comparable at that point. In other words, the two groups are considered to be 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.

The fundamental requirement of randomized experimentation is that program services be deliberately withheld from some people who are otherwise like the

⁸Evaluation designs often focus on units other than people, either aggregations of people (households, students in a school, the population of a county) or operating entities (program offices, schools, businesses). For simplicity of presentation, the present discussion generally refers to individuals rather than aggregations or other entities.

people receiving the service. Such a practice is generally prohibited in entitlement programs because law and regulation require that program benefits be provided to everyone who meets eligibility requirements and takes the necessary steps to qualify. Many FANPs are entitlement programs.

Saturation programs—those with sufficient funding and infrastructure to serve essentially all eligible people—pose similar problems. Whether a potentially eligible person can receive benefits from a nonentitlement program depends on the local availability of program funding and infrastructure. For many nonentitlement programs that approach full saturation, like the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), it can be virtually impossible to find a reasonably representative set of potential participants 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 for research purposes from people who might apply.

Given these challenges, it is not surprising that the literature reviewed for this report included only one study that used a randomized experiment to evaluate the impacts of a specific FANP. This study was completed by Metcalf and his colleagues (1985) during the early years of the WIC program. Random assignment was feasible because, at the time, the demand for WIC participation at the site in which the study was conducted exceeded the available funding.

A few studies have used randomized experiments to estimate the impact of demonstrations or pilot programs, rather than of the FANPs *per se*. These demonstrations typically represented policy initiatives that were tested on a limited scale before full-scale implementation. The most prominent examples are demonstrations of cashing out food stamps (the so-called “cash-out” studies (Fraker et al., 1992; Ohls et al., 1992) and studies of pilot projects in which school breakfasts were offered free to all school children (universal free breakfast projects) (for example, Peterson et al., 2003; McLaughlin et al., 2002; Murphy and Pagans, 2001).

Keep in mind when interpreting results of evaluations of demonstration projects that, in these evaluations, the counterfactual is not the absence of the program. Rather, it is the *status quo*, or the program as it exists without the innovation or modification introduced by the demonstration. Control subjects experience usual

program services but are not offered the new services specified in the intervention. In the case of the food stamp cash-out demonstrations, for example, the evaluations estimated the effects of receiving benefits in the form of checks rather than as food stamps, but not the overall impact of the Food Stamp Program (FSP) itself.

Quasi-Experiments

Virtually all the research that has examined the impact of FANPs on nutrition and health outcomes has identified counterfactual conditions without random selection into treatment and control groups. Such impact evaluation designs are 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. Most of the FANP research reviewed in this report used one of four quasi-experimental designs.

Quasi-Experiment 1: Comparing Participants With Nonparticipants

This design, referred to as “participant vs. nonparticipant” in the program-specific chapters, is the one most commonly used in the research summarized in this report. It calls for identifying comparable groups of participants and nonparticipants and interpreting the average difference in outcomes between the groups as the effect of the program. Nonparticipants must be potentially eligible—that is, people who apparently could have applied and qualified for the program, but did not—to be a credible representation of the counterfactual. In most, but not all, FANP studies, researchers apply an approximation of the means test to identify nonparticipants with incomes below the eligibility cutoff for the program in question.

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 people who have taken the actions necessary to participate in a program are compared with people who have not.

Selection bias often occurs because participants are more highly motivated to achieve the program-relevant outcomes than 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 nutrition advice included in brochures they pick up in the doctor's office—or getting to a doctor's office at all. If this supposition 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 with 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 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 of participants 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, and would be income-eligible for program participation for many years. Other people who qualify for the same programs are not permanently poor, but are at a temporary low point in a fluctuating income pattern. In an earlier period, their income was high enough that they

did not qualify for the program, and at some point, they will regain that level. These two types of people might have similar incomes 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.⁹ 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 WIC-Medicaid study conducted by Devaney et al. (1990 and 1991) to assess the impact of prenatal WIC participation on birth outcomes. Taking advantage of the fact that all Medicaid recipients were automatically eligible for WIC benefits, Devaney and her colleagues contrasted birth outcomes of Medicaid recipients who had participated in WIC during pregnancy with those who had not. The relevant dataset 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. The resulting linked WIC-Medicaid database included approximately 112,000 births to Medicaid mothers in five States over a 2-year period.

To minimize selection bias, Devaney and her associates used regression adjustments. The 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 typically happens, the researchers were limited to the variables available in existing datasets, which seldom measure all of the factors that might create different outcomes for participants and nonparticipants. Alternative attempts to counter selection biases led to quite drastic changes in estimates of the effects, without any clear indications of which attempt was more sensible.

⁹Another technique for dealing with selection bias is the use of propensity scores. Propensity scoring allows a more comprehensive and complex treatment of covariates than is possible with regression adjustment (Hamilton and Rossi, 2002). However, though propensity score methods have been used extensively in public health research, they were not used in the literature reviewed for this report.

Matched Pairs. Sometimes researchers construct a comparison group by matching participants and non-participants on characteristics thought to be related to selection tendencies. For each participant in the research sample, the researcher identifies a nonparticipant with identical or closely similar characteristics on key 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 datasets, 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) 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 non-participants 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 FSP, with benefits measured in dollars and a large number of actual benefit amounts, is the main candidate for dose-response analysis among the FANPs. A number of researchers have used this approach, although with considerable variation in the way it was applied. Some researchers have estimated models that exclude non-participants (for example, Neenan and Davis, 1978; Levedahl, 1991; Kramer-LeBlanc et al., 1997), while others include nonparticipants and specify the model to include both a variable representing the benefit amount and a variable representing participation per se (for example, Fraker, 1990; Devaney and Fraker, 1989).

The dose-response model requires that benefits vary across households that are similar in terms of the factors expected to affect their health and nutrition outcomes. The food stamp situation appears 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. However, 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.

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 two-stage 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 datasets. 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-stage 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 two-stage approaches have been used in evaluating FANPs, but with no clear consensus that any of them can be considered generally reliable. For example, Gordon and Nelson (1995) used three approaches (instrumental variables, Heckman two-step, and simultaneous equations) and a rich dataset to estimate WIC effects on birthweight. 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 Elderly Nutrition Program (ENP). The authors rejected all of the two-stage approaches and based their conclusions on the results of the simple, one-stage regression adjustment.

Caveats to Selection-Bias Adjustment

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

When researchers have compared 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, when La Londe and Maynard (1987) compared the findings from a randomized experiment with those obtained by using comparable nonparticipants as the counterfactual, they found that none of several methods for identifying comparable nonparticipants produced results consistent with the findings from the randomized experiment. However, subsequent work argued that specification tests could have led to a result approaching the estimate from the randomized 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 datasets. Even with special data collection, many of the factors pertain to the time period before the individual began participating (or not participating) and cannot be measured reliably on a retrospective basis.

Although the extent of 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

This simple design (referred to as “participants, before vs. after” in the program-specific chapters) eliminates some dimensions of selection bias but has other major vulnerabilities. Subjects are selected into the study before they have been meaningfully exposed to the program—for example, when 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.¹⁰ The subjects’ status on the outcome dimensions is measured upon their selection into the study and again after program exposure (long enough after exposure that effects are expected to be visible).

The subjects’ preparticipation status serves as the counterfactual. The design assumes that, without the program, the individual’s preprogram status would not change. If this assumption is valid, the before vs. after difference represents the effect of the program.

A prime example of the “participants before vs. after” design in FANP research is the work done by Yip et al.

¹⁰This 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 preenrollment requirements, such as looking for a job or visiting a doctor.

(1987) on anemia among preschool children. Yip and his colleagues studied infants and preschool children participating in WIC and contrasted hematocrit levels at the time of admission into the program with levels at a followup visit a few months later. The data showed a marked decrease in anemia over the few intervening months. Because the time frame was so short, it is unlikely that the effects were attributable to natural developmental processes or to long-term secular declines in 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. 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.¹¹ Many people become eligible for means-tested programs because they have experienced a temporary drop in income. Over time, many such people have an improved 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 within a few months.

Given this vulnerability, the participants 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 also occur, the design can neither refute the possibility nor control for it statistically.

¹¹A related issue is measurement error. If a measure is not fully reliable, that is, is not capable of producing the same result in repeated application, 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.

Although this design is usually applied prospectively, it can be applied retrospectively if panel datasets 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.

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

This design ("participant vs. nonparticipant, before and after") combines the strengths of the two previous quasi-experiments. It has less vulnerability to selection bias than a simple comparison of participants to nonparticipants and less vulnerability to bias from temporal effects than a before vs. after comparison.

In this design, outcomes for participants and nonparticipants are measured once before participation begins 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.

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 previously, 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.

A noteworthy example of this design is a study conducted by Kennedy and Gershoff (1982). The authors compared changes in hemoglobin and hematocrit levels of pregnant WIC participants and nonparticipants between the first and final prenatal visits.

Although this variation 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- and post-participation measures on both participants and nonparticipants, data collection can be complicated and very costly. Moreover, existing national surveys or administrative datasets that collect substantial amounts of nutrition and health outcome data are cross-sectional rather than longitudinal in design.

Quasi-Experiment 4: Time Series Analysis

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 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.

The version of time series analysis that has been used in FANP research is the cross-sectional time series. This approach uses time series on multiple units, such as series for individual States or counties. A good example is the study undertaken by Rush and colleagues (1988) to assess effects of the WIC program. 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 the program's growth on birth outcomes. They related the growth of the WIC program between 1972 and 1980 in a large number of counties to county-aggregate data on 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 its expansion. 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 time period and were especially pronounced for births to less-well-educated and minority women. The analysis covered 19 States and almost 1,400 counties.

Unlike all of the preceding research designs, time series analyses do not focus on outcomes for individual program participants. Rather, they focus on a 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 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 FANPs 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.

Outcome Measures

Existing research has examined the impact of FANP participation on a number of different outcomes. The outcomes are logically sequential, as summarized below, using the FSP as an example.

- *Household food expenditures* is the first outcome in the sequence. The FSP, which provides earmarked economic benefits, can be expected to have a direct impact on the amount of money a household spends on food.
- *Household nutrient availability* is the second outcome. If a household increases the amount of money it spends for food, it is expected to increase the availability to household members of food energy and at least some nutrients.
- *Individual dietary intake* is the next outcome in the sequence. For the FSP, the hypothesis is that increased availability of nutrients in the household leads to increased nutrient intake by individual household members. Programs like WIC and the school nutrition programs, which provide specific foods or meals to participants, are hypothesized to have a direct impact on individual dietary intake.
- *Measures of nutrition and health status other than dietary intake, which FANP participation may influence through the above pathways.* Such measures include, for example, birth outcomes, nutritional biochemistries, linear growth in children, and body weight. Relatively recent research on the School Breakfast Program has expanded this set of outcomes to include measures of school and academic performance.

With the exception of the WIC program and the ENP, relatively few FANP studies have examined the last group of outcomes. Moreover, conclusions from studies that have examined these outcomes must be interpreted with caution. Establishing causality between FANP participation and long-term nutrition and health outcomes requires that data support a logical time sequence. For long-term outcomes (measures that develop over time, such as linear growth and body weight), FANP participation must precede the outcome for a reasonable period and be of sufficient intensity to provide a plausible basis for a hypothesized impact. In addition, reliable assessment of impacts on measures such as linear growth and nutritional biochemistries requires at least two measurements, one before participation and one after. Finally, nutrition and health status are influenced by a complex interplay of diet, heredity, and environment, making the task of determining the specific impacts of FANPs on these long-term outcomes a challenge.

A few studies have examined the impact of FANP participation on health-related behaviors, including, specifically, the impact of the WIC program on breastfeeding and child immunizations and the impact of the ENP on socialization among the elderly.

A potential limitation for all outcome measures used in FANP research is the problem of measurement error. Estimation of key outcomes—including household food expenditures, household nutrient availability, and individual dietary intake—involves collecting detailed data over a day, multiple days, a week, or a month. The data are subject to errors associated with respondents' abilities, cooperation, and recall. These errors are assumed to affect participants and nonparticipants in FANP studies equally; however, the overall effect is a reduction in measurement reliability. In turn, reduced reliability increases the likelihood that differences between participants and nonparticipants will be obscured (Rossi, 1998).

The next sections of this chapter describe key outcome measures used in existing FANP research. Later program-specific chapters also include some discussion of the strengths and limitations of various outcome measures; however, the present chapter serves as the primary source of such information.

Household Food Expenditures

Most of the studies that have examined the impact of FANP participation on household food expenditures

have focused on the FSP. However, a handful of studies have assessed impacts on food expenditures relative to participation in the WIC program, the National School Lunch Program (NSLP), and the Nutrition Assistance Program (NAP) in Puerto Rico.

Although studies of the impact of FANP participation on food expenditures are conceptually similar, they vary substantially in how food expenditures were measured. Some studies were based on money spent on food for at-home use over the course of a week (or weekly food purchases), while others used the monetary value of food eaten out of household supplies over a week or a month. The former measure includes expenditures for foods not necessarily eaten during the week of purchase and excludes the value of foods used from household inventories during the recall period.

Another important difference relates to whether the measure considered expenditures only for food eaten at home or total food expenditures, including meals and snacks eaten away from home. Finally, some measures included the value of purchased food only, while others also included nonpurchased food (for example, home-grown foods and food received as gifts).

Some researchers analyzed expenditures for the household as a whole, while others normalized expenditures to account for the household's size, its age/sex composition, meals eaten away from home, meals served to guests, and/or economies of scale. Commonly used approaches standardize food expenditures based on "equivalent adults" (EAs), counting additional family members less heavily because of economies of scale, "adult male equivalents" (AMEs), counting family members according to caloric requirements, and "equivalent nutrition units" (ENUs), counting family members according to caloric requirements and percentage of meals eaten at home. In general, the more factors considered in normalizing expenditure data, the better. That is, ENUs provide a more precise assessment of expenditures per household member than the more basic EA measure.

In examining the impact of FANP participation on food expenditures, researchers have used both primary data collection and secondary analysis of data collected in national surveys, such as the Consumer Expenditure Survey (CES), the Panel Study of Income Dynamics (PSID), the Nationwide Food Consumption Survey (NFCS), and the Continuing Survey of Food Intakes by Individuals (CSFII). The latter two surveys are no longer conducted.

Household Nutrient Availability

Assessment of household nutrient availability is based on detailed records of household food use for an extended period, usually 1 week. Information on quantities of food withdrawn from the household food supply is translated into nutrient equivalents to represent the food energy and nutrients available to household members. Although household nutrient availability excludes the nutrient content of food eaten away from home, it is still an important measure because the FSP is specifically intended to improve in-home food consumption.

Nonetheless, nutrient availability at the household level is not equivalent to nutrient intake at the individual level. The relationship between the two measures is weakened by several considerations.

- Some household members will get nutrients from foods eaten away from home.
- Some of the food used from household supplies is wasted.
- Household members may unequally consume nutrients from household food supplies, relative to their needs, depending on their tastes and appetites.

Moreover, increased availability of food energy and nutrients at the household level does not necessarily translate into better diets—for example, lower intakes of nutrients and food components that tend to be overconsumed by many Americans (fat, saturated fat, cholesterol, and sodium) or greater adherence to recommended patterns of food intake (for example, eating fruits and vegetables or whole grains). For these reasons, one must examine the dietary intakes of individual household members to adequately assess nutrition-related impacts of the FSP.

In assessing household nutrient availability, the amount of energy and nutrients available in the foods withdrawn from the household food supply is evaluated relative to the Recommended Dietary Allowances (RDAs) and the household's size and composition. Household nutrient requirements are generally defined based on adult male equivalents (AMEs), which take into consideration the number of individuals in the household and their differing nutrient requirements based on age, gender, and pregnancy/lactation status, or equivalent nutritional units (ENUs), which further adjust for the number of meals each family member eats at home and the number of meals served to guests.

All studies of impacts on household nutrient availability have focused on the FSP. Research has included both primary data collection and secondary analysis of national survey data. Most of the secondary analyses used data from the 1977-78 NFCS (low-income supplemental sample) or data from a followup NFCS low-income sample that was collected in 1979-80.

Individual Dietary Intake

A number of techniques can be used to assess individual dietary intake (Thompson and Byers, 1994). In research on FANP impacts, the technique used most often is a single 24-hour recall or a single-day food record. Some studies collected multiple days of data, ranging from 2 to 7 days, using recalls, records, or a combination approach. Respondents usually reported on their own intakes, but parents or other caregivers served as proxy respondents for infants and young children.

Although all dietary data collection techniques have limitations, it is generally accepted that the more days of data available, the better the measure.¹² In addition, food records are generally believed to be more accurate than recall-only methods because respondents, at least in theory, record food intake on a prospective basis rather than recalling it retrospectively and have the opportunity to measure or carefully observe portions. Food records impose a significant response burden, however, and are particularly problematic for respondents with limited literacy. Moreover, the need to record food intake may alter respondents' eating behavior. For these reasons, recall-based data collection is preferred for assessment of low-income populations.

The 24-hour recall has three key disadvantages. First and most obvious, the method relies on memory, which tends to be imperfect. Second, 24-hour recalls have been shown to be subject to systematic underreporting by some subgroups, including individuals who are overweight (Briefel et al., 1997) and the elderly (Madden et al., 1976). Third, because intakes vary so much from day to day in highly industrialized countries, such as the United States, a single day's intake is unlikely to be representative of the respondent's usual diet (Beaton, 1983).

The accuracy of 24-hour recall data can be improved by careful, standardized interviewing techniques.

¹²There are limitations, however. Experience has shown that quality and completeness of data decrease as the number of days increases. Respondents tend to fill out records less carefully as time goes on, after approximately 4 or 5 days (Gersovitz et al., 1978).

Computer-assisted interviewing is one way to achieve a high level of standardization. One of the first applications of computer-assisted 24-hour recalls was developed for the third National Health and Nutrition Examination Survey (NHANES-III) (McDowell et al., 1989). The approach was refined and improved, based on methodological research, to better engage respondents in the interview process and to provide memory cues for accurate recall of food and beverage consumption (Moshfegh et al., 2001). A version of the improved system was used to collect data for the 1994-98 CSFII, and the final version is being used to collect data in NHANES-IV. A comparable system is included in the Nutrition Data System (NDS), managed by the Nutrition Coordinating Center (NCC) at the University of Minnesota (NCC, 2001).

Recent guidelines for dietary assessment issued by the Institute of Medicine (IOM, 2001) recommend that studies examining dietary intakes of groups collect a minimum of 2 nonconsecutive days or 3 consecutive days of data for a subgroup of the population(s) being studied. The additional data for the subgroup(s) can be used to adjust intake distributions for day-to-day, within-person variation (IOM, 2001).¹³ The adjustments provide reliable estimates of usual energy and nutrient intakes. These improved dietary assessment methods are just beginning to appear in FANP research (McLaughlin et al., 2002).

Nutrient estimates generated from dietary intake data generally include only the nutrients provided by the foods and beverages consumed. While studies may collect information on use of vitamin and mineral supplements, the contributions of supplements are seldom included in the estimates.¹⁴ None of the studies reviewed for this report included contributions from supplements.

Comparison to Reference Standards

Most studies that have examined the impact of FANPs on nutrient intakes assessed intakes in reference to

established intake standards rather than just comparing raw intakes in kilocalories, milligrams (mg) or grams (gm). At the time most of these studies were conducted, the standards used were the Recommended Dietary Allowances (RDAs) (National Research Council (NRC), 1989a). More recent studies have also used the *Dietary Guidelines for Americans* (U.S. Departments of Agriculture (USDA) and Health and Human Services (HHS), 2000). A few studies used the Healthy Eating Index (HEI) as a summary measure of dietary quality (Kennedy et al., 1995). Each of these reference standards is discussed in turn below.

Recommended Dietary Allowances. Most FANP researchers compared mean intakes of participants and nonparticipants, expressed as a percentage of age- and gender-appropriate RDAs. Some researchers compared the proportion of individuals in each group with intakes below a defined cutoff, generally between 70 and 100 percent of the RDA. The latter approach is less common, perhaps because an expert panel convened in the early 1980s by USDA specifically recommended against the use of fixed cutoffs relative to the RDAs as a means of assessing the prevalence of inadequate intakes (NRC, 1986).

In assessing program impacts, researchers generally deemed a significantly greater mean intake among participants or a significantly greater percentage of participants with intakes above a specified cutoff as evidence of a positive program effect. Effects were characterized as program participation leading to “increased intake(s).”

Although these interpretations are common in the available literature, differences in the mean percentage of the RDA consumed, or in the proportion of individuals consuming some percentage of the RDA, do not provide information on the underlying question: Is the percentage of FANP participants with adequate diets different than the percentage of nonparticipants with adequate diets? Even when mean nutrient intake of a group approximates or exceeds the RDA, significant proportions of the population may have inadequate intakes. On the other hand, use of RDA-based cutoffs seriously overestimates the proportion of a group at risk of inadequate intake because, by definition, the RDA exceeds the needs of nearly all (97-98 percent) healthy individuals in the group (IOM, 2001).

Thus, the available research provides an imperfect picture of both the prevalence of inadequate intakes and the substantive significance of differences in intakes of

¹³Adjustment of intake distributions is necessary to develop accurate estimates of the proportion of the population with inadequate intakes. If research goals are limited to estimates of mean intake for each group, additional days of data are not necessary as long as sample sizes are sufficient (IOM, 2001).

¹⁴This trend may be changing. There is increasing interest in basing assessment of nutrient intake on complete intake data, including vitamins and minerals provided by supplements. However, because supplement use can be intermittent and because most extant data have inconsistent reference time periods for dietary intake data and supplement data (previous 24-hours vs. use during preceding month or week), combining the two sources of data is not a straightforward task.

FANP participants and nonparticipants. That is, the available data provide information on whether FANP participants have “increased intakes” of food energy or key nutrients relative to nonparticipants but do not provide any information on whether these differences affect the likelihood that FANP participants consume adequate amounts of food energy or nutrients.

This imperfect picture of the risk of inadequacy reflects a limitation in the reference standards and dietary assessment methods available when most of the existing FANP research was conducted, rather than shortcomings in the research per se. This limitation has been addressed in the Dietary Reference Intakes (DRIs), a revised set of nutrient intake standards that has replaced the RDAs (IOM, 1999, 2000a, 2000b, 2002a, 2002b).

The development of the DRIs has led to statistically based guidance on estimating the prevalence of inadequate intakes of population groups (IOM, 2001). The recommended approach, referred to as the “EAR cut-point method,” differs in two important ways from the approach used in previous research. First, assessment of adequacy is based on the Estimated Average Requirement (EAR) rather than the RDA. The EAR is the level of intake estimated to meet the requirements of half of the healthy individuals in a given gender and life-stage group.¹⁵ It was developed specifically to provide a better standard for assessing the adequacy of nutrient intakes than is possible with the RDA.

Second, assessment is based on estimates of *usual* rather than observed intakes. As discussed above, estimation of usual intakes requires collecting 2 nonconsecutive or 3 consecutive days of intake data for a subgroup of the population(s) under study. These data are then used to adjust the distribution of intakes to remove within-person variation and better represent usual intake patterns.

Compared with estimates from previous research, the recommended approach to estimating the prevalence of inadequate intakes is likely to yield lower estimates of the prevalence of inadequacy because, as noted, using the RDA as a reference point for assessing adequacy

always leads to an overestimation of the problem (IOM, 2001).¹⁶ Similarly, using observed intakes rather than usual intakes tends to overestimate the percentage of individuals falling below a given cutoff because the distribution of observed intakes is usually wider than the distribution of usual intakes.

At the time this report was finalized, only one FANP study had used the EAR cut-point method to estimate the effect of FANP participation on the prevalence of inadequate intakes (McLaughlin et al., 2002). Applying the EAR, in combination with data on usual intakes, is not as straightforward as one might expect because (1) the procedures used to estimate usual intakes adjust *distributions* rather than individual estimates and (2) the IOM specifically cautions against using a binary variable to represent inadequacy in a standard regression model (IOM, 2001). The DRI applications report outlines an analysis strategy for assessing the impact of FANP participation on the prevalence of inadequate intakes (IOM, 2001).

Dietary Guidelines for Americans. The *Dietary Guidelines for Americans* (DGAs) were developed specifically to provide consumers with recommendations that could be used to plan healthful diets (USDA/HHS, 2000). The DGAs have been revised over the years but have always stipulated moderate intake of fat, saturated fat, cholesterol, and sodium.

Relatively few FANP studies have used the DGAs to assess dietary intakes of program participants vs. nonparticipants. Most research that has used the DGAs compared intakes of total fat and/or saturated fat, as a percentage of total energy intake, to DGA recommendations. Because early versions of the DGAs did not include quantitative recommendations for cholesterol and sodium intake, most studies used recommendations from the NRC, which include a maximum of 300 mg per day for cholesterol and a maximum of 2,400 mg per day for sodium (NRC, 1989b). The NRC recommendations for these two nutrients, which were incorporated into guidelines for nutrition labeling, are the ones now included in the DGAs.

The DRIs have defined a new reference standard for intake of total fat, referred to as an Acceptable

¹⁵For some nutrients, most notably calcium, available data were insufficient to establish an EAR. In these instances, a different DRI—an Adequate Intake or AI—was established. The AI is a level of intake that is assumed to be adequate, based on observed or experimentally determined estimates of intake. The DRIs also define ULs (Tolerable Upper Intake Levels) for selected nutrients. The UL is the highest intake likely to pose no risk of adverse health effects. The DRI applications report provides guidance on appropriate uses of AIs and ULs in assessing nutrient intakes of groups (IOM, 2001).

¹⁶For some nutrients, the estimated prevalence of inadequate intakes would be lower even if the old approach was replicated using the latest RDAs because the new RDAs for some nutrients differ substantially from previous RDAs. For example, for children ages 1-3, the 1989 RDAs for zinc and vitamin C were, respectively, 10 mg and 40 mg. The new RDAs for these nutrients are substantially lower, at 3 mg (zinc) and 15 mg (vitamin C).

Macronutrient Distribution Range (AMDR) (IOM, 2002b). AMDRs have also been defined for carbohydrates, protein, and specific types of polyunsaturated fatty acids. AMDRs have not been defined for saturated fat or cholesterol because these dietary components have no known beneficial effect in preventing chronic disease and are not required at any level in the diet (IOM, 2002b). DRIs for electrolytes, including sodium, are currently in development (IOM, 2003).

The Healthy Eating Index. Very few FANP studies have examined impacts of FANP participation on the HEI. Developed by USDA's Center for Nutrition Policy and Promotion (CNPP), the HEI is a summary measure of overall diet quality (Kennedy et al., 1995). It is based on 10 component scores, all of which are weighted equally in the total score. The component scores measure different aspects of a healthy diet, based on current public health recommendations. Five of the component scores are food-based and evaluate food consumption compared with the recommendations of the Food Guide Pyramid for grains, vegetables, fruits, dairy, and meat (USDA/CNPP, 1996). Four component scores are nutrient-based and assess compliance with the DGA recommendations for intake of fat and saturated fat (USDA/HHS, 2000), as well as with the NRC recommendations for intake of cholesterol and sodium (NRC, 1989b). The 10th component score is food-based and assesses the level of variety in the diet. Dietary variety is stressed in the Food Guide Pyramid, the Dietary Guidelines, and the NRC's diet and health recommendations (Basiotis et al., 2002).

Health-Related Behaviors

Breastfeeding

A handful of studies have examined breastfeeding initiation and duration among WIC participants and non-participants. Initiation is generally defined as ever having breastfed, regardless of frequency or duration. Duration is measured as total length of time and/or as the percentage of mothers who breastfed for 6 months or more.

Socialization Among the Elderly

The ENP was designed to address the psychological and sociological needs of the elderly as well as their nutritional needs. Studies that have compared socialization among ENP participants and nonparticipants have used two different approaches. Two studies classified respondents based on a five-point isolation index: (1) living alone, (2) reporting having too few

friends, (3) having no one to confide in, (4) having children that do not visit them, and (5) feeling lonely more often. A third study defined socialization based on number of social contacts (with relatives and/or friends) per month.

Other Measures of Nutrition and Health Status

While the majority of studies of the impact of FANPs on nutrition- and health-related outcomes have focused on food expenditures, household nutrient availability, and/or individual dietary intake, some studies have examined impacts on longer term measures of nutrition and health status. The most studied outcome in this group is birthweight (and related measures). Others include measures of food sufficiency/security/insecurity, nutritional biochemistries, linear growth in children, body weight, and school/academic performance.

Birthweight and Related Measures

Impacts of FANP participation on birthweight—perhaps the most fundamental measure of nutrition and health status in infants—has focused almost exclusively on the WIC program. This is an obvious and appropriate focus, given that one of the issues WIC was specifically designed to address is birth outcomes among low-income pregnant women. Note, however, that birthweight reflects multiple influences exerted both before and during a pregnancy. These include, but are not limited to, maternal health and nutrition, intrauterine exposures (tobacco, drugs, alcohol), and genetic factors.

Compared with the measures discussed in the previous sections, reliable and complete data on birthweight, which is routinely measured at birth and recorded on the birth certificate, is easy to collect. However, proper interpretation of data on birthweight depends on relating birthweight to the expected weight for the infant's gestational age (duration of pregnancy). Infants who are below the expected weight are classified as having intrauterine growth retardation (IUGR). IUGR infants are at increased risk for adverse birth outcomes, compared with those of low birthweight whose weight is appropriate for their gestational age. Infants born at full term (39+ weeks) with a birth weight of less than 2,500 gm (5.5 pounds) are classified as IUGR.

Another issue that affects interpretation of data on birthweight is the simultaneity of WIC participation and gestational age. Women who deliver early have less chance of enrolling in WIC than women who go to term. Consequently, both the decision to participate

in WIC and the length of participation are inexorably linked with gestational age, an important predictor of most birth outcomes. Moreover, women who enroll late in pregnancy will automatically have better outcomes than other women by virtue of their increased gestation. This simultaneity means that assessments of the impact of WIC on birthweight that rely on a binary indicator of participation are likely to overstate the impact of the program. Moreover, because the duration of WIC participation is also simultaneous with gestational age, a traditional dose-response approach employed by several studies—estimating WIC impacts based on number of months of WIC participation—is not a viable solution to the problem.

Gordon and Nelson (1995) studied several approaches to addressing the relationship between the timing of WIC enrollment and gestational age. These approaches included omitting very late WIC enrollees (enrolled after the eighth month), including gestational age as an independent variable in the birthweight regression, and defining several cohorts of WIC participants by gestational age (pregnancy duration) at the time of WIC enrollment. The authors found, however, that these approaches systematically underestimated the impact of WIC and suggested that results from analyses using a binary indicator (participant vs. nonparticipant) and results of analyses that compare various cohorts of WIC participants (e.g., early vs. late enrollees) bound the likely magnitude of the effect.

Food Security

In 1997, USDA released the 18-item Federal food security module, the currently accepted standard for measuring household and individual food security (Price et al., 1997; Bickel et al., 2000). Studies completed before 1997 used one or more of the questions included in the early food security assessment work done by Wehler et al. (1991) and by Radimer and her colleagues at Cornell University (1992). Studies completed after 1997 used either the early questions or the 18-item module.

Nutritional Biochemistries

Several studies have examined the impact of FANP participation on blood levels of key nutrients. The nutrient studied most often is iron. Iron deficiency is the most common known form of nutritional deficiency, affecting the entire age span from infancy to old age. In infancy and early childhood, iron deficiency is an especially important problem that may be associated with anemia as well as with delayed psychomotor development (de Andraca et al., 1997).

Impacts on nutritional biochemistries are best assessed using a design that compares participants (and potentially nonparticipants) before and after FANP participation. As described earlier, Yip and his colleagues (1987) conducted a widely recognized study of the impact of WIC on the prevalence of anemia among young children, using a classic “participants, before vs. after” design.

Studies that rely on single measures of iron status (or other nutritional biochemistries) are subject to significant selection bias, particularly WIC studies because low blood levels of iron and other nutrients are used to define eligibility for WIC participation.

Linear Growth in Children

One of the most fundamental measures of health status in preschool children is the attainment of normal growth. Failure to attain normal linear growth (stunting) is a highly sensitive indicator of underlying nutritional deficits or other health problems. Height-for-age is used to assess the adequacy of linear growth, relative to growth curves established by the Center for Disease Control and Prevention (CDC) (Kuczmarski et al., 2002). Height-for-age below the fifth percentile is indicative of growth retardation (HHS, 2000).

Similar to nutritional biochemistries, proper assessment of the impact of FANP participation on measures of linear growth in children requires at least two measurements, ideally collected for both treatment and control groups (World Health Organization, 1995). For example, children of Asian descent, many of whom came into the United States as refugees in the late 1970s and early 1980s, had an increased prevalence of growth stunting relative to other children in the WIC program. Over time, coincident with participation in WIC, the prevalence of stunting decreased significantly to levels approaching those of other low-income children served by WIC (Yip et al., 1993).

Body Weight

The substantial increase in the prevalence of overweight and obesity in the United States over the past several decades has heightened interest in this aspect of nutritional status among low-income Americans. Few studies have attempted to estimate the impact of FANP participation on this indicator, and none has studied the issue adequately. Development of overweight and obesity is a complex process that takes place over a long period and is influenced by a number of factors other than dietary intake, including levels of

physical activity/inactivity and genetics. Moreover, low-income and food-insecure individuals are more likely to be overweight or obese than higher income and food-secure individuals. This confounding makes it difficult to assess relationships between these characteristics using cross-sectional data.

For adults, overweight and obesity are defined based on body mass index (BMI), a measure of the relationship between height and weight that is commonly accepted for classifying adiposity (or fatness) in adults (CDC, 2003).¹⁷ For adults, a healthy weight is defined as a BMI of at least 18.5 but less than 25. Overweight is defined as a BMI between 25 and 30, and obesity as a BMI of 30 or more. A BMI of less than 18.5 indicates extreme thinness or underweight.

Classifying children as overweight is fundamentally different from classifying adults (Cole, 2001). Adults have traditionally been classified as overweight based on life insurance mortality data and data relating weight status to morbidity and mortality (Troiano and Flegal, 1998). These criteria cannot be used to define

overweight in childhood, however, because childhood mortality is not associated with weight, and weight-related morbidity in childhood is too infrequent to define meaningful cutoffs (Cole, 2001). Therefore, children are classified as overweight by comparing their weights and heights with appropriate reference populations. For children, overweight is defined as a BMI at or above the 95th percentile on CDC growth charts, which define BMI percentile distributions by age and gender (CDC, 2003). Children with BMIs between the 85th and 95th percentiles are considered to be at risk of overweight.

School Performance

A relatively recent body of research has examined impacts of breakfast consumption on school performance. Virtually all of these studies have evaluated the issue within the context of demonstration projects of “universal free” school breakfast programs—that made breakfast available to all students free of charge, regardless of household income. Measures examined include attendance and tardiness, academic achievement—generally measured with standardized test scores—cognitive functioning, student behavior, and referrals to school nurses.

¹⁷BMI is equal to [weight in kilograms] / [height in meters]².

References

- Basiotis, P., A. Carlson, S. Gerrior, et al. 2002. *The Healthy Eating Index: 1999-2000*. CNPP-12. USDA, Center for Nutrition Policy and Promotion.
- Beaton, G.H., J. Milner, V. McGuire, et al. 1983. "Sources of Variation in 24-hour Recall Data: Implications for Nutrition Study Design and Interpretation—Carbohydrate Sources, Vitamins and Minerals," *American Journal of Clinical Nutrition* 37:895-96.
- Bickel, G., M. Nord, C. Price, et al. 2000. *Guide to Measuring Household Food Security: Revised 2000*. USDA, Food and Nutrition Service.
- Briefel, R.R., C.T. Sempos, M.A. McDowell, et al. 1997. "Dietary Methods Research in the Third National Health and Nutrition Examination Survey: Underreporting of Energy Intake," *American Journal of Clinical Nutrition* 65(4, Supplement):1203S-09S.
- Campbell, J.T., and J.C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.
- Centers for Disease Control and Prevention. 2003. "CDC Growth Chart Training Modules." Available: <http://www.cdc.gov/nccdphp/dnpa>. Accessed May 2003.
- Cole, N. 2001. *The Prevalence of Overweight Among WIC Children*. USDA, Food and Nutrition Service.
- de Andraca, I., M. Castillo, and T. Walter. 1997. "Psychomotor Development and Behavior in Iron-Deficient Anemic Infants," *Nutrition Reviews* 55(4):125-32.
- Devaney, B., L. Bilheimer, and J. Schore. 1990. *The Savings in Medicaid Costs for Newborns and Their Mothers from Prenatal WIC Participation in the WIC Program: Volume 1*. USDA, Food and Nutrition Service.
- Devaney, B., L. Bilheimer, and J. Schore. 1991. *The Savings in Medicaid Costs for Newborns and Their Mothers from Prenatal WIC Participation in the WIC Program: Volume 2*. USDA, 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* 71(1):99-104.
- Fraker, T.M. 1990. *Effects of Food Stamps on Food Consumption: A Review of the Literature*. USDA, Food and Nutrition Service.
- Fraker, T.M., A.P. Martini, J.C. Ohls, et al. 1992. *The Evaluation of the Alabama Food Stamp Cash-out Demonstration: Volume 1: Recipient Impacts*. USDA, Food and Nutrition Service.
- Gersovitz, M., J.P. Madden, and H. Smiciklas-Wright. 1978. "Validity of the 24-hour Dietary Recall and Seven-day Food Record for Group Comparisons," *Journal of the American Dietetic Association* 73:48-55.
- Gleason, P., A. Rangarajan, and C. Olson. 2000. *Dietary Intake and Dietary Attitudes Among Food Stamp Participants and Other Low-Income Individuals*. USDA, Food and Nutrition Service.
- Gleason, P., and C. Sutor. 2001. *Children's Diets in the Mid-1990s: Dietary Intake and Its Relationship with School Meal Participation*. USDA, Food and Nutrition 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*. USDA, Food and Nutrition Service.
- Hamilton, W.L., and P.H. Rossi. 2002. *Effects of Food Assistance and Nutrition Programs on Nutrition and Health: Volume 1, Research Design*. FANRR-19-1. USDA, Economic Research Service.
- Heckman, J. 1979. "Sample Bias as a Specification Error," *Econometrica* 47:153-62.
- 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* 84:862-80.
- Institute of Medicine. 2003. "The Current Project System—Project Title: Dietary Reference Intakes for Electrolytes and Water." Available: <http://www.iom.edu/project.asp?id=3969>. Accessed July 2003.
- Institute of Medicine, Food and Nutrition Board. 2002a. *Dietary Reference Intakes: Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. Washington, DC: National Academy Press.

- Institute of Medicine, Food and Nutrition Board. 2002b. *Dietary Reference Intakes: Energy, Carbohydrates, Fiber, Fats, Fatty Acids, Cholesterol, Protein, and Amino Acids (Macronutrients)*. Washington, DC: National Academy Press.
- Institute of Medicine, Food and Nutrition Board. 2001. *Dietary Reference Intakes: Application in Dietary Assessment*. Washington, DC: National Academy Press.
- Institute of Medicine, Food and Nutrition Board. 2000a. *Dietary Reference Intakes: Thiamin, Riboflavin, Niacin, Vitamin B₆, Folate, Vitamin B₁₂, Pantothenic Acid, Biotin, and Choline*. Washington, DC: National Academy Press.
- Institute of Medicine, Food and Nutrition Board. 2000b. *Dietary Reference Intakes: Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press.
- Institute of Medicine, Food and Nutrition Board. 1999. *Dietary Reference Intakes: Calcium, Phosphorus, Magnesium, Vitamin D, Fluoride*. Washington, DC: National Academy Press.
- Kennedy, E.T., and S. Gershoff. 1982. "Effect of WIC Supplemental Feeding on Hemoglobin and Hematocrit of Prenatal Patients," *Journal of the American Dietetic Association* 80:227-30.
- Kennedy, E.T., J. Ohls, S. Carlson, et al. 1995. "The Healthy Eating Index: Design and Applications," *Journal of the American Dietetic Association* 95(10):1103-09.
- 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* 79(5):1600-07.
- Kuczmarski, R., C. Ogden, L. Guo, et al. 2002. *2000 CDC Growth Charts for the United States: Methods and Development*. Vital and Health Statistics Series 11, No. 246. Centers for Disease Control and Prevention, National Center for Health Statistics.
- La Londe, R., and R. Maynard. 1987. "How Precise are Evaluations of Employment and Training Programs: Evidence from a Field Experiment," *Evaluation Review* 11(4):428-51.
- Levedahl, J.W. 1991. *The Effect of Food Stamps and Income on Household Food Expenditures*. TB-1794. USDA, Economic Research Service.
- Madden, J.P., S.J. Goodman, and H.A. Guthrie. 1976. "Validity of the 24-hour Recall: Analysis of Data Obtained from Elderly Subjects," *Journal of the American Dietetic Association* 68:143-147.
- McDowell, M.A., R.R. Briefel, R.A. Warren, et al. 1989. "The Dietary Data Collection System: An Automated Interview and Coding System for NHANES-III," in P.J. Stumbo (ed.), *Proceedings of the Fourteenth National Nutrient Databank Conference*. Ithaca, NY: The CBORD Group, Inc.
- McLaughlin, J., L. Bernstein, M.K. Crepinsek, et al. 2002. *Evaluation of the School Breakfast Program Pilot Project: Findings from the First Year of Implementation*. USDA, Food and Nutrition Service.
- Metcoff, J., P. Costiloe, W.M. Crosby, et al. 1985. "Effect of Food Supplementation (WIC) During Pregnancy on Birth Weight," *American Journal of Clinical Nutrition* 41:933-47.
- Moshfegh, A.J., N.R. Raper, L.A. Ingwersen, et al. 2001. "An Automated Approach to 24-hour Dietary Recall Methodology," *Annals of Nutrition and Metabolism* 45(1, Supplement):156 (2.04.157).
- Murphy, J.M., and M. Pagano. 2001. *Effects of a Universally Free, In-Classroom School Breakfast Program: Final Report from the Third Year of the Maryland Meals for Achievement Evaluation*. Boston, MA: Massachusetts General Hospital.
- National Research Council. 1989a. *Recommended Dietary Allowances, 10th edition*. Washington, DC: National Academy Press.
- National Research Council. 1989b. *Diet and Health: Implications for Reducing Chronic Disease Risk*. Washington, DC: National Academy Press.
- National Research Council. 1986. *Nutrient Adequacy: Assessment Using Food Consumption Surveys*. Washington, DC: National Academy Press.
- 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.

- Nutrition Coordinating Center. 2001. *Nutrition Data System for Research*. Minneapolis, MN: University of Minnesota.
- Ohls, J.C., T.M. Fraker, A.P. Martini, et al. 1992. *The Effects of Cash-out on Food Use by Food Stamp Program Participants in San Diego*. USDA, Food and Nutrition Service.
- Peterson, K., M. Davison, K. Wahlstrom, et al. 2003. *Fast Break to Learning School Breakfast Program: A Report of the Third Year Results, 2001-2002*. Minneapolis, MN: University of Minnesota.
- Ponza, M., J.C. Ohls, B.E. Millen, et al. 1996. *Serving Elders at Risk: The Older Americans Act Nutrition Programs, National Evaluation of the Elderly Nutrition Program, 1993-1995, Volumes I, II, and III*. U.S. Department of Health and Human Services, Administration on Aging.
- Price, C., W.L. Hamilton, and J.C. Cook. 1997. *Guide to Implementing the Core Food Security Module*. USDA, Food and Nutrition Service. (The guide was revised and updated in 2000; see Bickel et al., 2000).
- Radimer, K., C. Olson, J.C. Greene, et al. 1992. "Understanding Hunger and Developing Indicators to Assess it in Women and Children," *Journal of Nutrition Education* 24(1, Supplement):36s-44s.
- Rossi, P.H. 1998. *Feeding the Poor: Assessing Federal Food Aid*. Washington, DC: The AEI Press.
- Rubin, D.B. 1979. "Using Multivariate Matched Sampling and Regression Adjustment To Control Bias in Observational Studies," *Journal of the American Statistical Association* 74:318-28.
- Rush, D., J. Leighton, N. Sloan, et al. 1988. "Historical Study of Pregnancy Outcomes," *American Journal of Clinical Nutrition* 48:412-28.
- Thompson, F.E., and T. Byers. 1994. "Dietary Assessment Resource Manual," *Journal of Nutrition* 124(11, Supplement).
- Troiano, R.P., and Flegal, K.M. 1998. "Overweight Children and Adolescents: Description, Epidemiology, and Demographics," *Pediatrics* 101:497-504.
- U.S. Department of Agriculture, Center for Nutrition Policy and Promotion. 1996. *The Food Guide Pyramid*. Home and Garden Bulletin No. 252.
- U.S. Department of Agriculture and U.S. Department of Health and Human Services. 2000. *Nutrition and Your Health: Dietary Guidelines for Americans, 5th edition*. Home and Garden Bulletin No. 232.
- U.S. Department of Health and Human Services. 2000. *Healthy People 2010: Understanding and Improving Health, 2nd Edition*.
- Wehler, C.A., R.I. Scott, and J.J. Anderson. 1991. *A Survey of Childhood Hunger in the United States. Community Childhood Hunger Identification Project*. Washington, DC: Food Research and Action Center.
- World Health Organization. 1995. *Physical Status: The Use and Interpretation of Anthropometry*. Geneva, Switzerland: World Health Organization, WHO Technical Report Series, No. 854.
- Yip, R., N. Binkin, L. Fleshood, et al. 1987. "Declining Prevalence of Anemia Among Low-income Children in the United States," *Journal of the American Medical Association* 258(12):1619-23.
- Yip, R., K. Scanlon, and F. Trowbridge. 1993. "Trends and Patterns in Height and Weight Status of Low-income U.S. Children," *Critical Reviews in Food Science and Nutrition* 33(4/5):409-21.