United States Department of Agriculture

Addressing the Challenges of Conducting Effective Supplemental Nutrition Assistance Program Education (SNAP-Ed) Evaluations: A Step-by-Step Guide

Nutrition Assistance Program Report **March 2014**

Food and Nutrition Service
Office of Policy Support

United States Department of Agriculture

Food and Nutrition Service, Office of Policy Support *March 2014*

Addressing the Challenges of Conducting Effective Supplemental Nutrition Assistance Program Education (SNAP-Ed) Evaluations: A Step-by-Step Guide

Authors:
Sheryl Cates
Jonathan Blitstein
James Hersey
Katherine Kosa
Laura Flicker
Karen Morgan
Loren Bell

Submitted by:
Altarum Institute
2000 M Street N.W. Suite 400
Washington, DC 20036

Project Director:
Loren Bell

Submitted to:
Office of Policy Support
Food and Nutrition Service
3101 Park Center Drive
Alexandria, VA 22302-1500

Project Officer:
Anita Singh

This guidebook is available on the Food and Nutrition website:
http://www.fns.usda.gov/research-and-analysis

Suggested Citation:
Cates, S., Blitstein, J., Hersey, J., Kosa, K., Flicker, L., Morgan, K., and Bell, L. *Addressing the Challenges of Conducting Effective Supplemental Nutrition Assistance Program Education (SNAP-Ed) Evaluations: A Step-by-Step Guide*. Prepared by Altarum Institute and RTI International for the U.S. Department of Agriculture, Food and Nutrition Service, March 2014.

Acknowledgements

This guidebook was prepared by the *SNAP Education and Evaluation Study: Wave II* project team at Altarum Institute and RTI International for the U.S. Department of Agriculture, Food and Nutrition Service (FNS), Office of Policy Support.

Many individuals made important contributions to this guidebook. From RTI International, the authors thank Georgina McAvinchey, Rosanna Quiroz, and Marjorie Hinsdale-Shouse for information they provided on readability and translation; Cathy Tokarski and Taya McMillian for their writing and editing assistance, and Sharon Barrell for preparing the document.

The authors would also like to thank the following individuals for their assistance in the study and development of this guidebook.

- U.S. Department of Agriculture, Food and Nutrition Service Project officer, Anita Singh, who provided valuable advice and guidance throughout the study and in the planning and editing of this report.
- U.S. Department of Agriculture, Food and Nutrition Service staff for their review of the draft reports: Danielle Berman from the Office of Policy Support; as well as Jane Duffield and Usha Kalro from SNAP.

Table of Contents

List of Exhibits

List of Appendixes

Chapter I • Introduction

A. Background

Nutrition education is an optional component of USDA's Supplemental Nutrition Assistance Program (SNAP), known as SNAP-Education or SNAP-Ed. The goal of SNAP-Ed is to improve the likelihood that SNAP participants and persons eligible for SNAP or other means-tested programs will make healthy food choices within a limited budget and choose physically active lifestyles in accordance with the Dietary Guidelines for Americans. SNAP-Ed Guidance also encourages all States to evaluate the effectiveness of their SNAP-Ed interventions. This can include formative, process, outcome, and impact evaluations.[1]

The *SNAP Education and Evaluation, Waves I and II*[2] studies were USDA Food and Nutrition Service (FNS)-initiated independent evaluations designed to identify potential models of effective SNAP education and evaluation. The overarching goal of these evaluations was to determine whether the selected projects can serve as good examples of effective nutrition education and promotion activities within SNAP-Ed by meeting the following criteria:

- positively impacting the nutrition and health behaviors of SNAP participants while adhering to FNS' SNAP-Ed Guiding Principles,
- exhibiting the potential to serve as models of effective nutrition intervention for large segments of the SNAP audience while requiring levels of resources that are manageable by a large percentage of SNAP-Ed implementing agencies (IAs), and
- providing methodologically robust yet logistically practical examples of project-level SNAP-Ed evaluation.

To accomplish the study goal, three complementary assessments were conducted for each demonstration project: a process evaluation, an impact evaluation, and an assessment of the IA's own impact evaluation. The process evaluations determined whether the interventions were implemented as intended. The data collected from the process evaluations also identified potential challenges and barriers to implementation. The impact evaluations assessed the impact of the programs on participants' average daily consumption of fruit and vegetables and other nutrition behavior outcomes. The assessments of the IA's self-evaluations considered the rigor of the self-evaluations and identified strengths, limitations, and areas for improvement. For the IAs that conducted impact evaluations,[3] the evaluations were technically sound and demonstrated most of the characteristics of a rigorous evaluation; however, the assessment of the self-evaluations identified a number of challenges that IAs face in conducting impact evaluations.

SNAP-Ed Guiding Principles call for SNAP-Ed programs that are evidence based and behaviorally focused and that States "demonstrate through research review or sound, self-initiated evaluation, if needed, that interventions have been tested and demonstrated to be meaningful for their specific target audience(s), implemented as intended or modified with justification, and shown to have the intended

[1] Prior to 2007, the SNAP-Ed Guidance encouraged States to evaluate the effectiveness of their nutrition education programming and provided links to evaluation resources and tools. In 2007, USDA expanded the guidance to encourage the use of a control or comparison group so that the impact of the program could be assessed and set a specific threshold for funding approval for impact evaluations.

[2] The final report for each wave of the study and the individual case study reports for each of the seven demonstration projects are available at http://www.fns.usda.gov/research-and-analysis.

[3] Two of the IAs in Wave I did not conduct impact evaluations; their self-evaluations were outcome evaluations; thus, they were unable to establish causality between the intervention and the nutrition behavior outcomes.

impact on behavior" (USDA, 2013). Based on the findings from the *SNAP Education and Evaluation, Waves I and II* studies, measuring and identifying the results of nutrition education in terms of measurable changes to dietary behaviors continue to be challenging.

The document *Nutrition Education: Principles of Sound Impact Evaluation* (USDA, 2005) provides SNAP-Ed IAs with guidance for conducting a sound impact evaluation (see sidebar). Given the range of available evaluation methodologies, the challenge to the evaluator is to choose a design that eliminates alternative explanations of program effects and establishes causality between the intervention and the dietary behavioral outcomes within the resource constraints of the IA. Based on the assessment of the Wave I and Wave II IAs' self-evaluations, as well as considering the types of resources and staff typically available to SNAP-Ed IAs, the final reports for the two studies offered recommendations for improving the impact evaluations conducted by SNAP-Ed IAs (Gabor et al., 2012a; Long et al., 2013).

Principles of Impact Evaluation

1. Make certain that the nutrition education intervention can be evaluated.
2. Build on available research.
3. Hold out for research designs with random assignment but use them selectively.
4. Choose impact measures that fit the intervention and that approach existing standards for credible assessment.
5. Observe standards for the fair treatment of study participants.
6. Collect impact data after startup problems get resolved but before implementation rolls out.
7. Report both positive and negative results, but do so accurately.
8. Share results to maximize their value.

Source: U.S. Department of Agriculture, Food and Nutrition Service. (2005). *Nutrition education: Principles of sound impact evaluation.* Retrieved from http://www.fns.usda.gov/nutrition-education-principles-sound-impact-evaluation

B. Purpose of this Guidebook

Based on the recommendations identified in the SNAP-Ed Wave II study, this guidebook provides IAs with step-by-step guidance to address the challenges they often face so that they can conduct rigorous and useful evaluations. Recognizing that the level of expertise and experience with evaluation will vary for each IA, this guidebook provides information on evaluation in nontechnical terms. We encourage you to reach out to those with expertise in a particular area, for example, instrument design, research design, or statistical analysis, if you or a team member lack the required capabilities and experience.

This guidebook first addresses the challenges of successfully implementing the intervention and its effect on evaluation. Next, the guidebook addresses the challenges of conducting effective impact evaluations. The challenges are organized into four key areas: study design and measures, instrumentation, data collection, and data analysis. Throughout the guidebook are sources you may want to consult for additional information as you plan and conduct evaluation studies. The guidebook also provides a glossary of key terms and several appendixes with examples and additional information on instrumentation development and testing, attrition analysis, and training of data collectors.

Chapter II • Addressing the Challenges of Implementation and Its Effect on Evaluation

Implementation challenges can affect program evaluation in several ways. Moreover, these challenges often don't appear until the evaluation has begun and it is too late to fix the problem. This chapter provides guidance on how to address the most common implementation challenges faced by implementing agencies (IAs).

A. Planning for Evaluation and Developing the Evaluation Plan

Why Evaluate?

Evaluations can serve many purposes. For example, they can help you adapt programs to best meet your participants' needs, guide efforts to enhance participant engagement, and monitor the progress of your participants in mastering key aspects of nutrition. Evaluations can also assess the effects of your program on nutrition-related behavior. This can provide information to ensure the accountability of your program and offer lessons that can be used to improve the program and to create evidence-based interventions.

Although evaluation is a beneficial activity, to truly be effective, you must talk to your stakeholders to better understand how to best communicate the issues that your program faces and the questions that the evaluation can help address. Talking with stakeholders helps focus the evaluation and engages stakeholders in how to use evaluation findings to increase their impact. Evaluation is a process as much as an individual study, and the information gained in one evaluation study can guide the ongoing improvement of your nutrition education efforts.

Nutrition Education: Principles of Sound Impact Evaluation (USDA, FNS, 2005), a nutrition education resource from FNS, describes four primary types of evaluation: formative research, process or implementation studies, outcome assessment, and impact evaluation. These are defined in Exhibit II-1 along with examples of the type of evaluation questions suited to each.

Each phase of your program will have a corresponding type of evaluation, and each type of evaluation will produce key information to better guide and inform your nutrition education efforts.

- **Formative research** is best suited for early intervention development periods when your program managers and IAs are interested in learning more about how to address target populations.
- **Process/implementation studies** can provide important feedback about how your program meets or fails to meet expectations.
- **Outcome assessments** offer evidence that your program is reaching its target audience and achieving its stated goals.
- **Impact evaluations** validate statements about the relationship between your program and nutrition and other outcomes.

Exhibit II-1.— Types of Evaluations, Definitions, and Example Questions

Type of Evaluation	Definition	Example Questions
Formative research	Application of qualitative and quantitative methods to gather data useful for the development and implementation of intervention programs.	• Do elementary-age school children served by our Implementing Agency (IA) eat the recommended daily servings of fruit and vegetables?
Process (implementation) study	The measurement and tracking of activities associated with the implementation and fidelity of an intervention program.	• How many SNAP participants and low-income eligibles are enrolled in the intervention? How many attended each of the six classes offered?
Outcome assessment	Examination of the extent to which an intervention program achieves its stated goals.	• Did the *Healthy Kid* program meet stated goals of increasing use of fat-free or 1% milk by 25% among participating families?
Impact evaluation	Measures the net change in outcomes for a particular group of people that can be attributed to a specific program.	• Did children in the *Color Your Plate* program increase the number and types of vegetables eaten by at least 0.25 cups per day compared with children who did not participate in the program?

A word of caution! Impact evaluations are often necessary, and many of your program managers may be eager to demonstrate the benefits of your program. But keep in mind that dietary behavior change is a complex process, and many variables can affect the outcomes you're seeking to evaluate. Moreover, implementing a nutrition education program can be challenging. It's easier than you may think for your evaluation efforts to produce results that suggest your program isn't really effective, when the real culprit is incomplete or less-than-acceptable levels of implementation. So watch out for poor implementation and/or overestimation of anticipated program efforts. Both can produce inconclusive data. If you still have unanswered questions, conduct a formative, process, or outcome assessment as needed, before conducting an impact evaluation.

Steps in Preparing an Evaluation Plan

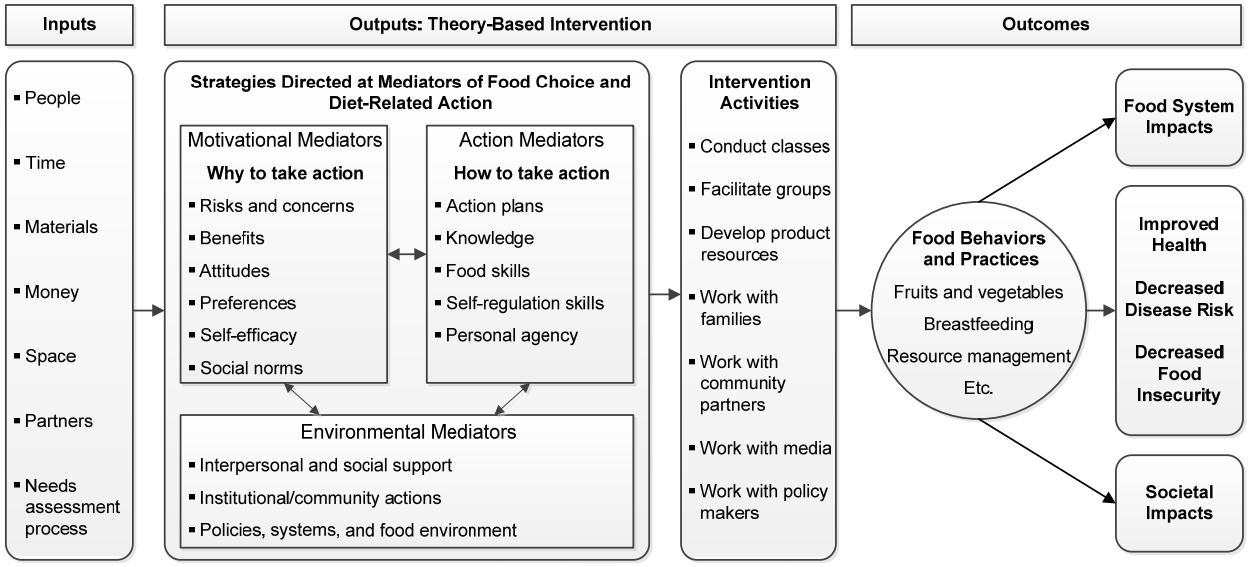
Prepare a Logic Model for the Program. In many cases, the nature of an evaluation may be tailored to the maturity of the program (Chen, 2005). A useful step in tailoring your evaluation is to first develop the logic model of the SNAP-Ed program. This step describes the process by which your program participants are exposed to nutrition education and the various steps expected to take place to affect changes in nutrition behavior.

These steps include:

1. participant awareness of nutrition education messages;
2. changes in knowledge and attitude and motivation; and
3. actions that can enable healthy nutrition behaviors, such as use of nutrition labels while shopping or making fruit and vegetables available for snacks.

Consider developing a graphic representation of your logic model that lays out the expected linkages between these various steps (see Exhibit II-2). Also, be sure to focus on the key pieces of your program's effects. This is a top priority for any evaluation study.

Exhibit II-2.— Logic Model for Nutrition Education Interventions



Source: From *Nutrition education: Linking research, theory, and practice* (2nd ed.), by I. Contento, 2011. Copyright (2011) by Jones & Bartlett Publishers. Reprinted with permission.

Tips to Create Your Evaluation Plan

- Prepare your logic model
- Develop an evaluation planning matrix
- Use multiple sources of information

In most evaluations, it is useful to gather information on the level of exposure (i.e., dosage) to the intervention in addition to behavioral effects, because that can help guide enhanced outreach efforts. For example, if a component of the SNAP-Ed program included parent education, you may want to measure how many minutes of parent education were delivered. For programs with children, it may be helpful to evaluate the effects on the nutrition-related behavior of parents of the students who received nutrition education. Finally, an evaluation may also want to assess the effects of the program on the nutrition behavior of children at home rather than only in the environment (i.e., school) where the intervention was delivered.

Develop an Evaluation Planning Matrix. Next, develop an evaluation planning matrix (see Exhibit II-3) that describes the following in logical sequence (Holden & Zimmerman, 2009):

- key evaluation questions to be addressed
- outcomes that might be assessed
- key data sources and/or data elements that can address these questions

Use Multiple Sources of Information to Corroborate Findings. Having laid out the potential data sources to address different evaluation questions, your evaluation team will work collaboratively with evaluation stakeholders to focus the evaluation (e.g., priority research questions; data sources, resources, and schedule requirements). Again, remember, it is important to engage stakeholders.

- Working closely with evaluation stakeholders (e.g., funders, program staff, and nutrition educators with an interest in evaluation decisions) can help you identify priority questions for your study and the most effective tradeoffs given constraints of time and resources.
- Close engagement of stakeholders in the evaluation design process can help promote the use of evaluation findings for program improvement.

Exhibit II-3.— Example of an Evaluation Planning Matrix

Question	Possible Outcome	Potential Source
Formative Research		
How relevant are the materials to the lives of participants?	• Perceived relevance	• Focus groups
Process Evaluation		
How effective is the nutrition education program at reaching the appropriate target population?	• Program reach • Program completion • Characteristics of participants	• Program records • Participant surveys
To what extent is the nutrition education program being implemented in the ways specified?	• Program fidelity	• Program materials • Interviews with implementers • Program information on mode and intensity of implementation
How much does the program cost?	• Program budgets • Program staffing	• Program budgets • Staff interviews
Outcome Assessment (May benefit from a control or comparison group)		
Does the program motivate participants to engage in healthier nutrition behaviors?	• Changes in participant attitudes and motivations for healthier foods	• Participant surveys
Do program participants master nutrition-related competencies?	• Changes in participant competencies (e.g., nutrition knowledge, self-efficacy) • Changes in consumption (e.g., use of 1% or fat-free milk)	• Participant surveys
Impact Evaluation (typically requires a control or comparison group)		
Is the nutrition education program effective?	• Changes in nutrition-related behavior • Changes in consumption	• Participant surveys • Surveys of parents of child participants
What are the program costs relative to its effectiveness?	• Cost-effectiveness	• Program impacts/unit of program cost

For More Information on Planning an Evaluation

- Chen, H. (2005). *Practical program evaluation*. Thousand Oaks, CA: Sage.
- Holden, D. J., & Zimmerman, M. A. (2009). *A practical guide to program evaluation*. Los Angeles, CA: Sage.
- Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation. A systematic approach.* (6th ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (1994). *Handbook of practical program evaluation*. San Francisco, CA: Jossey Bass.

You can use multiple sources of data to help increase confidence in the robustness of your evaluation findings. For instance, in assessing implementation, it helps to verify findings among interviews with your program director and program staff by reviewing written documents (e.g., logs of participation, exposure data for social marketing programs, instructor work plans) and nutrition education materials. If possible, it is also helpful to observe nutrition education classes and gather feedback about your program from teachers, parents, and/or students.

B. Process Evaluation and Monitoring Program Fidelity

Although the focus of this guidebook is on impact evaluation, careful assessment of program implementation (e.g., process evaluation) is another important element of an overall program evaluation. Exhibit II-4 lists important measures of a process evaluation and suggestions on how to collect data to assess program performance along these measures.

Exhibit II-4.— Process Evaluation: What to Measure and How

Process Measure	Rationale	Potential Sources
Partnerships		
• Partnership agreements are in place to ensure cooperation with implementation plan. • Facilities and support provided by partners are available on a consistent basis. • If facilities are provided, they are accessible to the target audience.	Need to provide consistent support to the Implementing Agency (IA) to assure that disruptions do not occur and that support promised is available when needed.	• Key informant interviews with partners, IA staff, and nutrition educators • Focus groups with clients
Target Audience		
• Formative research was conducted to understand both the needs of the target audience and their receptivity to the proposed approach. • Services are provided at a time and place that are convenient to the target audience. • Engagement of critical sectors of the target audience (e.g., parents of school children) is well planned and documented.	Recruitment and engagement of target audience members in the intervention is critical to successful behavior change. Ensuring that the target audience wants the intervention, is willing to commit the time to participate, and can access the services is critical to success.	• Program records • Formative research reports and plans • Interviews or focus groups with target audience members • Surveys of target audience members • Key informant interviews

(continued)

Exhibit II-4.— Process Evaluation: What to Measure and How (continued)

Process Measure	Rationale	Potential Sources
Implementation Logistics		
• Staff has a clear understanding of the nature of the intervention, how it is being implemented, and their role in all aspects of the program implementation. • Staff has received appropriate training to implement the intervention. • Systems are in place to handle unanticipated problems and provide administrative and supervisorial support to line staff.	Staff needs a clear understanding of implementation logistics and their role in the process to effectively implement the interventions. Support systems must be in place to provide necessary training, supervision, and problem solving.	• Key informant interviews with staff • Review of training plans • Onsite observation of implementation • Staff meeting notes
Budget		
• Budget is adequate to support implementation as planned. • Contingency plans are in place to address budget shortfalls.	Staying on budget and ensuring resources are available are key to successful implementation.	• Review budget and expenditure documents • Key informant interviews

Assessing Program Reach and Dosage

Process evaluation can help document the reach of your program—the proportion of the target population who participated in the SNAP-Ed program. For instance, you might want to know the proportion of eligible preschool children who participated in your program. In the case of programs with children, reach is often influenced by the degree of engagement of parents with the program.

Process evaluation can also help document the dosage or intensity of engagement of participants in your program. Process information on the dosage of nutrition education can often provide valuable feedback. For example, an earlier evaluation of a SNAP-Ed program with low-income preschool children found that the childcare centers faced challenges with parent engagement (Gabor et al., 2012b). To address this challenge, the program added a local "promotion" effort in the preschools during the weeks prior to program implementation to increase parent involvement. This example illustrates how information on dosage can help identify the effects of nutrition education in an impact evaluation.

Reach versus Dosage

Reach helps quantify the proportion of the target population that participated in the program. **Dosage** measures participants' exposure to the program, that is, the amount of the program received.

For example, consider a SNAP-Ed program that offers six 30-minute classroom lessons to preschool children. Reach can be defined as the percentage of eligible children who attended at least one of the lessons, and dosage can be defined as the mean number of lessons attended (e.g., 3.5 classes) or the mean number of minutes spent in the classroom (105 minutes).

Assessing the Fidelity of Program Implementation

The recent proliferation of evidence on the importance of high-quality program execution suggests that there is wisdom in assessing the "**fidelity** of program implementation." Faithful implementation of all program components is necessary to ensure that the program your audience receives matches the program you intended to deliver. Fidelity assessment can take the form of interviews with educators and other key staff as well as random, unannounced observations of classroom practices.

Even veteran educators will need time and experience to fully grasp all the details of new program materials. What's more, over time, program providers may drift from the intended delivery if they feel their experience trumps the original program design. Although well intentioned, that kind of variation can undermine program effectiveness. Fidelity assessment as a form of monitoring can also be used to provide positive feedback and professional development for your staff.

Fidelity refers to how closely the intervention was implemented as designed. For example, it's difficult to draw conclusions about the impact of a program if the nutrition educators did not follow the procedures they received in training.

Although program fidelity is important, it is equally important to understand the unique nature of your audience and setting (e.g., cultural differences) and the ways that they vary from the audience and setting reported in the original research. There may be times when you feel that you need to adapt parts of a program to ensure that the program's messages will resonate with your audience. Sticking too closely to a program that does not meet the needs of your audience is not likely to lead to success. If you do decide to tailor an evidence-based program, be sure to carefully and thoroughly document the changes. Make sure that the changes are understood and implemented by all staff members.

Assessing the Impact of Social Marketing Efforts

Increasingly, nutrition education involves a combination of direct nutrition education with social marketing activities, reinforcing nutrition education through a variety of media or channels (e.g., television, radio, transit posters, billboards, posters and flyers, print sources, point-of-purchase promotions, community and afterschool events, and communications via social media such as Twitter or Web sites promoted by email). Because nutrition-related behavior involves an ongoing series of choices, the cues provided by social marketing efforts can help motivate and reinforce changes in nutrition behavior.

Program planners are often interested in the level of exposure to different types of media channels. Because different media channels require different investments in time and resources, it is important to know that exposure information can help you plan future efforts and understand the effects of your social marketing campaign. The various channels of the social marketing component of a SNAP-Ed program implemented by the Iowa Nutrition Network (INN) are shown in Exhibit II-5.

Exhibit II-5.— Example of Social Marketing Elements Monitored in the Evaluation of the INN BASICS Program

Component	Description
Billboards	Fourteen billboards in Supplemental Nutrition Assistance Program Education (SNAP-Ed)-qualified low-income census tracts displaying campaign messages and imagery
Bus shelters	Signage featuring campaign messages and imagery displayed on seven bus shelters serving passengers on bus lines in SNAP-Ed-qualified low-income census tracts
Television	Ads broadcast on television stations with viewers in the target demographic
Radio	Ads broadcast on radio stations with listeners in the target demographic
Family nights out	One weeknight event at intervention schools to provide families with hands-on, fun nutrition and physical activity education as well as resources to help them develop healthy habits

(continued)

Exhibit II-5.— Example of Social Marketing Elements Monitored in the Evaluation of the INN BASICS Program (continued)

Component	Description
Materials in schools	Signage featuring campaign messages
Materials in the community	Signage featuring campaign messages and imagery posted at locations such as WIC offices and YMCAs
Point-of-purchase intervention	Signage featuring campaign messages and imagery in milk and produce departments at participating retail grocery stores over a period of 7 months; two food demonstrations per month at each store (coordinated with curriculum classroom tastings)

Media like television can be relatively expensive; still, well-executed ads on TV can be memorable. Conversely, less well-executed ads can easily be lost in the clutter of ads for commercial products. Radio ads have the advantage of being more easily targeted to a specific audience and may have advantages in terms of reaching households with a high proportion of SNAP-eligible participants.

The objective of using process evaluation in your social marketing outreach is to generate information on the reach of your social marketing effort. In other words, you want to find out the proportion of a target audience (e.g., low-income families) who saw an ad a least one time and the frequency, or number of times, that audience saw an ad. Early research suggests that at least three exposures are required to influence purchase behaviors; more exposures may be needed to influence nutrition-related behaviors (Hersey et al., 2005).

Typically, media outlets can provide detailed information on the potential reach of their media channel (e.g., the number of individuals who watch television or listen to radio at the time when ads are broadcast, or the number of people who view a billboard or transit poster, or the number of individuals who read the newspaper or magazine). Because this type of information is used to price advertising, it is generally available, and it is often possible to get estimates of exposure among broad age or income groups.

However, just because a nutrition education message is broadcast does not necessarily mean that people pay attention to it. To assess awareness of a campaign, an evaluation of social marketing efforts will survey a target audience to assess awareness of nutrition education messages or messages in different media channels.

Awareness versus aided awareness? In evaluating awareness of broadcast media using telephone or in-person surveys, it's smart to use measures of "confirmed awareness." In this kind of evaluation, you will describe portions of an ad and ask a respondent to describe other things that happened in the ad. The respondent's answers are used to confirm awareness of exposure.

"Aided awareness" measures can be used with Web or paper-based mail surveys. They ask respondents if they have seen an ad with a particular tag line. Those measures often yield higher degrees of positive response than confirmed awareness measures so that it is helpful to include a bogus tag line as a way to gauge the level of possible overreporting. Evaluation of health promotion media campaigns have found that both measures of exposure and measures of awareness contribute to the impact of social marketing efforts on health behavior (Hersey et al., 2005).

Newspapers, newsletters, and brochures can be evaluated by finding out who read the materials through the use of surveys of the target population. A process evaluation may be useful in helping to assess

awareness of different media channels, because that can help planners determine the best mix of media channels to deliver nutrition education messages.

For instance, in the evaluation of INN's BASICS program, a survey of parents reported relatively high recall of materials at food assistance programs (e.g., food pantries or Special Supplemental Nutrition Program for Women, Infants and Children (WIC) clinics). INN also worked closely with a supermarket chain to promote the availability of fruit, vegetables, and low-fat milk. One advantage of the community and social marketing efforts is that they reach a broader number of low-income families in the community, rather than just the child's immediate family. Thus, these efforts may help reinforce broader change toward healthier nutrition behaviors.

Remember, process evaluations can help identify the types of outreach activities that have the greatest impact per dollar of funds expended. Well-targeted, clear messages, even if delivered through a relatively low-cost medium, can have considerable reach.

For More Information on Program Evaluation

- Centers for Disease Control and Prevention. (n.d.). *Social marketing—nutrition and physical activity*. Retrieved from http://www.cdc.gov/nccdphp/dnpa/socialmarketing/training/
- Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Adolescent and School Health. (2005–2011). *Program evaluation resources: Evaluation briefs*. Atlanta, GA: CDC. Retrieved from http://www.cdc.gov/healthyyouth/evaluation/
- Centers for Disease Control and Prevention. (2011). *Introduction to program evaluation for public health programs: A self-study guide.* Retrieved from http://www.cdc.gov/eval/guide/index.htm
- Contento, I. R. (2011). *Nutrition education: Linking research, theory, and practice* (2nd ed.). Sudbury, MA: Jones & Bartlett Publishers.
- Leeman, J., Sommers, J., Vu, M., Jernigan, J., Payne, G., Thompson, D., Heiser, C., Farris, R., & Ammerman, A. (2012). An evaluation framework for obesity prevention policy interventions. *Preventing Chronic Disease, 9*, 110332. Retrieved from http://www.cdc.gov/pcd/issues/2012/11_0322.htm
- National Collaborative on Childhood Obesity Research (NCCOR). (2009). *Policy evaluation webinar series*. Retrieved from http://nccor.org/resources/nccor/webinars.php#fourpart
- U.S. Department of Agriculture, Food and Nutrition Service. (2005). *Nutrition education: Principles of sound impact evaluation.* Alexandria, VA: USDA. Retrieved from http://www.fns.usda.gov/nutrition-education-principles-sound-impact-evaluation
- U.S. Department of Agriculture, Food and Nutrition Service. (2013). *SNAP-Ed strategies and interventions: An obesity prevention toolkit for states—Evidence-based policy and environmental change in child care, school, communities and family settings.* Alexandria, VA: USDA. Retrieved from http://snap.nal.usda.gov/snap/SNAP-EdInterventionsToolkit.pdf
- U.S. Department of Agriculture, Food and Nutrition Service, Supplemental Nutrition Assistance Program, Western Region. (2013, June). *Nutrition, physical activity, and obesity prevention outcomes evaluation framework.* Retrieved from http://snap.nal.usda.gov/snap/WesternRegionEvaluationFramework.pdf
- U.S. Government Accountability Office. (2012). *Designing evaluations: 2012 revision*. Washington, DC: GAO. Retrieved from http://gao.gov/assets/590/588146.pdf

Chapter III • Addressing the Challenges of Conducting Effective Impact Evaluations: Study Design and Outcome Measures

A. How to Select and Define Outcome Measures

Once you've decided to conduct an impact evaluation, you need to decide what you will measure to gauge the success of your program. The outcome or outcomes you select should be related to the content of your program and sensitive enough to change following exposure to your program. In the sections that follow, we use the words "outcome" and "outcome measure" to denote the intended consequence of the intervention. In many cases, outcomes will include changes in dietary intake (e.g., consumption of fruit and vegetables) measured using food frequency questionnaires or 24-hour dietary recalls. Only when these outcomes are evaluated against information obtained from a control or comparison group are they a measure of program impact.

Identify Key Outcomes First

Bring your program stakeholders together to find out what they want to learn about your program and figure out the questions that you want to address. Program stakeholders are individuals or groups of people with significant interest in how well your program performs. For example, you may want to know what type of decisions your stakeholders will make using the results of the evaluation. Who is the audience for the results of the evaluation, and what does that audience need to know?

An **outcome measure** is the desired benefit, improvement, or achievement of a specific program goal or objective. More specifically, it is the amount of change expected among an individual, a group of people, or population that is associated with a program or intervention activity within a given timeframe (Posavac & Carey, 1997).

Next, develop a logic model or a flow diagram that describes your program's components and desired outcomes, as described in Chapter II. At a minimum, your logic model should outline the program's inputs, activities or processes, outputs, and outcomes (see Exhibit II-1). Keep in mind that outputs and outcomes are different:

- Outputs describe the number of individuals served by a program.
- Outcomes record the results of the program's interaction with the individual.

Also include any external factors that may influence the program's inputs, activities, or outcomes.

Use SMART Outcome Measures

Once you have identified your key outcomes, it's time to create your specific outcome measures. Outcome measures quantify your desired results or achievements and should be clearly defined and

Outcome measures quantify the desired achievements and should be clearly defined and "SMART" (Meyer, 2003)

- **S**pecific
- **M**easurable
- **A**chievable
- **R**elevant
- **T**ime-bound

"SMART" (Meyer, 2003). SMART outcome measures are as follows:

- **S**pecific: Clearly state your desired outcome, achievement, or accomplishment by addressing the five W's: who, what, when, where, and why.

> Good example:
> - ✓ ***Specific outcome:*** At least 85% of the teens enrolled in the program will increase their daily consumption of whole grains by at least 0.3 ounce equivalent of whole grains per 1,000 calories by May 2015.
>
> Bad example:
> - x ***Nonspecific outcome:*** To get teens to eat more whole grains.

- **M**easurable: Next, identify the level or amount of change that you expect as a result of your program's activity. Be sure to include numeric or descriptive measures that define quantity, quality, cost, and so on. Remember that measurable outcomes guide evaluation design, track progress, and document success.

> Good example:
> - ✓ ***Measurable outcome*:** To increase fruit and vegetable consumption among third graders in the Wake County school district by 25% by June 2015.
>
> Bad example:
> - x ***Nonmeasurable objective:*** To ensure that the third graders in the Wake County school district eat more fruit and vegetables.

- **A**chievable: The behaviors that you desire to see should be realistic and attainable with your current resources and timeframe. You can stretch a bit, but they should still be feasible.

> Good example:
> - ✓ ***Achievable objective:*** To increase the number of cups of vegetables consumed by adults aged 60 to 75 enrolled in the program by 0.3 cups per day by December 2014.
>
> Bad example:
> - x ***Nonachievable objective:*** To increase older adults' vegetable consumption by 2 cups in 1 month.

- **R**elevant: Your goals should be critical to the mission of your organization and align with other goals. Keep in mind that synergistic goals get support (e.g., resources, a champion) and drive the team, department, and organization forward.
- **T**ime-bound: Clearly state a definitive date when the desired behavior will be achieved.

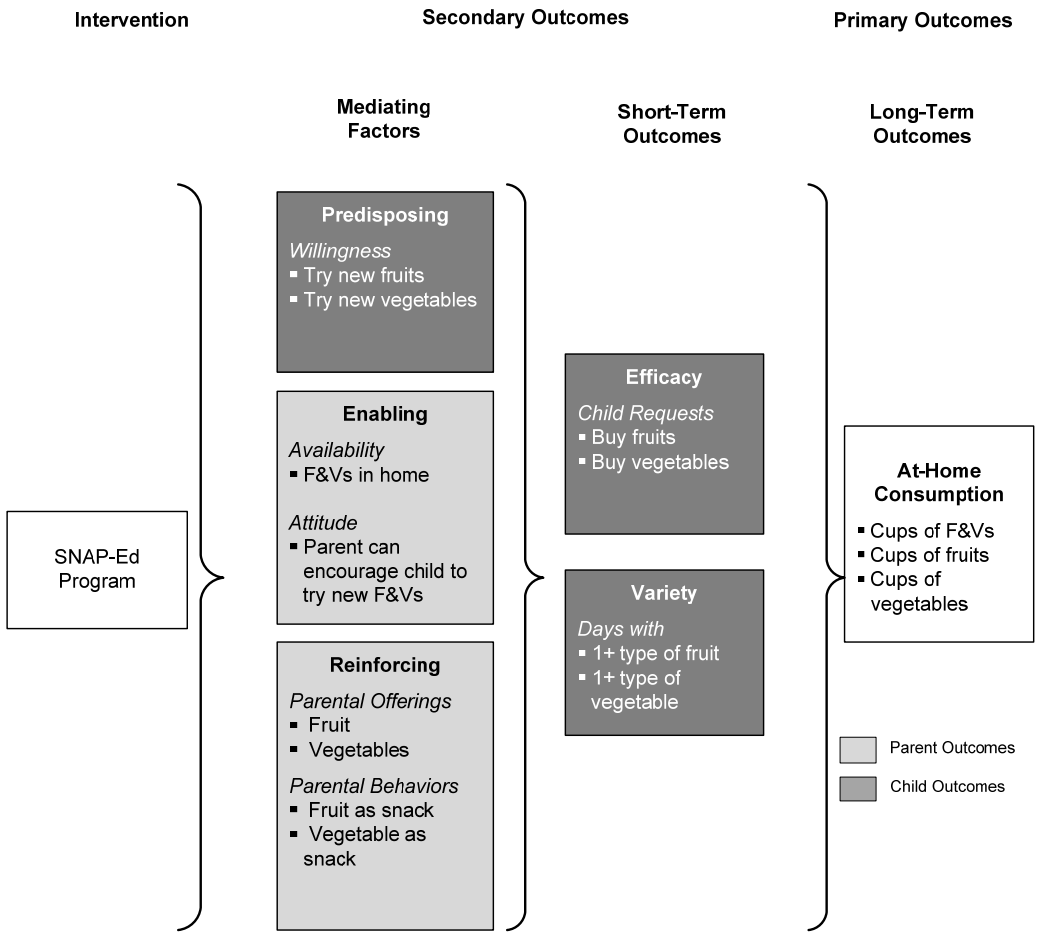
Once you've written your desired outcomes, take a moment to identify when (e.g., at baseline and 2 weeks after the end of the intervention) and how each of the outcomes will be measured.

Include Primary and Secondary Outcome Measures

You should consider including secondary outcome measures in addition to your primary outcomes. When done correctly, secondary outcomes are important because they link your intervention to your long-term outcome, creating a clear snapshot of your program's impact on participant behaviors.

The framework presented in Exhibit III-1 was used by the independent evaluator for the SNAP-Ed Waves I and II evaluations and was adapted from Green et al. (1980). It has been applied in other studies to capture the main types of secondary outcomes associated with changes in nutrition behavior (Mullen, Hersey, & Iverson, 1987). Secondary outcomes capture the complex nature of the behavior change process, include mediating factors and short-term outcomes.

Exhibit III-1.— Conceptual Framework for the SNAP-Ed Wave II Independent Evaluations



Note: F&Vs = fruits and vegetables

Example: Suppose the objective of your program is to increase children's combined at-home fruit and vegetable consumption by 0.3 cups per day. The program curriculum includes several components with a focus on encouraging children to eat a variety of fruit and vegetables and to try new fruit and vegetables by offering tastings and cooking demonstrations at school.

Additionally, your program provides parent newsletters with information on how to shop for nutritious foods and ways to offer fruit and vegetables as alternatives to sweet or salty snacks.

As shown in Exhibit III-1, three main types of mediating factors can influence changes in dietary consumption.

- **Predisposing factors** include the knowledge and attitudes of an individual as they relate to the motivation to act. In our example, a predisposing factor is the willingness of a child to try new fruit and vegetables.
- **Enabling factors** include the skills and resources needed to engage in good nutrition. In our example, an enabling factor is the availability of fruit and vegetables in a child's home.
- **Reinforcing factors** include factors that help reinforce healthy nutrition. In our example, a reinforcing factor is a parent or caregiver offering fruit and vegetables as options for snacks or at dinner.

These mediating factors can affect diet-related behaviors that include the following short-term outcomes: the child asking the parent or caregiver to buy certain fruit or vegetables and the daily variety of fruit and vegetables eaten by the child. According to the framework, a greater willingness to try new fruit and vegetables may influence how often a child eats a variety of fruits and vegetables. Changes in these short-term outcomes may influence the primary outcome: at-home consumption of fruit and vegetables.

Express Evaluation Objectives in Quantifiable Terms

Evaluation objectives should always be expressed in quantifiable terms. If not, it is hard to know if your program failed to observe changes in dietary behavior because of implementation failures or because of statistical and measurement issues.

It may be helpful to examine measures from prior program implementations; for example, this may help you decide how much change in fruit and vegetable preference is realistic and achievable. Alternatively, you can review published literature to get a better sense of the magnitude of change in programs similar to your intervention that have been evaluated by others. Regardless of whether your prior estimates come from your work or the work of others, you can use these values and your best judgment to quantify expected program impacts. With these prior estimates, you can make best- and worst-case estimations to help in other facets of program planning. If no prior estimates are available, you might step back and consider if it's the right time to move forward with a full impact evaluation or if it might make more sense to conduct a smaller pilot study first.

Use Systematic Reviews to Inform the Development of Outcome Measures

Systematic reviews, such as the meta-analyses published by Knai, Pomerleau, Lock, & Mckee (2006) and Thomson & Ravia (2011), can be very useful. These review papers provide a range of values for studies similar to many SNAP-Ed programs.

The review by Snyder and colleagues (Snyder, Hamilton, Mitchell, & Kiwanyka-Tondo, 2004) provides insight into the potential effects of social marketing programs.

For More Information on Selecting and Defining Outcome Measures

- Davidson, E. J. (2013). *Actionable evaluation basics: Getting succinct answers to the most important questions*. Auckland, New Zealand: Real Evaluation Ltd.
- Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (2010). *Handbook of practical program evaluation* (3rd ed.). San Francisco, CA: Jossey-Bass.
- Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

B. How to Select a Research Design

There is rarely, if ever, a single answer to the question "What is the right evaluation design?" Good evaluation designs are derived from an open and dynamic process that involves working with stakeholder groups, program managers, and funding agencies. Thus, your design choice should be based on:

- knowledge of the program, its objectives, and its target population;
- a clear understanding of the needs and ultimate uses of the information that will be derived from your evaluation; and
- an agreed-upon scope that summarizes your available resources.

Exhibit III-2 illustrates the steps in choosing a research design. The process begins with a clearly defined research question that addresses the information needs of your key stakeholders. Energy and resources are often wasted by failing to spend enough time considering these issues. As the program evaluator, it is your responsibility to:

- present proposed research questions to your program implementers and other potential users of the evaluation results,
- seek their input, and
- revise the research questions to ensure that the reported findings are viewed as valuable and credible.

There are many factors to consider during the process of developing your research questions. Here are a few key things to consider:

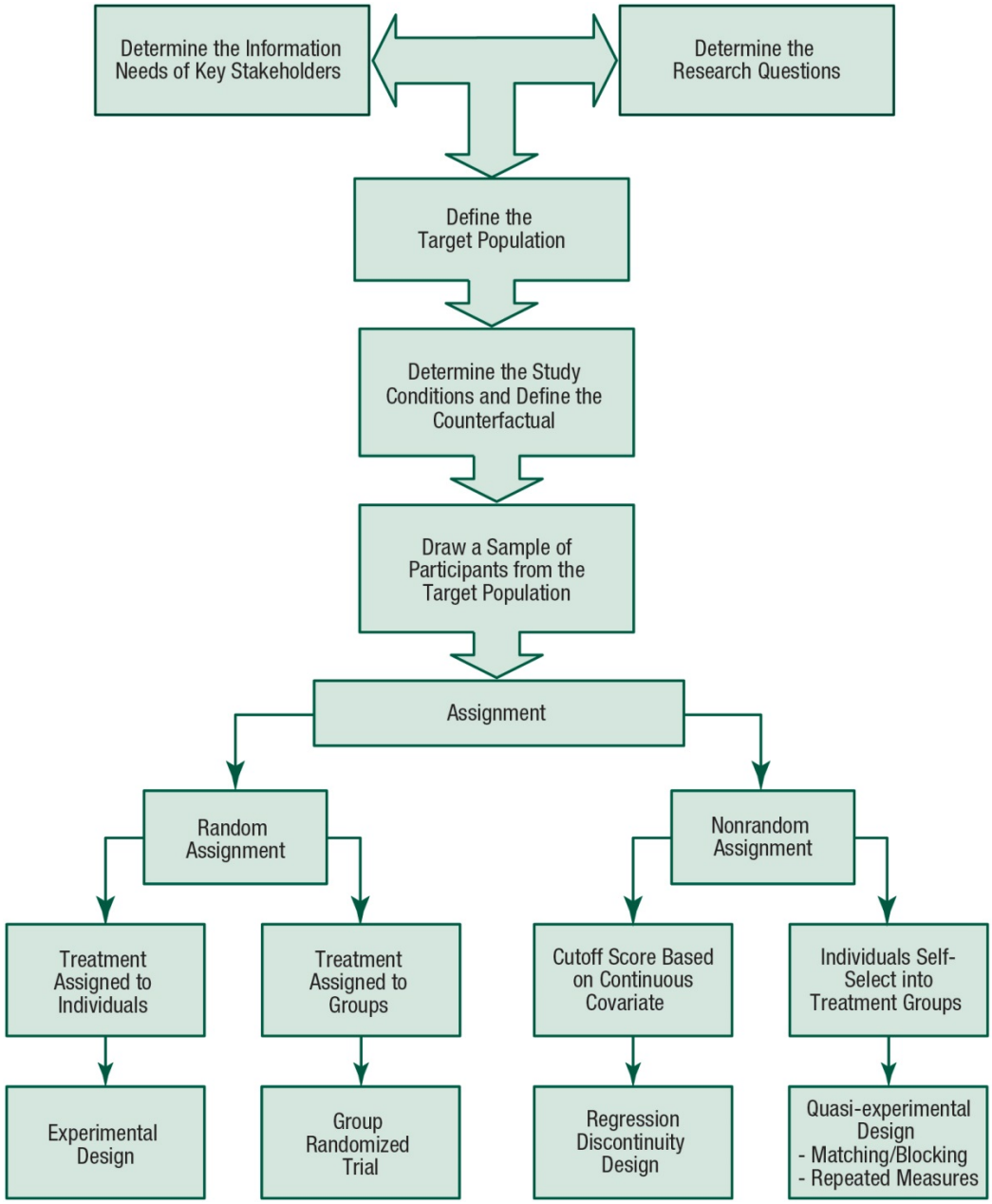
- What is the underlying theory of behavior change?
- Should important intermediate changes like attitudes toward vegetables or grocery shopping habits be evaluated?

The selection of a target population and the sampling strategy may also influence the research design. For example:

- Does the program target particular age groups?
- Are there linguistic or cognitive complexities that might make the information conveyed by the program less accessible to specific groups?

The following sections further explore these questions and include additional issues that must be addressed when selecting your research design.

Exhibit III-2.— Choosing a Research Design



Factors to Consider in Selecting a Research Design

Include a control or comparison group. The hallmark of impact evaluation design is a control or comparison group that allows you to concurrently measure what happened after exposure to the SNAP-Ed program with what would have occurred in the absence of exposure to the program. Your control or comparison group sets up the null hypothesis, or "no effect" outcome (i.e., the **counterfactual**), and allows you to estimate the net impact of exposure (U.S. Government Accountablity Office, 2012). The

impact evaluation is designed to demonstrate that differences between the outcomes in the intervention group and the control (or comparison group) could not have occurred simply by chance by comparing the "null effect" to the program effect.

Use a design with good internal validity. The strength of your research design is reflected in its internal validity. Internal validity refers to the strength of the conclusion that can be drawn about the relationship between your program and its effects on the outcomes of interest. Good designs rule out alternative explanations that might challenge the validity of your causal claim.

The term **counterfactual** refers to the control or comparison condition in an evaluation design. It is the state of affairs (e.g., the measured outcome) in the absence of the intervention.

As shown in Exhibits III-3 and III-4, a number of possible design options, some better than others, provide good internal validity for impact evaluation. Because this guidebook is focused on impact evaluation, the discussion focuses on study designs that include a control or comparison group.

Incorporate randomization when possible. When feasible and practical, a randomized design is typically preferred. Exhibit III-3 includes definitions and issues to consider when selecting a randomized design.

Randomized experimental designs provide strong evidence supporting cause-and-effect statements through the controlled application of an intervention or treatment. Randomization also

Randomized designs provide strong evidence to support cause-and-effect statements.

- increases confidence that the treated and untreated groups are as similar as possible (both on measurable and unmeasurable factors that might confound or distort the results of your study) and
- reduces alternative explanations for program impacts.

Quasi-experimental designs can include the controlled application of an intervention but rely on strategies other than randomization to rule out alternative explanations of program effects. Exhibit III-4 includes definitions and issues to consider when selecting a quasi-experimental design.

Quasi-experimental designs are a reasonable alternative when randomization is not possible. They provide the ability to rule out many potential validity threats such as maturation and history. However, because these designs do not use random assignment, selection bias cannot be completely ruled out.

Consider a Quasi-Experimental Design

If randomization isn't an option, consider a quasi-experimental design which will allow you to rule out potential validity threats.

Exhibit III-3.— Randomized Research Designs, Definitions, and Factors to Consider

Definition	Factors to Consider
Randomized Experimental Design	
Randomized experimental designs provide strong evidence supporting cause-and-effect statements. These designs use controlled application of an intervention or treatment and randomization to provide evidence of the missing counterfactual and support claims of causal inference. They offer strong internal validity, but they may not provide external validity (i.e., generalization) when participants do not match the broader Supplemental Nutrition Assistance Program (SNAP) audience and when the evaluation setting is idealized or highly controlled (Shadish, Cook et al., 2002).	• Requires the ability to assign treatment to some individuals but not to other individuals by a random process (e.g., flip of a coin). This ensures that the control group is as similar as possible to the intervention group in every way except for exposure to the intervention, which helps to reduce the plausibility of selection bias. • Probability of assignment must be a known characteristic of the evaluation design. – For a simple randomized design, individuals are assigned to one of two conditions with equal probability. – Assignment probabilities can be adjusted as long as each individual faces the same probabilities of assignment. For example, 3-to-5 assignment can be applied with only limited impact on statistical power if the program is expensive to implement and only a limited number of participants can be assigned the program.
Group randomized trial (GRT)	
With a GRT, the program is randomly assigned to the school or community center with observations of key outcomes occurring at the individual level.	• Useful option to consider when it is not possible to randomly assign treatment to individual participants. This is common when delivering Supplemental Nutrition Assistance Program Education (SNAP-Ed) in a school-based setting or in community centers. In these contexts, it is not possible to randomly assign individuals or even classrooms to treatment and control conditions because participants mingle and interact, leading to strong concerns that the intervention may "spill over" into the control condition, reducing the ability to assess program impact. • Presents more analytic challenges because the unit of assignment (e.g., school, center) is not the same as the unit of observation (e.g., individual). • Requires the need to account for variations in settings and in program implementation and fidelity, which may influence measures of program effectiveness.

Exhibit III-4.— Quasi-Experimental Research Designs, Definitions, and Factors to Consider

Definition	Factors to Consider
Regression discontinuity	
Assignment to the intervention group is based on a cutoff score on a measured covariate. For example, families with a household income at or lower than a given cutoff score could be assigned to receive the intervention, while families with household incomes above the cutoff score would be in the comparison group. The effect of the intervention is measured as the discontinuity between intervention and comparison regression lines at the cutoff.	• Considered to be one of the strongest quasi-experimental approaches because, like fully randomized designs, it relies on a known mechanism of assignment, in this case, a continuous measure closely related to the outcome of interest. • Provides very good internal validity and allows researchers to rule out many plausible alternative explanations of program effects.
Quasi-experimental designs with repeated measures	
Design with repeated measurement of outcomes with one or more intact groups, with an experimental treatment inserted between at least two of the measurements of at least one group.	• Use of repeated measures is recommended to ensure that both the intervention and comparison groups are as similar as possible at baseline and to allow for difference-in-difference estimation (i.e., net effects) of program impacts. • If resources allow, use multiple baseline measures to increase confidence that the exposed and nonexposed groups were both similar at baseline and demonstrate comparable patterns over time before the introduction of the SNAP-Ed program.
Quasi-experimental designs with matching or blocking strategies	
The term "matching" refers to a one-to-one correspondence between the treatment and comparison groups, while the term "blocking" is used to refer to a many-to-many correspondence. This approach requires the identification of covariates that are highly correlated with the impact measure and the use of those covariates to create two comparison groups that are similar with respect to the measured covariates.	• Reduces the potential influence of selection. • Matching the intervention group and the comparison group on observed covariates reduces the likelihood that these variables will be associated with bias in the estimated impact.

Research Designs for Evaluation of Social Marketing Campaigns

Evaluation designs for social marketing campaigns are worth special attention. If the social marketing component of your SNAP-Ed program involves mass media, in-store placement, or the distribution of materials that are difficult to control, it will be challenging—if not impossible—to randomly assign individuals to treatment and control (or comparison conditions) because of the risks of intervention "cross-over" (i.e., contamination).

Instead, it will be helpful to

- define the geographic areas (e.g., communities) where your social marketing activities occur as part of the treatment condition and
- define the "nonexposed" geographic areas (or areas where your social marketing activities did not occur) as the control condition.

This approach represents an example of the group randomized trial design in Exhibit III-3. When possible, you should identify and recruit a number of communities and randomize them into treatment and control conditions. If you are successful, your design will be fully experimental. However, in many cases, your social marketing activity will occur in certain communities for practical or logistical reasons so that random assignment is not possible. In this situation, you should select communities that are as similar as possible to your treatment communities to provide a reasonable quasi-experimental alternative. In either case, a random sample of individuals within the community should be selected to represent the community as closely as possible.

Address-based sampling or telephone directory lists are highly recommended when selecting a representative sample. However, these approaches will require additional financial and technical resources. If cost is an issue, you may want to consider using intercept surveillance, which refers to placing trained data collectors in key locations (e.g., supermarkets) with high levels of foot traffic by members of the target audience. Data collectors are provided with a protocol or instruction set that describes the process for selecting interviewees. For example, a protocol may indicate that data collectors approach the *nth* individual who appears to be between the ages of 18 and 64 and who exits through the southwest doorway. The protocol should also include instructions for what to say to the potential respondent to increase the likelihood of participation and to reduce the possibility of introducing bias into the data collection process.

For More Information on Research Designs for Program Evaluation

- Bamberger, M., J. Rugh, et al. (2004). Shoestring evaluation: Designing impact evaluations under budget, time, and data constraints. *American Journal of Evaluation, 25*(1), 5–37.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. New York, NY: McGraw-Hill, Inc.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Co.
- Trochim, W. (1990). Regression-discontinuity design in health evaluation. *Research Methodology: Strengthening causal interpretations of nonexperimental data*. L. Sechrest, E. Perrin, & J. Bunker. Washington, DC: U.S. Department of Health and Human Services.
- West, S. G., Duan, N., et al. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health, 98*(8), 1359–1366.

C. How to Conduct a Power Analysis (and Determine Statistical Power)

As you begin to plan your impact evaluation, you may ask yourself, "How many people do I need to survey?" or "How big of an effect will my program have in the target population?" These questions are closely tied to the issue of statistical power.

Statistical power refers to the ability of a test to detect relationships that exist in the population. It is defined as the probability that a statistical test will reject the null hypothesis (no effect) when it is false (Shadish, Cook et al., 2002). In other words, statistical power is an indicator of the likelihood that you will correctly reject the premise that your program did not have an impact on the outcomes based on statistical evidence.

Low statistical power is a major threat to your ability to accurately conclude that your statistical test results are not due to chance; this is commonly referred to as statistical conclusion validity (Shadish, Cook et al. 2002). If your statistical power is low, you run the risk of failing to reject the null hypothesis (no effect), concluding that your program did not have impact. However, if your study had been properly powered (e.g., if you had a larger sample), you would have demonstrated statistically significant results.

You should conduct a power analysis during your study planning phase. This will increase the confidence that your sample size, the anticipated program impact, and other study characteristics are aligned to provide enough statistical power to reject the null hypothesis when real change is observed.

In practice, when evaluators approach a power analysis, they are most often carrying out a sample size calculation because this design feature is within your control and is one of the most influential determinants of statistical power.

Statistical power is the probability that you will reject the null hypothesis when it is false.

If you find that you need to conduct a sample size calculation, you can follow a few simple steps to gather information from the literature, program developers, and even other evaluators who have conducted similar studies. By leveraging information from previous studies that are comparable to the study you are planning, you will increase your likelihood of avoiding the mistake of failing to reject the null hypothesis (i.e., no program effect), when there was an actual program effect. Exhibit III-5 identifies the type of information you need, provides a definition, and indicates where you can find this information.

Exhibit III-5.— Information Needed to Estimate Sample Size

Type of Information	Definition	Source
Program impact (Δ)	The difference between the intervention group mean ($\bar{y}_{tx}$) and the control group ($\bar{y}_{cx}$) mean for the measured outcome of interest ($\bar{y}_{tx} - \bar{y}_{cx}$). We are most often interested in the smallest practical difference, because larger differences are easier to observe. Accordingly, the program impact of interest is often referred to as the **minimum detected effect**.	• Reports in the published literature of studies using the same or similar program. • The program developers or implementers.
Variance of the outcome measure (σ_y^2)	A measure of the variability around the average score for the measured outcome of interest estimated as the sum of the squared deviations around the measured outcome.	Reports in the published literature of studies using the same or similar program.
Type I error rate (α)	The probability of rejecting the null hypothesis when it is true.[a]	Statistical textbooks. The numeric value needed is determined by the Type I error rate and the degrees of freedom for the test statistic.
Type II error rate (β)	The probability of retaining the null hypothesis when it is in fact false.[b]	Statistical textbooks. The numeric value needed is determined by the Type II error rate and the degrees of freedom for the test statistic.

[a] A Type I error rate of α = 5% is desired, but higher Type I error rates are acceptable depending on the risk associated with claiming that statistical differences are significant when they are not.

[b] The Type II error rate determines statistical power and vice versa. Power (1-β) is the probability of rejecting the null hypothesis when it is false. By convention, researchers strive for $\beta \geq 80\%$. Lower levels of power lead to Type II errors that are at an unacceptable rate.

Example: You have been asked to conduct a study assessing the impact of a program designed to increase fruit and vegetable consumption among children aged 9 to 11 years. The study will include one group of children exposed to the *Healthy Eats Program* and one group of children who are similar in many respects but have not been exposed to the program. The survey includes a measure of cups of fruit and a measure of cups of vegetables that will be used to create an index of cups of fruit and vegetables consumed. The index is continuous and should provide a normal distribution of scores. You selected a randomized design and need to determine how many children to include and survey in each group.

Step 1. Determine the minimum program impact. (Δ) Based on your review of the literature, you find that previous studies have shown that the *Healthy Eats Program* has produced statistically significant increases of 1/3 of a cup of fruit and vegetables per day in preteen children. After discussions with the program implementers (*always* talk to the program implementers to really find out what's happening), you believe that the current study is likely to replicate this finding.

Step 2. Determine the variability of the outcome measure. Variability is used to describe the spread among a group of scores around the mean. It's also a measure of "noise" or imprecision in the data you've collected. This information will sometimes be provided as

- the standard deviation of the outcome $\left(\sigma_y\right)$ or
- the standard error of the mean $\left(\sigma_{\bar{X}}\right)$ or
- a measure of variance $\left(\sigma_y^2\right)$.

These values take different forms but contain the same information and can easily be transformed by applying simple formulas available in most statistical textbooks.

Your literature review of the *Healthy Eats Program* suggests that the measured fruit and vegetable outcome you selected has a standard deviation of $\sigma_y = 1.17$, or a variance of $\sigma_y^2 = 1.37$.

Step 3. Determine the null hypothesis and the Type I (α) and Type II (β) error rates. It is typical to set the Type I error rate at 0.05, allowing only a 1-in-20 chance of incorrectly rejecting a false null hypothesis. It is also common to set the Type II error rate at 0.20, allowing a 1-in-4 chance of incorrectly retaining a false null hypothesis***.*** *By setting the Type II error rate at 0.20, you aim for a study with 80 percent statistical power.* If you had set your Type II error rate at 0.10, you would have set your sights on 90 percent statistical power.

The Type I and Type II error rates define the regions of rejection for the null hypothesis. Numerical values (i.e., critical values) defining these regions are available in tables in most statistical textbooks. Keep in mind that the error rates you establish at this point are for planning purposes. The actual error rates—and the actual power of your study—will be determined by characteristics of the data you collect and analyze.

Step 4. Apply the information. The equation below provides an easy-to-use formula for sample size estimation. In this equation, n is the number of participants per condition. Most studies involve a comparison of two conditions, so the total number of participants will be 2n. Notice that all you need to solve for n is the four pieces of information just reviewed.

$$n = \frac{2\hat{\sigma}_y^2\left(t_{crit(\alpha/2)} + t_{crit(\beta)}\right)^2}{\hat{\Delta}^2}$$

The values for $t_{crit(\alpha/2)}$ and $t_{crit(\beta)}$ are determined by the error rates you established and the expected **degrees of freedom**. Of course, the expected degrees of freedom are determined by sample size, so you need a good working assumption to get started. You can start with the conventional values, $t_{crit(\alpha/2)} = 1.96$ and $t_{crit(\beta)} = 0.84$. These are the values under the null hypotheses associated with a 0.05 Type I error rate and a 0.20 Type II error rate when sample sizes are very large. If you only need a medium-to-small sample (say, fewer than 120 people per condition), you may need to adjust the values you use for $t_{crit(\alpha/2)}$ and $t_{crit(\beta)}$.

Degrees of freedom (df) refers to the number of scores that are free to vary when estimating a sample mean. In general, n−1 scores are free to vary.

…and the answer is:

$$n = \frac{2(1.37)(1.96 + 0.84)^2}{0.33^2} = 197$$

The number solved for n is the sample of participants per condition. Assuming you have an intervention and a control or comparison group, you will need a minimum of 394 individuals (or 197 x 2) to estimate program impacts.

n is the number of participants per condition. Most studies involve a comparison of two conditions, so the total number of participants will be 2n.

This sample size estimation indicates the number of complete observations required for the impact analysis. It is highly unlikely that everyone will respond to the survey, and even those who do may provide incomplete data (i.e., missing data). Therefore, your initial sample size will need to be *larger* than this minimum needed.

For example, you might only expect a 60 percent completion rate. At that rate, assuming all responses are complete, you would need to survey about 660 individuals to achieve your needed minimum of 394. If you're uncertain about completion rates, you should consider consulting with data collection specialists.

This example assumes that the impact of the study will be assessed by comparing the difference between two condition means using a t-statistic. Different assumptions are required for more complex comparisons or when different statistical tests will be used.

For More Information on How to Conduct Sample Size Estimation and Statistical Power Analysis:

- The classic textbook on statistical power is *Statistical Power Analysis for the Behavioral Science* (Cohen, 1988). This book provides an excellent introduction to the topic of statistical power along with numerous tables and formula for different types of comparisons.

- Other useful textbooks include:
  - *Essentials of Behavioral Research—Methods and Analysis* (Rosenthal & Rosnow, 1991)
  - *Fundamentals of Biostatistics* (Rosner, 2000)
  - *Statistical Principles in Experimental Design* (Winer, Brown, & Michels, 1991)

- A number of online resources are available that can assist you with power analysis and sample size calculation:
  - Inference for Means. Based on Rosner's (2000) *Fundamentals of Biostatistics*, this calculator provides the sample size for a prespecified level of statistical power or determines statistical power for a prespecified sample size for a two mean comparison. (http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html)
  - Power Analysis. This Web site hosted by the University of Wisconsin–Madison provides links to several Web page calculators and downloadable applets. (http://psych.wisc.edu/henriques/power.html)
  - G*Power. G*Power is a free online and downloadable power analysis software program. It can perform power analysis tests for the most common statistical tests in behavioral research. (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register/index_html)

Chapter IV • Addressing the Challenges of Conducting Effective Impact Evaluations: Instrumentation

A. How to Develop Effective Evaluation Instruments

Developing Effective Evaluation Instruments: Tips for Best Results

- Use multiple measures and/or data sources
- Include only important variables
- Use valid and reliable measures
- Use existing instruments … don't "reinvent the wheel"
- Use measures that are sensitive to change
- Use measures that are appropriate for the audience's literacy level
- Follow guidelines if you write original questions

Many resources are available to help you design a questionnaire to evaluate the impact of your program. (Detailed sources are available at the end of this section.) Even with this input, you'll want to keep several concepts in mind that will help you develop effective questionnaires or survey instruments.

Here are the most important ones:

Use Multiple Measures and/or Data Sources

By using multiple measures of the same underlying construct that are assessed in different ways and drawn from different sources, you can arrive at conclusions that can greatly strengthen your overall evaluation effort (Posavac, 2010). Consider the case of dietary intake of fruit and vegetables for schoolchildren. Some schools may be able to provide data on what children purchased for lunch. This information is objective, not subject to recall bias, and can provide a day-to-day snapshot of children's preferences for fruit and vegetables. However, because what children select and what they actually eat can vary considerably, you might also collect self-reported data from children on their consumption of fruit and vegetables. This information is subject to recall bias as well as other issues that need to be considered when collecting self-reported data. Taken together, the two data sources, will provide a clearer picture of children's dietary intake for fruit and vegetables, than just one source on its own. If resources permit, this is a good approach to use.

Include Only Important Variables

Ask yourself what questions truly need to be answered, and then limit the variables you include in your survey instrument only to those that are essential to the evaluation.

Resist the "kitchen sink" approach and the urge to include questions just because you think it might be interesting to know this information.

For example, your essential questions may focus on the amount of fruit and vegetables your target population consumes in a given time period. Although you may also want to know about whole grain consumption for this population, including those nonessential questions will lengthen the time that respondents need to complete the survey. When respondents perceive questions to be unimportant or off-track, they may become inattentive and less willing to complete the survey accurately. Ultimately, asking nonessential questions can weaken your results when many respondents do not answer particular questions or decline to complete the survey.

A good rule of thumb: Limit interviewer-administered surveys to about 15 to 20 minutes and self-administered surveys to about 5 to 10 minutes.

Use Valid and Reliable Measures

Survey items or instruments are considered to be **valid** when they accurately measure the activity, behavior, or opinion they're intended to assess. For example, does a self-efficacy scale really measure self-efficacy? Or does it measure a similar, but not the same, construct, such as one's intention to reach a goal? Validity applies not to the survey items or instruments themselves, but rather to the purpose for which they are being used. This means that an instrument could be valid for measuring one certain outcome, such as self-efficacy, but entirely invalid for measuring a different outcome, such as intentions (Iversen, Gudmund, & Helmut, 1976).

The **reliability** of an instrument or survey items is shown by the extent to which it produces the same result when applied to the same person under the same conditions. Reliability is often assessed using the test-retest method in which the same set of subjects is measured twice under similar conditions and the strength of the relationship between the two measurements is examined (Iversen, Gudmund, & Helmut, 1976).

Reliability versus Validity

Reliability refers to the consistency of the measures, and **validity** refers to the accuracy of the measures.

Although reliability is necessary, by itself it can be insufficient. A survey instrument must be both reliable and valid. For example, if your scale is off by 5 pounds, it reads your weight every day as being 5 pounds higher than it really is. Although the reliability (consistency) of this scale is very good because it consistently reports the same weight every day, it is not valid (accurate) because you actually weigh 5 pounds less.

If you are using a survey instrument to measure fruit and vegetable consumption, you can test the instrument for reliability (consistency) by having a subset of participants complete the instrument a second time. If their responses are consistent, your instrument is reliable. To find out if your instrument is valid (accurate), have participants also complete a food diary or a 24-hour dietary recall to measure if their survey responses are accurate and valid. The next section (Chapter IV.B) briefly describes how to conduct validity and reliability testing when developing new survey items.

Use Existing Instruments … Don't Reinvent the Wheel

Evaluating validity and reliability requires empirical evidence, which can be costly and take a lot of time to collect. To save your time and resources, use existing instruments that have proven to be valid and reliable for measuring your outcomes of interest.

Use What's Out There

Whenever possible, use existing instruments that have been previously demonstrated to be valid and reliable for measuring the outcomes of interest.

You can identify valid and reliable measures and instruments by reviewing the literature. Many researchers publish papers that describe the results of the validity and reliability testing with specific populations and data collection modes (e.g., paper-and-pencil, telephone). Exhibit IV-1 provides an example of how validity testing and reliability testing were conducted and the type of results reported for the Food Behavior Checklist (FBC) (Townsend, Kaiser, Allen, Joy, & Murphy, 2003).

Exhibit IV-1.— Understanding Validity and Reliability: An Example Using the Food Behavior Checklist

Townsend and colleagues developed and tested a Food Behavior Checklist (FBC) for a limited-resource audience. The 22-item checklist collected information on the dietary topics included in the Expanded Food and Nutrition Education Program (EFNEP) and Food Stamp Nutrition Education (now referred to as SNAP-Ed) interventions such as increasing fruit, vegetable, and calcium intake; decreasing fat intake; and assessing food security status. Data were collected from 132 limited-resource women at two points in time to measure reliability of the FBC and sensitivity to change (following a nutrition intervention). Three 24-hour dietary recalls were conducted and serum sampled to measure validity of the FBC. The results of the validity and reliability testing are reported in Townsend, Kaiser, Allen, Joy, & Murphy (2003) and summarized below. Based on the results of this testing, the 22-item checklist was revised to a 16-item checklist and is available at http://townsendlab.ucdavis.edu.

Reliability: Reliability coefficients (*r* values) showed that 20 of the 22 items met the criterion ($p < 0.05$) for reliability.

Internal consistency: The Cronbach alpha correlation coefficients for the fruit/vegetable subscale ($\alpha = 0.80$) and diet quality subscale ($\alpha = 0.60$) had acceptable internal consistency values.

Criterion validity: The fruit and vegetable subscale showed a significant correlation with serum carotenoid values ($r = 0.44$, $p < .001$), indicating acceptable criterion validity.

Convergent validity: Convergent validity was defined as the Spearman rank order correlation coefficient of the subscale score with the hypothesized nutrients and food groups from the mean of the three 24-hour dietary recalls. Milk, fat/cholesterol, diet quality, food security, and fruit/vegetable subscales showed significant correlations with dietary variables.

Sensitivity to change: Eleven of the 15 items expected to show change following the intervention demonstrated sensitivity to change.

Source: Townsend, M. S., Kaiser, L. L., Allen, L., Joy, A., & Murphy, S. (2003). Selecting items for a food behavior checklist for a limited-resource audience. *Journal of Nutrition Education and Behavior, 35*(2), 69–82.

The *Nutrition Education Impact Measurements/Instruments Literature Review*, conducted as part of the SNAP-Ed Education and Evaluation Study (Wave I) (Gabor et al., 2012a), is a useful resource for outcomes related to consumption of fruit and vegetables and low-fat/fat-free milk. Appendix B of the Wave I report describes the approach used to conduct the literature review and develop the templates that were used to summarize information on the measurements and instruments reviewed. The populated templates with summary information for each instrument reviewed are available at http://www.fns.usda.gov/research-and-analysis.

For fruit and vegetables outcomes, information is provided on food frequencies and food behavior checklists to measure dietary intake, as well as items used to measure other nutrition behaviors related to fruit and vegetable consumption, such as

- availability,
- accessibility,
- willingness to try,
- preferences, and
- self-efficacy.

For low-fat/fat-free milk outcomes information is provided for items used to measure

- intake,
- availability,

- self-efficacy, and
- willingness to serve.

Each summary provides information on the following (see Exhibit IV-2 for an example):

- *Instrument:* developer and audience
- *Survey administration:* mode, length, and languages
- *Measurement properties:* cognition, reliability, validity, and sensitivity to change

Exhibit IV-2.— Example of an Entry from the *Nutrition Education Impact Measurements/Instruments Literature Review*: Fruits, Fruit Juices, and Vegetables (FJV) Preference Measure

Instrument	
Type	Questionnaire
Developer	Jaramillo and colleagues
Original audience	African American and Hispanic preschoolers
Topic and number of items	Fruit (15) and vegetable (15) intake
Survey Administration	
Year	Not reported
Study population and size	n = 198 African American and Hispanic preschool children enrolled in 12 Head Start Centers in Houston
Modification	New instrument
Mode	Self-administered on computer
Length of administration	15 minutes
Other languages	Spanish
Measurement Properties	
Cognition	Expert panel review and pilot tested with children
Reliability (internal consistency, test-retest)	For internal consistency, Cronbach $\alpha_{overall}$ = 0.87, Cronbach α_{fruit} = 0.77, Cronbach $\alpha_{fruit\ juices}$ = 0.58, Cronbach $\alpha_{vegetables}$= 0.82. The overall coefficient for test-retest was 0.73 ($p < 0.001$); 0.49 for fruit ($p < 0.0001$); 0.37 for fruit juices ($p < 0.01$); and 0.73 for vegetables ($p < 0.0001$).
Validity (convergent validity, criterion validity)	Not evaluated
Sensitivity to change	Not evaluated
References	Jaramillo et al., 2006.
Notes	Mean fruit and vegetable consumption was significantly higher in children who reported higher preferences for fruit and vegetables compared with those who reported lower fruit and vegetable preferences ($p < 0.02$).

Appendix A summarizes dietary and nutrition behavior instruments that have been used with low literacy and limited resource audiences. These instruments may be a useful starting point as you develop the impact instrument for your evaluation.

Other valid, reliable nutritional assessments have been developed by the National Cancer Institute (NCI). The Diet History Questionnaire (DHQ II) is a detailed food frequency questionnaire that provides information on total dietary intake (available at http://appliedresearch.cancer.gov/dhq2/). The Dietary Screener Questionnaire (DSQ) is a much shorter instrument that may be useful for impact evaluations. It was administered in the National Health and Nutrition Examination Survey (NHANES) NHANES 2009-2010 (available at http://appliedresearch.cancer.gov/studies/nhanes/dietscreen/).

Conducting 24-hour dietary recalls is another option for measuring dietary intake that you may want to consider. Although 24-hour dietary recalls are considered by many to be the gold standard for measuring dietary intake, they can be time-consuming and relatively expensive to administer. NCI's Automated Self-Administered 24-hour Recall, or ASA24™, is a free Web-based tool that allows respondents to make multiple self-reports of food consumed in the previous 24 hours (available at http://appliedresearch.cancer.gov/tools/instruments/asa24). An associated Researcher Web site allows study managers to handle study logistics and analyze. Additionally, you will need someone with experience analyzing 24-hour dietary recall data.

Use Measures That Are Sensitive to Change

The term "sensitivity" refers to the ability of a measure to detect variation associated with the measured outcome before and after the intervention. Continuous measures of an outcome (e.g., a 5-point Likert scale) are more sensitive to detecting change than dichotomous (yes/no) measures. This is because continuous measures of an outcome tend to have smaller standard deviations. The denominator of the statistic used to assess a program's impact is generally influenced by two factors:

- sample size
- measurement error

When you can't afford to recruit or collect data on a large number of participants, you can enhance your ability to identify program-related change by carefully selecting your measurement tools.

Exhibit IV-3 provides two examples using continuous vs. dichotomous measures for measuring change in an outcome.

Exhibit IV-3.— Use Continuous Instead of Dichotomous Measures so That the Measures Are Sensitive to Change

Example 1: Child's Preferences for Specific Fruit and Vegetables

Dichotomous Measure

For each food pictured:

Circle the  if you like it.

Circle the  if you do not like it.

Circle the  if you do not know what it is.

Continuous Measure

For each food pictured, circle a number to show how much you like (smiley face) or do not like the food (sad face).

0 1 2 3 4 5 6 7 8 9 10

There are two possible concerns with the dichotomous measure shown. First, the use of the question mark is equivalent to a response of "do not know" and is treated as missing data. The presence of missing data is problematic when attempting to develop multi-item scales (e.g., using the preference information for individual fruits to develop an overall preference measure for fruits), because varying degrees of "missingness" across items can lead to the creation of biased measures. Second, the dichotomous measure may not be appropriately sensitive to change as previously noted. The use of the continuous response set, such as the visual analog shown with appropriate "faces" for the endpoints, will allow you to detect more nuanced changes in children's preferences.

Example 2: Meal Planning by Adults

Dichotomous Measure	Continuous Measure
Do you plan meals?	How often do you plan meals?
☐ Yes ☐ No	☐ Almost never ☐ Often
	☐ Once in a while ☐ Almost always
	☐ Sometimes

The dichotomous measure of meal planning is not very sensitive to change. However, the continuous measure shown, a categorical question with five levels, will be more sensitive to change. For analysis purposes, convert the 5-level categorical variable into a continuous variable by assigning "almost never," a value of 1, "once in a while" a value of 2, "sometimes" a value of 3, "often" a value of 4, and "almost always" a value of 5.

Use Measures That Are Appropriate for the Audience's Literacy Level

Making sure your audience understands the questions you are asking in your survey is another important, and sometimes overlooked, consideration in your evaluation. For evaluations of SNAP-Ed interventions, questions should be easily understood by low-literacy and limited-resource audiences. For best results, use existing instruments that have been tested for literacy, especially ones that have been used with limited-resource audiences (see Appendix A).

Follow Guidelines If You Write Original Questions

If existing instruments don't meet your needs, you may need to develop a new instrument or introduce new questions. In these situations, be sure to test the original questions with persons from the target population. Testing will help assess whether questions are understood as written or need revisions (see Chapter IV.B below). As you go through the question-writing process, follow these commonly accepted guidelines:

- **Begin by asking questions that clearly relate to the program's topics and goals to build face validity and engage respondents.** For example, when asking respondents for their opinions about program quality, ask about specific domains (e.g., printed materials, family nights out) before asking a general summary question. Ask sensitive items near the end of the questionnaire and conclude with questions that confirm demographics (age range, race/ethnicity). Be sure to thank respondents for their time on the last page of a self-administered instrument or at the end of an interview.

Developing New Questions? Follow these guidelines when developing new questions. Be sure to test your questions to make sure they are understood by your target audience.

- **Consider whether respondents are willing or able to answer all the questions you would like to ask them.** It can be difficult for an individual with irregular living habits or in a transient living situation to recall detailed information. Respondents might not have formed attitudes about topics that researchers find important. Respondents may also have a difficult time accurately reporting the behavior or attitudes of other household members. For these and other reasons, avoiding certain types of questions is an important part of questionnaire design.
- **Avoid open-ended questions whenever possible,** including "Other, Specify" responses. Low-literacy populations can find writing to be a challenging process, and their written responses may be difficult to read or interpret. Also, answers written in text take more time to code and quantify for reporting purposes.
- **Avoid numerous, wordy, or complicated questions that involve many skips and changes in topic.** Respondents may become fatigued and break off before completing the survey or resort to "satisficing," or responding only superficially to complete the task.
- **Write questions that address one subject at a time; avoid double-barreled questions.** Present response options that are mutually exclusive. Doing so avoids ambiguity as the respondent attempts to map his response onto the available options.
- **Avoid complicated syntax, technical jargon, or advanced vocabulary.** If slang terms must be used for a given topic or study population, introduce them accordingly (e.g., "…sometimes known as soda or pop.")
- **Word questions neutrally to avoid response bias.** When presenting a series of questions that ask for respondents to rate their agreement, ask a mix of positively and negatively worded items (e.g., "The information was useful," "It was hard to schedule an appointment").
- **Be specific about reference periods.** For example, use "in the past 7 days" instead of "in the past week."
- **Explain that survey data will be handled confidentially and follow through with that promise.** Use a system that has unique study IDs to avoid naming respondents on paper forms. Report data in summary form so that individuals cannot be identified.
- **Ask colleagues to review your draft questions to get a fresh perspective.** Rewrite items as needed and return them to the original reviewer and a naive reviewer, if possible. You may have to go through this process several times. Reread questions aloud to yourself. Get an estimate of how long the survey will take to complete by giving the questions to someone unfamiliar with the evaluation. Make final edits and format the instrument before you test it with people recruited from the target population.

For More Information on Developing Effective Evaluation Instruments

- Converse, J., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Newbury Park: Sage Publications.
- Devellis, R. F. (2012). *Scale development: Theory and applications* (3rd Ed.). Thousand Oaks, CA: Sage Publications.
- Fink, A. (2004). *Evaluation fundamentals* (2nd Ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Fowler, Jr., F. J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage Publications.
- National Collaborative on Childhood Obesity Research (NCCOR). (2012). Measures registry. Retrieved from http://nccor.org/projects/measures/index.php
- Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65–88.
- Sudman, S. & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

B. How to Test Evaluation Instruments

If you need to develop new questions or modify existing questions, you should test your evaluation instrument with members of your target population by conducting cognitive interviews or one-on-one sessions with a respondent and tester/interviewer. Other important steps you'll need to take include ensuring that new questions can easily be read by your target population, making sure your questions are valid and reliable, and providing translation of the instrument if it will be administered in a language other than English.

Conduct Cognitive Interviews to Evaluate Respondent Understanding of the Survey Questions

Through one-on-one sessions with a respondent and tester/interviewer, the cognitive interviewing approach allows you to examine the thought processes of respondents, which can, in turn, affect the quality of the respondent's answers to survey questions (Willis, 2005).

Cognitive interviews can provide useful information on the

- manner and degree to which respondents understand the words and phrases in the questions,
- how respondents recall information and appropriate cues to help aid recall,
- cognitive complexity of the questions and the strategies used by respondents to answer the questions,
- extent to which respondents are answering questions as intended,
- ability of respondents to make any calculations and judgments, and
- whether any important responses are missing from the question response list.

There are two approaches to conducting cognitive interviews:

1. The respondent reads aloud from the draft survey (especially if the instrument will be self-administered) or
2. You (the interviewer) read the questions aloud to the respondent.

You can then use what is called the "think aloud" technique in which you ask a respondent to describe his/her thought process while answering the survey questions. Remember to use a written protocol that includes the survey questions and follow-up "probes" to learn how respondents understand the meaning of questions. These probes often ask targeted questions about anticipated problems with the survey questions (see Exhibit IV-4).

Exhibit IV-4.— Sample Follow-Up Probes for Cognitive Interviews

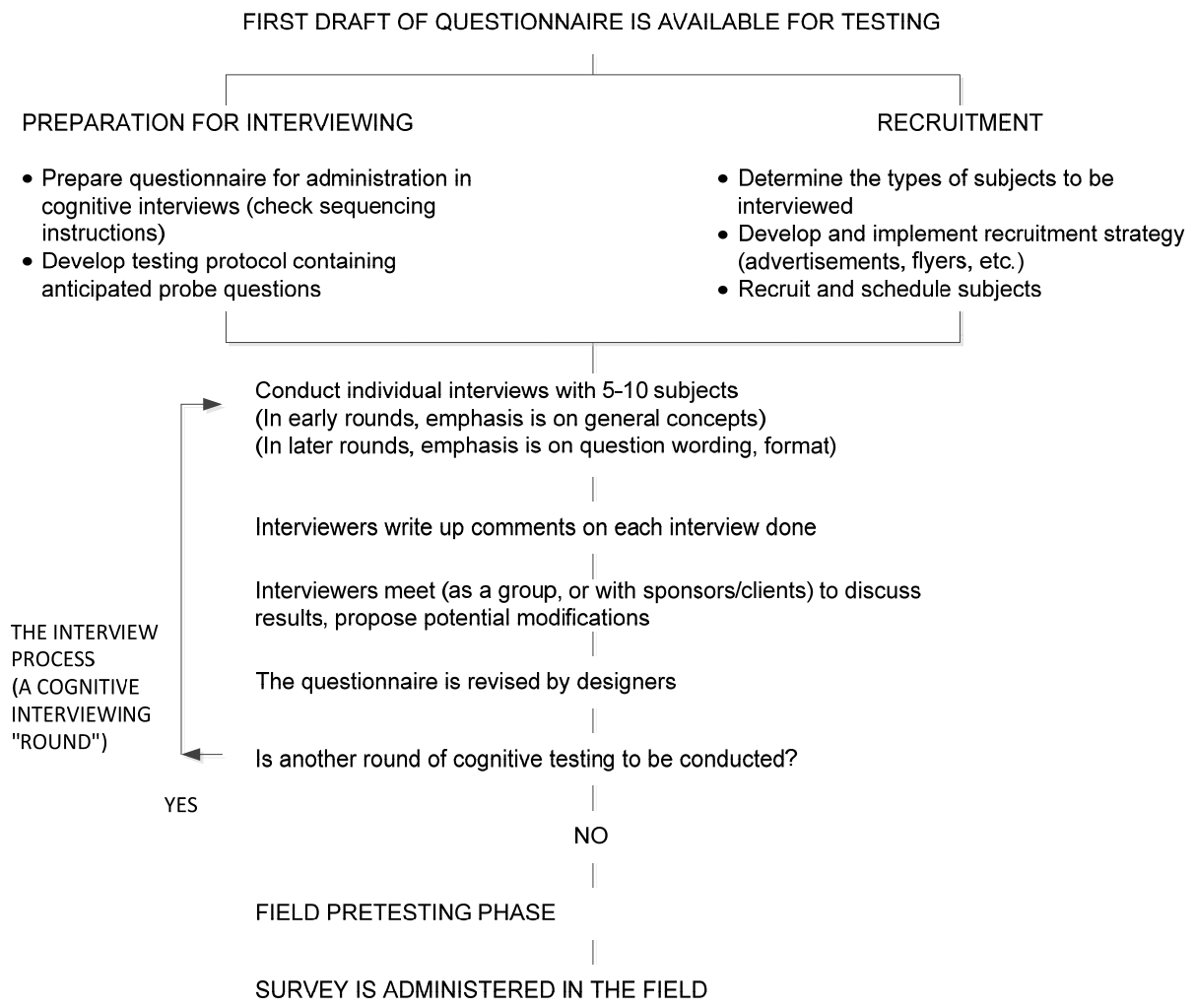
- *"How would you put that into your own words?"* for instructional text, complex questions, or technical terms
- *"How far back did you think when answering that question?"* for a question asking about past behavior
- *"What response options should be added there?"* for a question asking respondents to select one or more responses
- *"How does that make you feel?"* for a sensitive question
- *"Where would you go next in the survey?"* for a skip instruction

In concurrent probing, you will ask probes while the respondent answers the questions. In retrospective probing, the probes are asked at the end of a survey module (set of questions) or at the end of the survey. If retrospective probes are used, you can keep track of the amount of time that it takes to complete the questionnaire to estimate survey burden.

The independent evaluator for the Waves I and II SNAP-Ed evaluations conducted cognitive interviews of the instruments using probes prepared before the interview. Probing was retrospective, or conducted after the respondent completed the questionnaire. Appendix B provides the debriefing guide used by the independent evaluator to test the instrument used for the evaluation of the child-focused programs in Wave II.

Exhibit IV-5 shows the steps in cognitive testing within the framework of instrument development (Willis, 2005). Be sure to allow adequate time in the schedule to conduct cognitive interviews. If substantial changes are made to the survey questions, you may want to conduct an additional round of interviews before the full-scale study. In some cases, you may want to conduct field pretesting. Field pretesting involves administering the survey to a small number of people from the target audience, using the same procedures as the full-scale study. This allows you to test the process of survey administration and to identify and correct any potential problems.

Exhibit IV-5.— Steps in Conducting Cognitive Interviews Within the Framework of Instrument Development



Source: From *Cognitive interviewing: A tool for improving questionnaire design*, by G. B. Willis, 2005. Copyright (2005) by Sage Publications. Reprinted with permission.

For More Information on Conducting Cognitive Interviews

- Willis. G. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.

Make Your Instructions and Questions Understandable to Your Audience

All of your efforts in carefully designing a survey instrument won't deliver good results if your target population can't read or understand the questions or how to complete the survey.

A variety of tests are available to assess reading ease and grade level.

Target your questions and instruction to a reading level of 8th grade or less for the SNAP population and other low-literacy and limited-resource audiences.

- The Flesch Reading Ease test (Flesch, 1948) assesses reading ease; higher scores indicate that the material is easier to read.
- Other readability tests, such as the Flesch–Kincaid Grade Level Formula (Kincaid et al., 1975), the Gunning Fog index (Gunning, 1968), the Coleman–Liau index (Coleman & Liau, 1975), the Fry Test (Fry, 1968), and the Automated Readability Index (ARI) (Senter & Smith, 1967), provide a measure of readability in terms of grade level. For example, the Flesch–Kincaid Grade Level Formula considers the average number of words per sentence and the average number of syllables per word within a given passage in order to estimate the complexity of the text. The formula then converts that complexity level into a score that corresponds to a U.S. school grade level.
- The Coleman–Liau index and the ARI rely on a factor of characters per word, instead of the usual syllables per word to estimate the complexity of the text and provide a score that corresponds to a U.S. school grade level.

The Web site http://www.readability-score.com/ can perform several different readability tests (and provides an average across all tests). Microsoft Word® can conduct the Flesch Reading Ease and Flesch-Kincaid Grade Level tests.

Exhibit IV-6 provides an example of two questions that ask about the same construct (self-efficacy) and the readability test scores for each. The second question uses words with fewer syllables (1.2 syllables per word versus 1.5) and fewer words per sentence (14 versus 18). It is easier to read and has a lower reading level (4.6 versus 9.1 grade level).

Exhibit IV-6.— Results of Readability Testing for Two Questions on Self-Efficacy

Harder to Read	**Easier to Read**
How confident are you that you can serve your child fruit to eat as a snack instead of foods like cookies or potato chips every single day of the week? ☐ Not confident ☐ Somewhat confident ☐ Very confident	How sure are you that you can serve your child fruit for a snack instead of cookies or chips every day? ☐ Not sure ☐ A little sure ☐ Very sure
Flesch Reading Ease (higher score indicates easier to read): 61.7 Flesch–Kincaid Grade Level Formula: 9.1	Flesch Reading Ease: 86.9 Flesch–Kincaid Grade Level Formula: 4.6

For More Information on Assessing Readability and Literacy

- Albright, C. L., Bruce, B., Howard-Pitney, B., Winkleby, M. A. & Fortmann, S. P. (1997). Development of a curriculum to lower dietary fat intake in a multiethnic population with low literacy skills. *Journal of Nutrition Education, 29*(4), 215–223.
- Klare, G. R. (1974–1975). Assessing readability. *Reading Research Quarterly, 10*(1), 62–102.
- Parker, R. I., Hasbrouck, J. E., & Weaver, L. (2001). Spanish readability formulas for elementary-level texts: A validation study. *Reading & Writing Quarterly, 17*(4).
- Wang, L. W., Miller, J. J., Schmitt, M. R., Wen, F. K. (2013). Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Research in Social and Administrative Pharmacy*, *9*(5), 503–16.

Measure the Reliability and Validity of New Survey Items

Suppose you want to develop a new survey item that seeks to measure the amount of fruit that members of a household consume each day. You'll want to test your survey items to make sure they are reliable (consistent) and valid (accurate). Exhibit IV-7 provides a brief description of different types of reliability and the suggested statistical test. Although these tests can be done manually, you'll probably want to use statistical software program (e.g., SAS or Stata). When testing for reliability, correlation coefficients (r values) or Cronbach's coefficient α are generally considered good if they equal or exceed 0.70.

Exhibit IV-7.— Types of Reliability

Type of Reliability	Description	Statistical Test
Test-retest	Indicator of the stability of responses over time that is measured by having the same set of respondents complete the survey (a single item or group of items or scale within an instrument) at two different time points.	Correlation coefficient or r value is calculated to compare the two sets of responses.
Internal Consistency	Indicator of how well a group of items in a scale measures the same characteristic or concept.	Measured by calculating Cronbach's coefficient α, a statistic that reflects the homogeneity of the scale.
Interobserver or inter-rater	Indicator of how well two or more interviewers agree in their assessment of a variable; should be used whenever there is a subjective component in the measurement of an external variable.	Measured as a correlation coefficient between different data collectors.

Source: *How to measure survey reliability and validity*, by M. S. Litwin, 1995. Copyright (1995) by Sage Publications. Adapted with permission.

Exhibit IV-8 provides a brief description of different types of validity. Validity is usually expressed as a correlation coefficient, or r value, between two sets of data. Levels of 0.70 or more are generally accepted as representing good validity.

Exhibit IV-8.— Types of Validity

Type of Validity	Description
Face	Casual review of how good an item or group of items appears as assessed by individuals with no formal training in the subject matter of interest.
Content	Formal expert review of how good an item or series of items appears, usually assessed by individuals who are experts in the subject matter of interest.
Concurrent criterion	Measures how well the item or scale correlates with the "gold standard" measure of the same variable. For example, for measuring dietary intake, the "gold standard" would be 24-hour dietary recalls.
Predictive criterion	Measures how well the item or scale predicts expected future observations.
Construct	Theoretical measure of how meaningful a survey instrument is, and is usually determined after years of experience by numerous investigators (more like hypothesis testing than calculation of correlation coefficients).

Source: *How to measure survey reliability and validity*, by M. S. Litwin, 1995. Copyright (1995) by Sage Publications. Adapted with permission.

For More Information on Measuring Survey Reliability and Validity

- Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Litwin, M. S. (1995). *How to measure survey reliability and validity*. Thousand Oaks, CA: Sage Publications.
- Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications*. Thousand Oaks, CA: Sage Publications.

Hablo español (I Speak Spanish) … How to Translate Survey Instruments

If a large percentage of study participants speak a language other than English, the survey instrument and supporting materials (e.g., consent forms, recruitment and follow-up letters) need to be available in that language. If the intervention materials are translated into another language, then the survey should be translated into this language so that the evaluation results are representative of all participants.

The focus of translation should be **cross-cultural and conceptual**, rather than linguistic/literal equivalence.

However, not all survey translations are created equal.

A good translation conveys the meaning of the original text rather than mechanically transferring words from one language to another. This will ensure a functionally equivalent representation of a survey. Translated documents should be easily understood by a native speaker of the target language; grammatically and terminologically correct; and free of any omissions, additions, or typographical errors. A translated instrument that is not **functionally equivalent** to the source language (i.e., English) version can lead to biased results, and poor translations can lead to increased variance (Weidmer, 1994).

Adapting the translated instrument to the culture of target respondents (i.e., **cross-cultural adaptation**) could require changes to ensure that the instrument reflects the cultural assumptions of the target respondent, not just those of the researchers. For example, don't assume that a construct in English has an equivalent in another language. Questions in English and another language are considered to be

functionally equivalent if they measure the same construct and take into account fundamental differences between language and cultural groups (Smith, 2002). Questions or answer categories can be grammatically correct, use common or literal translations, and not be functionally equivalent.

To address these concerns and to produce a good translation, translators should use an approach called **forward translation** and then conduct cognitive interviews to test the functional equivalence of the translation. Based on the findings of the cognitive interviews, translators can then adapt the translation as necessary. Cognitive interviews will help validate how respondents interpret the translation of key questions or terms and determine whether respondents' interpretations are consistent with the English instrument. This will help improve the quality of the data.

If possible, use a native-speaking translation professional to translate the survey instrument, not someone who simply speaks the other language. Also use a professional editor who is a native-speaking translation professional to review and edit the translated document. Exhibit IV-9 provides a step-by-step approach for conducting forward translations.

Exhibit IV-9.— Steps in Conducting Forward Translations

Step 1: Prepare for the forward translation
Provide the translator with the source documents (i.e., survey instrument and other survey materials) and the intervention materials (English and translated version). The translator should prepare for the translation task by pulling out the key terminology and researching any unfamiliar vocabulary used in the source documents. The translator should review the intervention materials so that translated words and phrases are consistent with the intervention materials. For example, if referring to the intervention by name, be sure the translated name of the intervention (used in the intervention materials) is the same name used in the survey instrument. In some cases, the name of the intervention may not be translated in the intervention materials; thus, this same approach should be used in the survey instrument.
Step 2: Conduct the forward translation
Provide the translator with instructions on the approach to translating, emphasizing conceptual rather than literal translations, and the need to use natural and acceptable language for the broadest audience. As recommended by the World Health Organization (WHO), the guidelines shown on page 40 should be considered in this process. The focus of the translation should be on cross-cultural and conceptual, rather than on linguistic/literal, equivalence.
Step 3: Edit the revised document
The editor reviews the entire translation against the source documents, and if she/he has questions, the editor contacts the project staff to resolve any issues relating to style, meaning, or terminology. The editor is responsible for finalizing the translation making the necessary corrections and improvements.
Step 4: Conduct cognitive interviews to test the translation
Using the finalized documents, the translator or another native-speaking individual, conducts cognitive interviews with native speaking individuals from the target audience to test the functional equivalence of the translation and to validate how respondents interpret the translation of key questions or terms.
Step 5: Finalize the translated instruments and other survey materials
Based on the findings of the cognitive interviews, the translator adapts the translation as necessary. If significant changes are required, conduct another round of interviews (with a smaller number of participants), time and resources permitting.

The World Health Organization (WHO) Guidelines for Forward Translation

1. Translators should always aim at the conceptual equivalent of a word or phrase, not a word-for-word translation (i.e., not a literal translation). They should consider the definition of the original term and attempt to translate it in the most relevant way.
2. Translators should strive to be simple, clear, and concise in formulating a question. Fewer words are better. Long sentences with many clauses should be avoided.
3. The target language should aim for the most common audience. Translators should consider the typical respondent for the instrument being translated and what the respondent will understand when s/he hears the question.
4. Translators should avoid the use of any jargon. For example, they should not use
    a. technical terms that cannot be understood clearly and
    b. colloquialism, idioms, or vernacular terms that cannot be understood by common people in everyday life.
5. Translators should consider issues of gender and age applicability and avoid any terms that might be considered offensive to the target population.

Source: World Health Organization (WHO). (n.d.). *The process of translation and adaptation of instruments*. Retrieved from http://www.who.int/substance_abuse/research_tools/translation/en/

For More Information on Translation of Data Collection Instruments and Supporting Materials

- Behling, O., & S. Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. Thousand Oaks, CA: Sage Publications, Inc.
- Esposito, N. (2001). From meaning to meaning: the influence of translation techniques on non-English focus group research. *Qualitative Health Research, 11*(4), 568–79.
- Harkness, J. A., van de Vijver, F. J. R., Mohler, P. P. (2002). *Cross-cultural survey methods*. New York, NY: Wiley Publishers.
- Pan, Y., & de la Puente, M. (2005). Census Bureau guideline for the translation of data collection instruments and data collection instruments and supporting materials: Documentation on how the guideline was developed. Research Report Series (Survey Methodology #2005-06). Washington, DC: Census Bureau.
- World Health Organization (WHO). (n.d.). *The process of translation and adaptation of instruments*. Retrieved from http://www.who.int/substance_abuse/research_tools/translation/en/

Chapter V • Addressing the Challenges of Conducting Effective Impact Evaluations: Data Collection

A. How to Effectively Recruit/Enroll Study Sites and/or Participants

Although recruiting and enrolling study sites or individual participants may seem like a daunting task, you can break it down into straightforward steps. In this chapter, we take a look at a few key considerations before you start and then outline the steps to make the recruitment and enrollment work as smoothly as possible.

Which Approach Do You Need?

If your intervention is site based, you'll need to recruit the intervention and control/comparison sites (e.g., schools or centers) and individual participants, so it's a two-step process. For interventions that are not site based, you'll only need to recruit individual participants. Two key considerations for either approach are to

- recruit a sample of participants that adequately represents the target population and
- recruit a sufficient number of participants to meet the sample size and power requirements of the study (as described in Chapter III.C).

The sidebar outlines six steps evaluators can take to facilitate the enrollment process, which are briefly described below.

Six Steps for Recruitment Success!

1. Plan ahead
2. Establish strong working relationships with partners
3. Establish and adhere to enrollment criteria
4. Train recruiters
5. Monitor recruitment closely
6. Have a back-up plan

1. Plan Ahead

Whether you want to recruit a study site (such as a school or center) or enroll individuals to take part in a study intervention, your first step is the same: plan ahead. Many of the topics in this section apply to recruiting both sites and individual participants, so we examine them together. Planning ahead is critical, because the recruitment/enrollment process is time-consuming and can change on a moment's notice.

Recruiting Study Sites. Because recruiting a study site is often the most time-consuming part of this process, it should be your first order of business. The size of study sites will play a role in how much advance planning you need to do. For example, it can take up to a year to get cooperation from all levels of a public school district and to get your program on the school calendar. Larger school districts may require that you submit a research application to a review board before giving approval.

However, getting the go-ahead from smaller study sites, such as childcare or senior centers, may take only a few weeks or months. Giving small centers months of lead time isn't necessarily a good thing because staffing changes or new priorities can cause their interest to wane. Aim for a balance of a reasonable lead time and efficiency.

Whether you're recruiting a large or small study site, remember to

- pay attention to your own staffing levels in the planning and recruitment process,
- avoid an overly ambitious delivery schedule, and
- keep in contact with your study sites so that they stay engaged and informed until the study is ready to begin.

Developing Recruitment Materials. Developing concise, clear recruitment materials for study site administrators and participants can help you gain—and keep—their interest and cooperation.

If you are recruiting a study site, the person reading your materials is most likely a busy administrator who may not be familiar with your project. To keep his or her attention, you'll want to

- prepare a concise, one-page summary of your project;
- emphasize your efforts to minimize any administrative burdens the host site may be concerned about;
- obtain the direct email or mailing address for your point of contact; and
- follow up with a phone call within the same week you send your materials.

Materials for individual participants also need to be clear and concise. You can help participants better understand your project by

- summarizing key information in a "Frequently Asked Questions" (FAQ) brochure,
- writing in an easy-to-understand way (preferably 3rd- to 5th-grade reading level),
- explaining the consent process in simple language and separate from the consent form itself, and
- including a local and/or toll-free phone number.

If possible, try to provide incentives for study sites and individuals who participate in the study. Cash or gift cards may not be feasible, but some form of recognition (e.g., a certificate of completion or appreciation) is usually welcome.

Did you know an incentive can be as simple as a certificate of completion? Most participants are happy to receive the recognition.

Recruiting Participants. Recruiting individual participants, especially for a site-based intervention, is a relatively quick process. After your materials are prepared, allow between 2 and 2 ½ weeks for individuals to enroll, especially if the site encourages participation. Beyond this time frame, enrollment among participants can decline and cooperation by the study site can wane.

2. Establish Strong Working Relationships with Partners

Think of staff at your study site as your "feet on the ground" in helping to recruit participants. Therefore, establishing strong relationships with partners can make a big difference in your project's success.

Here are some tips for establishing successful working relationships:

- Establish partnerships early on.
- Identify a main "point of contact" at each site and keep in touch with partners on an ongoing basis.
- Establish rapport with administrative staff or others who may be more knowledgeable of day-to-day happenings and routines.

- Identify where each site might perceive burden in cooperating with your intervention and/or evaluation.
- Strategize with staff to minimize these burdens without compromising the integrity of your study.
- Repeatedly express appreciation for their cooperation and support.

Exhibit V-1 summarizes the key elements of successful partnerships. Taking the time to help partners understand the mutual benefits of partnering, establishing clear channels of communication, and developing an understanding of respective roles can help provide the foundation for strong relationships.

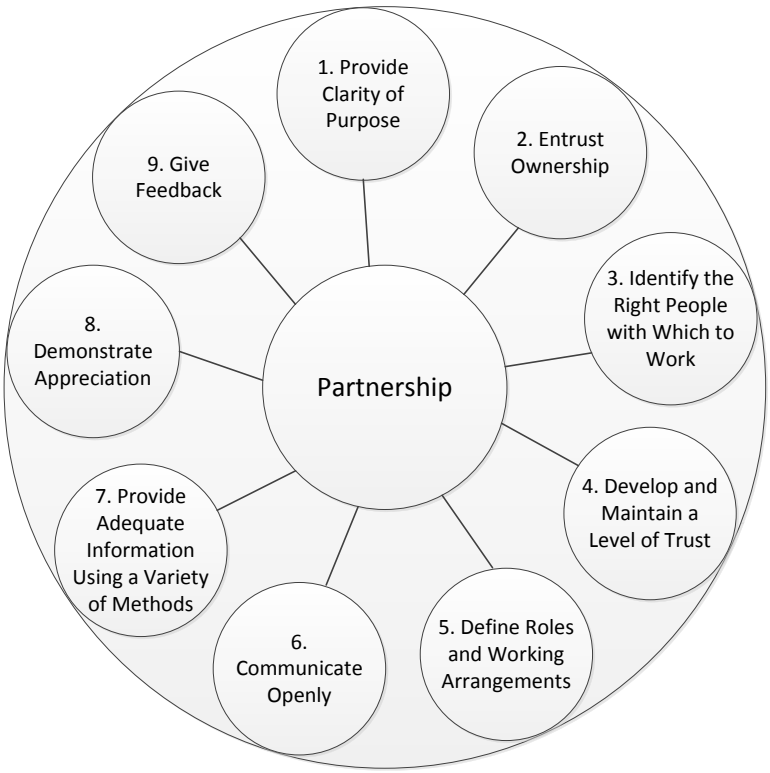
3. Establish and Adhere to Enrollment Criteria

Enrollment Criteria Are Two-Fold. They are made up of "**inclusion criteria**" or criteria that must be met for people to participate in your study and "**exclusion criteria**" or criteria that eliminate people from participating in your study.

To make sure your study results are scientifically sound, you'll need to establish and stick to specific enrollment criteria. This includes inclusion criteria (criteria that must be met for people to participate in the invention/evaluation study) and exclusion criteria (criteria that eliminate people from participating in the study). Enrollment criteria often vary for sites and participants.

For example, for a SNAP-Ed program targeted to seniors to increase fruit and vegetable consumption, the exclusion criteria for sites may be sites that are assisted living facilities and sites that serve more than one meal per day. That's because seniors who live in these centers may have limited opportunities to eat more fruit and vegetables at meal and snack time because they are not preparing these meals. The inclusion criteria for participants might be healthy adults between the ages of 60 and 75, which would exclude people outside the targeted age range of the program curriculum.

Exhibit V-1.— Nine Key Elements of Successful Partnerships



Staff members in charge of enrolling participants need to understand and follow the enrollment criteria to ensure that participants represent the study population. In some cases, it may appear rude or discriminatory (e.g., based on age) to exclude certain people from study participation. In that case, these people may be allowed to participate in study-related activities, but they should be clearly identified so that their data can be excluded from your analysis. Also, it's a good idea to recruit more participants than you need so that the number of participants with usable data is sufficient for the analysis based on your power calculations.

4. Train Recruiters

Recruiters play an important role in the success of your intervention. They are the people who are responsible for identifying and getting the buy-in of study sites and participants. Providing them with thorough, timely training will ensure that you convey a clear and consistent message, use appropriate strategies, reflect a wide variety of approaches, and prepare answers to likely questions. Typically, study team members are in charge of recruiting study sites, and study team members or employees of the study sites recruit study participants.

Training should provide the following information:

- overview of the intervention and evaluation study
- recruitment procedures, including the enrollment criteria
- strategies for marketing the study to sites and/or participants
- techniques for handling noncooperative sites
- methods for working with reluctant participants as well as those who refuse to participate and the refusal conversion process (if applicable)
- evaluation protocol (with study team staff)

Fine-Tune Recruitment Skills With Role-Playing

Role-playing provides trainees multiple opportunities to fine-tune their recruitment skills. In a group setting, try modeling various approaches by having your trainees work in pairs.

5. Monitor Recruitment Closely

Start your monitoring efforts right after recruitment begins and continue it throughout the entire recruitment process. Careful monitoring will help make sure that sites and participants meet your enrollment criteria and that you are meeting your recruitment targets (e.g., by week 3, 50 percent of sites and/or participants have been enrolled). Also, if you are randomizing participants to study groups (instead of to study sites), check early in the process to make sure that the randomization process is being done right.

6. Have a Back-Up Plan if Not Meeting Enrollment Targets

Despite the best-laid plans, you may have problems enrolling enough sites or participants. If that's the case, it's best to already have a back-up plan in place. First, try to determine the reasons why enrollment targets are not being met:

- Are the enrollment criteria so restrictive that there is a lack of eligible sites or participants, or are eligible sites or participants declining to participate?
- If the concern is due to eligibility, consider whether the criteria can be relaxed while still maintaining the integrity of the study design.

- If sites or participants are eligible but unwilling to participate, what are the reasons they are refusing? Can these concerns be addressed? For example, if the time of the sessions is not convenient, can additional sessions be offered at a different time?
- Do other burdens exist—real or perceived—that can be resolved?

With careful planning and by starting recruiting efforts early, you should have enough time to put the back-up plan in place, if needed. The back-up plan may include turning to other venues to recruit participants or using different recruiting methods.

B. How to Develop and Implement Standardized Procedures for Data Collection

We've all heard the saying, "garbage in, garbage out." To avoid this fate, you need to develop and implement standardized procedures for data collection. Two steps can help you standardize data collection and ensure consistency in data collection across sites, minimize bias, and yield high-quality data:

1. Prepare a data collection manual that clearly documents the data collection procedures.
2. Carefully train data collectors.

Working With Subcontractors

If a subcontractor conducts the data collection, both parties (the implementing agency [IA] and the subcontractor) need to understand and agree with the data collection protocol. When you set up a subcontract agreement:

- Include the protocol in the statement of work and as part of the contractual agreement.
- Monitor the performance of the subcontractor carefully to ensure that the data are being collected according to the protocol.
- Require that the subcontractor provide status reports in a timely manner (e.g., daily during the first several weeks of data collection, and then weekly thereafter) so that you can identify and quickly address any problems. This can help you avoid more serious problems down the road.

Preparing a Written Data Collection Manual

A good data collection manual

- describes the procedures for recruiting participants and for collecting and recording data at pre- and post-intervention,
- provides management and administrative details of the data collection process, and
- outlines systematic procedures for checking data quality to ensure that high-quality data are collected by all data collectors.

Ask key team members to review the manual to make sure that everyone agrees with the study procedures and then share it with all study staff members and data collectors. Exhibit V-2 identifies the topics to address in the manual. Other items may need to be added, depending on the study design and the mode of data collection.

Exhibit V-2.— Suggested Topics to Address in a Data Collection Manual

<table>
<tr>
<td>Data Collection Manual
Background
• Purpose of the study (background, goals, objectives, sponsor)
• Project management/organizational structure
• Communication protocols for study team and contact information
Recruitment
• Methods for establishing and maintaining rapport with site staff and satisfying logistical requirements
• Selection and recruitment of participants
• Strategies for gaining cooperation and avoiding refusals
Data Collection
• Data collection schedule
• Materials needed by data collectors (e.g., surveys, incentives, pens, consent forms)
• Informed consent procedures
• Potential challenges and solutions</td>
<td>Procedures
• Instructions for instrument administration
• Instructions for use of computer and/or other study-provided equipment
• Description of differences in data collection for the intervention and control/comparison sites and at pre- and post-intervention (if applicable)
• Procedures for payment of incentives (if provided)
• Employment-related administrative tasks
Respondents
• Participant retention procedures
• Strategies for maximizing valid responses and high response rates (and refusal conversion procedures, if being used)
• Safety concerns (for in-person data collection)
• Distressed respondent protocol (if applicable)
Data Delivery
• Data security and privacy procedures
• Data management and submission procedures</td>
</tr>
<tr>
<td colspan="2">Appendixes
• Copies of recruiting materials
• Copies of pre- and post-intervention instruments
• Copies of other survey materials (e.g., pre-notice letters)
• Copies of informed consent forms
• Receipt (if providing incentives)
• Forms for documenting data submission
• Answers to frequently asked questions (FAQs)
• Glossary of study-related terminology</td>
</tr>
</table>

Training Data Collectors

Investing time in training data collectors will help ensure that data are collected according to study protocols, procedures, and standards. Exhibit V-3 outlines steps for selecting, training, and monitoring data collectors.

Exhibit V-3.— Steps in Selecting, Training, and Monitoring Data Collectors

Step 1: Select the Data Collectors

- Develop a plan that includes the number and location of data collectors needed and the skill set required.
- Determine if data collectors will be current IA employees or site-based staff (e.g., classroom teachers), or if you will need to hire data collectors specifically for your study.
- When selecting data collectors, consider the following
  - Ability to follow the study protocol and work cooperatively with site-based staff.
  - Ability to establish rapport and engage respondents (special consideration may be needed if working with children, seniors, and/or racially and ethnically diverse populations).
  - Transportation requirements and requisite physical capabilities (e.g., ability to carry study equipment/materials).
  - If access to schools or childcare centers will be required, certain health screenings (e.g., tuberculosis) and enhanced criminal background checks may be necessary.
- If multilingual capabilities are needed, test both written and oral competencies of individuals.
- Once you have identified your staffing needs, select capable, reliable, and, if possible, experienced data collectors.
- If resources permit, consider whether the scope of the study warrants designating a data collector supervisor to monitor data collection daily. A suggested ratio for data collection is one full-time supervisor to every 14 data collectors.

Step 2: Plan the Training

- Determine the training mode that is most appropriate, most efficient, and cost-effective. The training session(s) may be conducted in person, by telephone, via the Web (e.g., Adobe Connect), through home study, or using a combination of these approaches.
- Develop a training plan in accordance with the data collection schedule and project resources. The plan should include details for handling logistics, such as date, time, and location of the training session(s). Be sure the training plan allows time for data collectors to become familiar with any software and hardware to be used for data collection.

Step 3: Develop Training Materials

- Develop documentation to train staff, which may include
  - a training agenda with scheduled breaks and meals;
  - a data collectors' manual;
  - an instructor training guide;
  - multiple training exercises, including mock interviews and survey administration;
  - an explanation of data security and confidentiality procedures and requirements;
  - home study training materials;
  - certification exams; and
  - PowerPoint slides (for lecture-based content).
- Include instructions for every aspect of data collection in your planning and materials, even mundane activities like stuffing envelopes.
  - Use an active learning approach to support differing learning styles; include participatory exercises and audiovisuals.
- Have appropriate staff (e.g., the study Principal Investigator [PI], evaluation manager) review and approve all training materials.
- If using bilingual data collectors, translate and thoroughly review all documentation before the training.
- Prepare materials specific to data collection supervisors that provide a clear understanding of their duties and responsibilities.

(continued)

Exhibit V-3.— Steps in Selecting, Training, and Monitoring Data Collectors (continued)

Step 4: Train the Data Collectors

- Limit the number of trainees to 15 to 20 people per training session to keep everyone engaged and to allow sufficient time for practice and certification. Schedule multiple training sessions if necessary.
- Each session should have at least one assistant trainer to help those who fall behind or have equipment problems. Try to switch out roles, even if just to lead an activity, to incorporate some variety and give the lead trainer an opportunity to circulate and see who might be having problems. If your training schedule is tight and/or the information to be covered is extensive, distribute the training manual 1 to 2 weeks in advance so trainees can review before the training.
- Have trainees demonstrate their knowledge of the most pertinent general concepts by completing a quiz and submitting it to the trainer for review before moving on to the next topic.
- If using multiple lead trainers, hold a "train-the-trainer" session with all training team members 2 to 4 weeks prior to data collector training to ensure consistency.
- Appendix C provides other general guidelines and useful suggestions for training data collectors.

Step 5: Certify the Data Collectors

- Do not authorize data collectors to begin work until they have demonstrated an acceptable level of proficiency with all study-related activities.
  - Employ written quizzes to test basic knowledge of the study (e.g., who is eligible to participate, who is sponsoring the study) and certain procedures (e.g., assignment of study IDs and transmittal of completed surveys).
  - Require each data collector to role-play with an instructor all the tasks necessary for the data collector job.
- Allow those who do not pass the first time an opportunity to repeat certification after some additional practice/review.

Step 6: Monitor the Data Collectors

- Delegate a person to monitor the data collectors, particularly during the first few weeks of data collection to identify and troubleshoot any problems or concerns.
- Establish weekly and cumulative production and response rate goals (e.g., X number of completed surveys by X date by study site).
- Develop a standardized report format to record actual production and response rates. Review reports daily and compare with goals. Closely monitor the cost of data collection as well to stay within budget.
- At least weekly, discuss production, caseloads, costs, data quality, weekly goals, schedule, and supply needs with data collectors.
- Consider scheduling a meeting with data collectors on a regular basis (e.g., biweekly) by teleconference or via a Web-enabled conference to provide updates, supplement training, and discuss lessons learned.
- Determine what actions are needed if field data collectors do not communicate regularly with their supervisor or do not meet production, cost, and quality requirements. Determine strategies for addressing nonresponse so plans can be implemented in a timely manner and not delay data collection.

For More Information on Developing and Implementing Standardized Procedures for Data Collection

- De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (Eds.) (2008). *International handbook of survey methodology*. New York, NY: Taylor & Francis.
- Dillman, D. A., Smythe, J. D., & Christian, M. L. (2009). *Internet, mail, and mixed mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: John Wiley & Sons.
- Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.
- Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Labowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley.

C. How to Ensure Quality Control During Data Collection

To ensure the quality of your data, build quality control into all aspects of the data collection process. This includes selecting, training, and supervising data collectors as well as designing the data collection procedures and materials.

Some General Tips

With a pre- and post-intervention design, procedures should be in place so that the data can be correctly linked and that participants receive the correct version of the materials (i.e., the intervention or control version). Although it's human to err, mistakes like forgetting to use IDs so that the data can be linked or giving the control group the questionnaire intended for the intervention group can be disastrous.

Here are a few steps you can take to ensure your data collection goes smoothly:

Whenever possible, **build in redundancies** so that human errors can be corrected and any loss of participant data is avoided.

- Make sure that case identification numbers are accurate and securely linked to the right person and his/her follow-up contact information. Use pre-assigned ID numbers and/or ID numbers preprinted on materials to help minimize assignment and transfer errors.
- For computer-assisted interviews and Web-based surveys, use programming logic and a case management system to ensure that appropriate versions of the survey are administered.
- Follow these additional steps to safeguard in-person data collection:
  - Use a color-coding system for surveys (e.g., yellow cover for intervention group and green cover for control group) to help prevent using the wrong version.
  - Implement a distinct range of case ID numbers for the intervention and control respondents (e.g., intervention IDs start with an "I" and control IDs start with a "C.").
  - Use color-coded folders to assist with the handling of materials to keep surveys safe and separate from consent forms and contact information.
- Use a step-by-step data collection checklist to reinforce protocol adherence.
- Use transmittal forms when shipping surveys to minimize omission and loss of forms and documentation.

Listed below are quality control procedures or activities that take place during and after data collection and are specific to different modes of data collection.

Quality Control for Telephone Data Collection

If available and resources permit, use a call center for telephone data collection. Call centers generally have established quality control procedures, like monitoring live or recorded interviews to confirm study protocol is followed by all interviewers. Key project team members can also monitor interviews to ensure that all interviewers have properly gained informed consent, are adequately and accurately responding to participant questions, and are reading all survey questions as written. As part of the monitoring process, evaluate an interviewer's tone, attitude, and ability to engage the respondent. Sharing feedback with interviewers will help to improve their performance and maintain desirable practices.

If you use in-house or site-based staff for telephone data collection, keep a record of calls for each respondent. Record the number of attempts, including date and time. If possible, record and review a proportion of the interviews to confirm that all interviewers are following the study protocol.

If recordings are not available, randomly select a proportion of respondents (5 to 10 percent) and call again to verify that the interview took place, key questions were asked, and the interviewer behaved in a professional manner. If you encounter any issues or concerns, randomly select another sample of respondents to call back.

Regardless of whether you use call center, in-house, or site-based staff, hold regular "quality circle" meetings in person, by teleconference, or via a Web-enabled conference to reinforce study protocol adherence, update progress, answer questions, and share lessons learned.

Quality Control for In-Person Data Collection

If your study relies on in-person data collection methods, you can observe interviews to confirm all interviewers are following the study protocol. The interviews can also be recorded and reviewed.

Similar to telephone data collection quality control, a random sample of respondents (5 to 10 percent) should be selected and verification calls made. Other methods of quality control can include routinely examining the data file for unusually short administration times (both overall and by section), unusually high numbers of "refused" or "don't know" responses, or other anomalies.

Quality Control for Self-Administered Hard Copy Surveys

If your study uses paper-and-pencil questionnaires, the data must be keyed into a data entry system or a spreadsheet template. Train keyers on the system, including how to

- enter data accurately and consistently (e.g., how to handle a range when a single number is required or multiple responses when only one is allowed);
- document any data anomalies; and
- identify respondents, groups, and sites.

All data should be keyed into the system twice. The two data sets should be compared for accuracies and/or discrepancies. If resources do not permit 100 percent verification, randomly select at least 20 percent of the completed surveys to verify the accuracy of the data entry.

If the level of discrepancies is acceptable (e.g., less than 5 to 10 percent), then additional verification is not necessary; however, if there are numerous discrepancies, then 100 percent verification is recommended; otherwise, the quality of the data is suspect.

Chapter VI • Addressing the Challenges of Conducting Effective Impact Evaluations: Analysis

A. How to Conduct an Attrition Analysis

What is attrition?

At the end of your study, you may be left asking yourself, "Where'd everyone go?" You conducted your sample size calculation and recruited enough participants to meet the statistical power needs of the analysis, but not everyone provided the follow-up data that you need.

Hopefully, you anticipated this loss and oversampled at baseline. Unfortunately, even with additional sampling at baseline, this "loss to follow-up" may threaten the representativeness of the sample, meaning that the sample may no longer be a good snapshot of the broader population. What's more, attrition may also lead to biased estimates of your program's impact.

Attrition or loss to follow-up occurs when participants assigned to one of your study conditions fail to complete the post-intervention survey.

Attrition is a greater concern for randomized trials, where the random assignment of individuals is designed to ensure that the people exposed to the SNAP-Ed program and those not exposed are similar in every way, except for the exposure. That said, it is important to assess the potential effect of attrition before conducting your impact analyses.

Expect to have some attrition with any evaluation. As a rule of thumb, minimal (i.e., less than 10% or so) attrition is generally not a problem.

Conducting an attrition analysis can uncover patterns and shed light on how attrition may affect the generalizability of your impact findings. In other words, attrition may hinder your ability to say that your findings are generally or universally applicable.

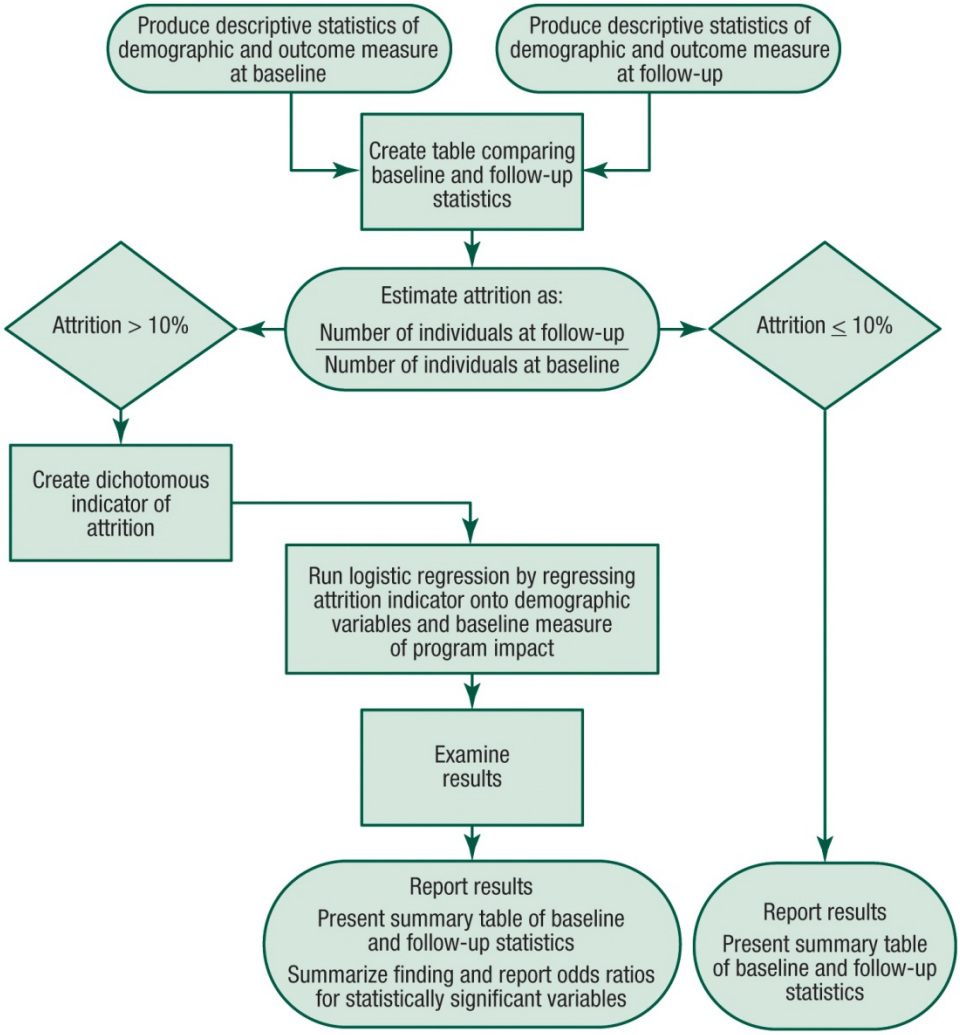
For example, in the analysis of a school-based SNAP-Ed program for early elementary school students, younger parents were less likely to provide data at post-intervention. The evaluators hypothesized that this may be because younger parents have more demands on their time or simply that younger parents are less willing to participate in an evaluation. Based on this result, evaluators concluded that the findings of their evaluation may not be universally true for younger parents.

Attrition analysis can also highlight potential biases in your impact findings. If attrition among younger parents is due to some factor that has nothing to do with nutrition behaviors, then attrition is unlikely to have influenced your measured impact. On the other hand, if attrition is indeed correlated or "linked" with the primary impact (e.g., younger parents are less likely to purchase fruit and vegetables), then it is possible that attrition may have influenced the measured impact.

How to Examine the Effect of Attrition

Exhibit VI-1 summarizes the recommended approach for how to examine attrition and, if necessary, how to perform an attrition analysis.

Exhibit VI-1.— Steps in Examining Attrition and Performing an Attrition Analysis



To begin, prepare a table with summary statistics of the study participants' demographic characteristics at baseline and again at follow-up. Next, estimate attrition as the proportion of respondents who provided follow-up data relative to the respondents who provided baseline data.

As a general rule of thumb, if attrition is less than 10 percent, the summary table should be sufficient. If attrition is greater than 10 percent, it's a good idea to run a simple logistic regression model whether or not respondent attrition is related to any of your measured demographic variables.

Keep in mind that the suggested 10 percent threshold is a rule of thumb and not a hard-and-fast criterion. You may also want to assess the effects of attrition at lower levels, especially as you observe systematic differences in your summary tables.

For this model, a dichotomous indicator—an indicator with two classifications—identifies participants as completers or noncompleters (at follow-up). It is regressed on the set of demographic variables, and regression coefficients represent the odds ratios of nonparticipation.

The following equation presents an example of a logistic regression equation that could be used to assess whether attrition is related to a person's sex, race/ethnicity, or reported income:

$$Y_i = \mu_0 + \beta_1 SEX + \beta_2 ETHNICITY + \ldots + \beta_k INCOME \quad .$$

In this model, $Y_i = 1$ for all respondents lost to follow-up (i.e., attrition) and $Y_i = 0$ for all respondents who provided data at the follow-up data collection period. The covariates *SEX*, *ETHINICITY*, and *INCOME* are the demographic variables collected at baseline. The ellipsis indicates that additional covariates can be added to the model. The model coefficients $\beta_{(1\ldots k)}$ express the relationship between each covariate and the likelihood of *not providing* data at follow-up with statistically significant results suggesting that one or more covariates are related to participant attrition.

It's important to remember that this is a logistic regression where Y_i is expressed as a logarithmic odds function $Y_i = \log\left(\frac{p_i}{1-p_1}\right)$

This information can be summarized to highlight demographic differences between those who provide data at follow-up and those who do not (see Appendix D for an example). These differences not only help you understand who benefited from your program and the limitations of your findings, but they may also help you improve your program's reach by showing you who is unwilling or unable to fully participate.

B. How to Conduct an Impact Analysis When Using a Clustered Research Design

Birds of a Feather …

A group randomized trial (GRT) assigns treatment to larger social units while assessing change on the individuals within those social units.

The old adage "birds of a feather flock together" describes the tendency of people with similar interests or characteristics to spend time together. This comment on human nature can pose a serious problem when you want to evaluate interventions or programs that have been assigned to groups of people who may be found in shared settings such as schools, workplaces, or community centers and not to groups of people (e.g., individuals). Keep in mind that social and economic facts of life, shared history, or common geography can lead to a social sorting in which people who are more similar to each other tend to be grouped together.

If your program takes place at the group level (e.g., school, community center) and you collect impact measures from individuals within those groups, your data are not independent.

Why? Because the same factors that led people to "flock together" will also lead them to respond more like other people in their group than like people in another group. Therefore, evaluation designs that

assign treatments or interventions to larger social units rather than individuals are referred to as group randomized trials (GRTs).

What's Different About Clustered Data?

1. Nonindependence. Most research designs are developed around individual assignment. These designs are referred to as clinical random trials (CRTs). In these designs, a group of *N* participants or individuals are assigned to one of two groups:

1. **intervention or treatment group**, where they receive the intervention as part of the treatment condition or
2. **comparison or control group**, where they do not receive an intervention.

Data are *independent and identically distributed (IID)* when:

- The data are randomly "scattered" around the mean value.
- The amount of "scatter" is generally uniform across the observations.
- The data consists of independent observations.

In practice, both GRTs and CRTs may involve nonrandomized, quasi-experimental approaches as well.

In a balanced design, half of the participants or $\left(N/2\right)$ are assigned to each condition. In fact, most common analytic approaches are based on the assumption that data are **independent and identically distributed** (this is often abbreviated as IID). Violating this assumption can threaten the integrity of your analysis, and when groups rather than individuals are assigned to a study condition and data are collected from participants within these groups, your assumption is almost *guaranteed* to be invalid.

Nonindependence will lead to correlated data. Under IID conditions, each person's response to an outcome of interest $\left(Y_i\right)$ can be expressed as a function of the mean response $\left(\mu_0\right)$ and some random error unique to the respondent $\left(e_i\right)$. If the outcome of interest is a continuous measure, it's common to assume that the errors are

- normally distributed (*N*),
- with a mean of 0, and
- vary around the mean by some known amount $\left(\sigma_m^2\right)$.

The subscript *i* indicates any person from the sample. This will often be expressed as a simple linear regression:

$$Y_i = \mu_0 + e_i, \quad e_i \sim N\left(0, \sigma_m^2\right).$$

However, when groups are assigned to conditions and data are collected from individuals within the groups, the individual's response ($Y_{i:j}$) is more complicated, as indicated by the subscript (*i:j*). The colon denotes nesting (i.e., one thing is uniquely located in another thing) and indicates that person *i* is a member of group *j*. Now each person's response is expressed as

- a function of the mean response $\left(\mu_{0:0}\right)$,
- a random effect $\left(u_{0:j}\right)$ that describes the effect of being a part of group *j*, and
- some random error unique to the respondent $\left(e_{i:j}\right)$.

The random effect associated with group membership has the same characteristics as the random error associated with each respondent. Specifically, it is assumed that the errors are normally distributed (*N*),

with a mean of 0 and that they vary around the mean by some known amount $\left(\sigma_g^2\right)$. This will often be expressed as a slightly more complicated mixed effect linear regression:

$$Y_{i:j} = \mu_{0:0} + u_{0:j} + e_{i:j}\text{, } e_{i:j} \sim N\left(0,\sigma_m^2\right) \text{ and } u_{0:j} \sim N\left(0,\sigma_g^2\right).$$

The main point to take away from this discussion is the fact that the term $u_{0:j}$ is a common term for all persons in group *j*, which highlights the lack of independence among individuals in the same group. Therefore, the larger σ_g^2 is relative to the total variation in the outcome measure, the greater the commonality among members.

2. Increased Variation. The total variation in the outcome of interest will be larger under the GRT than the CRT. The additional variation is a result of the group-level component of variance σ_g^2 and is indexed by the **intraclass correlation coefficient (ICC**). The ICC is defined as **the proportion of variance attributable to the group over and above the variation attributable to the individuals within the group and is expressed as a proportion:**

$$ICC = \frac{\sigma_g^2}{\sigma_m^2 + \sigma_g^2}.$$

As you'll recall, we defined σ_m^2 as the individual-level variation or the amount of variation among persons within the same group and σ_g^2 as the group-level variation or the amount of variation between groups. Additionally, we can define $\sigma_m^2 + \sigma_g^2 = \sigma_y^2$, where σ_y^2 is the total variation in the measured outcome.

The **ICC** is a measure of the degree to which **individuals within one group are "more similar" to each other** than they are to individuals in another group.

In practice, it is common to find larger ICCs among more tightly knit groups (e.g., families) and small ICCs among more diverse groups (e.g., communities). The magnitude of the ICC will also be influenced by the relationship between the measured outcome and the characteristics of the social group. For example, in school settings, it is common to find larger ICCs related to educational outcomes than to dietary outcomes. The Variance Inflation Factor (VIF) describes the impact of the ICC on total variation. The VIF is expressed as:

$$VIF = 1 + \left(m - 1\right) ICC,$$

where *m* represents the number of persons per group. Exhibit VI-2 demonstrates how even small ICCs can have a substantial influence on overall variation.

Exhibit VI-2.— Understanding Intraclass Correlation Coefficients (ICC) and Its Effect on Variation

No. Participants Per Group	ICC				
	0.01	0.03	0.05	0.07	0.09
50	1.49	2.47	3.45	4.43	5.41
75	1.74	3.22	4.7	6.18	7.66
100	1.99	3.97	5.95	7.93	9.91
125	2.24	4.72	7.20	9.68	12.16

As you can see in Exhibit VI-2, an ICC as small as 0.05 combined with groups of 50 participants leads to almost 3.5 times the variation in the GRT design as there would have been had this same study been conducted with individuals assigned to condition.

3. Fewer Degrees of Freedom. Statistical nonindependence also affects the effective sample size and, by extension, the degrees of freedom (*df*) for the test of the program's impact.

When individuals are assigned to a study condition, each person is independent, and *df* are determined by the number of individuals in each condition, specifically:

$$df = c(n-1),$$

where *c* indicates the number of study conditions and *n* indicates the number of persons per study condition.

However, under "clustered designs," the independence assumption has been violated, and *df* are based on the number of groups assigned to condition.

Accordingly, *df* for the test statistic are determined as $df = c(g-1)$, where *c* again represents the number of study conditions and *g* indicates the number of groups per condition.

How does clustering affect degrees of freedom (*df*)?

In a study with 100 participants randomly assigned to 1 of 2 study conditions (treatment vs. control), *df* would be calculated:

df = c(n-1) = 2(50-1) = 98

If those same 100 participants were nested in 10 schools and 5 schools were randomly assigned to each study condition, *df* would be calculated:

df = c(g-1) = 2(5-1) = 9

Quite a difference!

What Can You Do?

First of all, have no fear—you're not alone. Systematic reviews suggest that despite more than 30 years of warnings about the impact of clustering, many researchers and evaluators in many fields continue to ignore this phenomenon when they analyze data (Simpson et al., 1995; Eldridge, Ashby et al., 2004; Murray et al., 2004).

If you understand the issues, you can take appropriate steps to address them. It is important to keep in mind that you should not use standard Analysis of Variance (ANOVA)-based (i.e., General Linear Model, or GLM) analyses if you have a lack of independence, increased variation, and limited *df*. These methods are likely to underestimate overall variance and overestimate the available *df*, which will lead to an inflated Type I error rate (i.e., assuming statistically significant differences when they do not really exist).

A two-stage ANOVA-based approach may be appropriate. This approach is a simple extension of the GLM that estimates group means in the first step and then uses those group means in a second GLM analysis to estimate program impacts across conditions.

Analyzing correlated data can be challenging. If you need to do this, work with an experienced statistician or social scientist.

However, two-stage approaches have generally been replaced by the use of more sophisticated mixed-effect models, or hierarchical linear models (HLMs). Methods that rely on robust variance estimation, or Generalized Estimating Equations (GEE), may also be used; however, these approaches are not recommended for small studies (fewer than 40 groups).

Classical statistical texts such as Snedecor and Cochran (1989) and Rosner (2000) provide an overview of the issues and methodology. More advanced texts like Murray (1998) and Bryk and Raudenbush (1992) provide more detail on statistical programs and advanced regression modeling techniques. Analytic methods that account for the clustered and correlated nature of the data are available in common statistical software packages like SAS, Stata, and SPSS. Once you have a basic understanding of regression modeling and time to explore the documentation of the appropriate software program (e.g., PROC MIXED or GLIMMIX in SAS), you'll be able to begin running basic HLM models.

If you don't have experience with regression modeling or if you need more complex modeling for your evaluation, you may find it helpful to consult a statistician or a social scientist with appropriate quantitative training.

For More Information on Conducting an Impact Analysis When Using a Clustered Research Design

- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.
- Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R., & Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clinical Trials, 1*(1), 80–90.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.
- Murray, D. M., Varnell, S. P., & Blitstein, J. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health, 94*(3), 423–432.
- Rosner, B. (2000). *Fundamentals of biostatistics*. Pacific Grove, CA: Duxbury.
- Simpson, J. M., Klar, N., & Donnor, A. (1995). Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health, 85*(10), 1378–1383.
- Snedecor, G. W., & Cochran, W. M. (1989). *Statistical methods*. Ames, IA: The Iowa State University Press.

Chapter VII • Conclusions

SNAP-Ed Guiding Principles call on implementing agencies (IAs) to conduct impact evaluations of funded SNAP-Ed programs to assess whether the programs are meaningful for their specific target audience(s), implemented as designed or modified with justification, and have the anticipated impact on nutrition behaviors (USDA, 2013). Conducting rigorous evaluations of SNAP-Ed programs can be a key resource in determining whether programs are achieving their intended outcomes of improving participants' nutrition and other behaviors. Evaluations can also help policymakers and program administrators strengthen the design and operation of programs and demonstrate the benefits of the program to the funding source.

Based on the assessment of the Wave I and II self-evaluations (Gabor et al., 2012a; Long et al., 2013), as well as considering the types of resources and staff typically available to SNAP-Ed IAs, this guidebook provides guidance to IAs on how to conduct rigorous evaluations of SNAP-Ed programs. This guidebook offers step-by-step procedures to address many of the challenges faced by program administrators and evaluators starting with guidelines on planning and developing an evaluation, through careful design and implementation of the data collection to collect information on key outcomes, to appropriate techniques to analyze and report the results of the impact analysis. This guidance has been tailored to meet the needs of SNAP-Ed program administrators and evaluators. Throughout the guidebook examples are provided to illustrate key points, as well as lists of resources for those seeking additional information on a specific topic.

Recommended Practices for Impact Evaluation

- Build on available research.
- Determine the anticipated size of the program impact on the target audience before conducting the intervention.
- Use a comparison or control group and, to the extent possible, randomly assign units to either the treatment or comparison/control group.
- Conduct a power analysis to determine the minimum sample size needed for the evaluation study.
- Choose impact measures that fit the intervention and that approach existing standards for credible assessment.
- Use existing instruments/survey questions that are demonstrated to be valid and reliable and are sensitive to change, and conduct pretests of newly developed questions.
- Establish standardized procedures for data collection and quality control and observe standards for the fair treatment of study participants.
- Match the analytic strategies to the characteristics of the evaluation design.
- Share results to maximize their value. Report both positive and negative results, but do so accurately.

Sources: Long, V., Cates, S., Blitstein, J., Deehy, K., Williams, P., Morgan, R., Fantacone, J., Kosa, K., Bell, L., & Hersey, J. (2013). *Supplemental Nutrition Assistance Program education and evaluation study (Wave II)*. Retrieved from the U.S. Department of Agriculture, Food and Nutrition Service website: www.fns.usda.gov/research-and-analysis

U.S. Department of Agriculture, Food and Nutrition Service. (2005). *Nutrition education: Principles of sound impact evaluation.* Retrieved from http://www.fns.usda.gov/nutrition-education-principles-sound-impact-evaluation

For some IAs, the lack of funds and expertise on the part of local project staff and subcontractors may be a barrier to conducting rigorous impact evaluations. Thus, if feasible, some IAs may need to secure additional funding (e.g., joint State funding or grant funding) or partner with evaluators or statisticians at

a local university to conduct a rigorous impact evaluation. Keep in mind that other program administrators and evaluators may be facing the same challenges, so consult with your colleagues and share lessons learned, both successes and failures. IAs' continued efforts to conduct rigorous evaluations of SNAP-Ed programs will help build the research basis for evidence-based, behaviorally focused interventions to improve the nutrition behaviors of SNAP participants and low-income individuals eligible to receive SNAP and other means-tested program benefits.

References

Bamberger, M., J. Rugh, Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time, and data constraints. *American Journal of Evaluation, 25*(1), 5–37.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (2nd ed.).Thousand Oaks, CA: Sage Publications, Inc.

Chen, H. (2005). *Practical program evaluation*. Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ, Lawrence Erlbaum Assoc.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*, 283–284.

Contento, I. (2011). Nutrition education: Linking research, theory, and practice (2nd ed.) Sudbury, MA: Jones & Bartlett Publishers, LLC.

Cullen, K. W., Baranowski, T., Owens, E., Marsh, T., Rittenberry, L., & de Moor, C. (2003). Availability, accessibility, and preferences for fruit, 100% fruit juice, and vegetables influence children's dietary behavior. *Health Education and Behavior, 30*(5), 615–626.

Dave, J. M., Evans, A. E., Pfeiffer, K. A., Watkins, K. W., & Saunders, R. P. (2010). Correlates of availability and accessibility of fruits and vegetables in homes of low-income Hispanic families. *Health Education Research, 25*(1), 97–108.

Economos, C. D., Sacheck, J. M., Ho Chui, K. K., Irizzary, L., Guillemont, J., Collins, J. J., & Hyatt, R. R. (2008). School-based behavioral assessment tools are reliable and valid for measurement of fruit and vegetable intake, physical activity, and television viewing in young children. *Journal of the American Dietetic Association*, *108*(4), 695–701.

Edmunds, L. D., & Ziebland, S. (2002). Development and validation of the Day in the Life Questionnaire (DILQ) as a measure of fruit and vegetables questionnaire for 7-9 years old. *Health Education and Behavior, 17*(2), 211–220.

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R., & Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clinical Trials, 1*(1), 80–90.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*(7), 513–516, 575–578.

Gabor, V., Cates, S., Gleason, S., Long, V., Aponte Clarke, G., Blitstein, J., Williams, P., Bell, L., Hersey, J., Ball, M., & Olson, S. (2012a). *SNAP education and evaluation (Wave I): Final report*. Retrieved from the U.S. Department of Agriculture, Food and Nutrition Service, Office of Research and Analysis website: www.fns.usda.gov/research-and-analysis

Gabor, V., Williams, P., Cates, S., Blitstein, J., Bell, L., & Hersey, J. (2012b). *SNAP education and evaluation case study report: New York State Department of Health's Eat Well Play Hard in Child Care Settings Program.* Retrieved from www.fns.usda.gov/research-and-analysis

Green, L. W., Kreuter, M. W., Deeds, S. G., & Partridge, K. B. (1980). *Health education planning: A diagnostic approach*. Palo Alto, CA: Mayfield Publishing Co. Gunning, R. (1968). *The technique of clear writing* (Rev. ed.). New York, NY: McGraw-Hill**.**

Hersey, J. C., Niederdeppe, J., Ng, S., Mowery, P., & Farrelly, M. C. (2005). How state counter-industry campaigns help prime perceptions of tobacco industry practices to promote reductions in youth smoking. *Tobacco Control, 14*, 377–383.

Holden, D. J., & Zimmerman, M. A. (2009). *A practical guide to program evaluation*. Los Angeles, CA: Sage.

Iversen, G. R., & Norpoth, H. (1976). *Analysis of variance*. Beverly Hills, CA: Sage Publications.

Jamelske, E., Bica, L. A., McCarty, D. J., & Meinen, A. (2008). Preliminary findings from an evaluation of the USDA fresh fruit and vegetable program in Wisconsin schools. *Wisconsin Medical Journal, 107*(5), 225–230.

Jaramillo, S. J., Yang, S.-J., Hughes, S. O., Fisher, J. O., Morales, M., & Nicklas, T. A. (2006). Interactive computerized fruit and vegetable preference measure for African-American and Hispanic preschoolers. *Journal of Nutrition Education Behavior, 38*(6), 352–359.

Kincaid, J. P., Fishburne Jr., R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel* (No. RBR-8-75). Millington, TN: Naval Technical Training Command, Research Branch.

Knai, C., Pomerleau, J., Lock, K., & McKee, M. (2006). Getting children to eat more fruit and vegetables: A systematic review. *Preventive Medicine*, *42*(2), 85–95.

Litwin, M. S. (1995). *How to measure survey reliability and validity*. Thousand Oaks, CA: Sage Publications, Inc.

Long, V., Cates, S., Blitstein, J., Deehy, K., Williams, P., Morgan, R., Fantacone, J., Kosa, K., Bell, L., & Hersey, J. (2013). *Supplemental Nutrition Assistance Program education and evaluation study (Wave II)*. Retrieved from the U.S. Department of Agriculture, Food and Nutrition Service website: www.fns.usda.gov/research-and-analysis

Marsh, T., Cullen, K. W., & Baranowski, T. (2003). Validation of a fruit, juice, and vegetable availability questionnaire. *Journal of Nutrition Education Behavior, 35*(2), 100–104.

Meyer, P. J. (2003). *What would you do if you knew you couldn't fail? Creating S.M.A.R.T. Goals. Attitude is everything: If you want to succeed above and beyond*. Meyer Resource Group, Incorporated.

Moore, G. F., Tapper, K., Murphy, S., Clark, R., Lynch, R., & Moore, L. (2007). Validation of a self-completion measure of breakfast foods, snacks and fruits and vegetables consumed by 9- to 11-year-old schoolchildren. *European Journal of Clinical Nutrition*, *61*(3), 420–430.

Moore, L., Tapper, K., Dennehy, A., & Cooper, A. (2005). Development and testing of a computerised 24-h recall questionnaire measuring fruit and snack consumption among 9-11 year olds. *European Journal of Clinical Nutrition, 59*, 809–816.

Mullen, P. D., Hersey, J. C., & Iverson, D. C. (1987). Health behavior models compared. *Social Science and Medicine*, *24*(11), 973–981.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Murray, D. M., Varnell, S. P., & Blitstein, J. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health, 94*(3), 423–432.

Ontai, L., Ritchie, L., Williams, S. T., Young, T., & Townsend, M. S. (2009). Guiding family-based obesity prevention efforts in children, Part 1: What determinants do we target? *International Journal of Child and Adolescent Health*, *2*(1), 19–30.

Posavac, E. J. (2010). *Program evaluation: Methods and case studies* (8th ed.). Saddle River, NJ: Pearson Education, Inc.

Posavac, E. J., & Carey, R. G. (1997). *Program evaluation methods and case studies* (5th ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Rosenthal, R. & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York, NY: McGraw-Hill, Inc.

Rosner, B. (2000). *Fundementals of biostatistics*. Pacific Grove, CA: Duxbury.

Senter, R. J., & Smith, E. A. (1967, November). *Automated readability index*, AMRI.-TR-66220. Wright-Patterson Air Force Base, OH.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Co.

Simpson, J. M., Klar, N., & Donnor, A. (1995). Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health, 85*(10), 1378–1383.

Smith, T. (2002, November 14–17). *Developing cross-national survey instruments.* Paper presented at the meetings of the International Conference on Questionnaire Development, Evaluation, and Testing Methods, Charleston, SC.

Snedecor, G. W., & Cochran, W. M. (1989). *Statistical methods*. Ames, IA: The Iowa State University Press.

Snyder, L. B., Hamilton, M. A., Mitchell, E. W., Kiwanyka-Tondo, J., Fleming-Milici, F., & Proctor, D. (2004). A meta-analysis of the effect of mediated health communication campaigns on behavior change in the United States. *Journal of Health Communications, 9*(Suppl 1), 71–96.

Thomson, C. A., & Ravia, J. (2011). A systematic review of behavioral interventions to promote intake of fruit and vegetables. *Journal of the American Dietetic Association, 111*(10), 1523–1535.

Townsend, M. S., & Kaiser, L. L. (2005). Development of a tool to assess psychosocial indicators of fruit and vegetable intake for 2 federal programs. *Journal of Nutrition Education and Behavior*, *37*(4), 170–184.

Townsend, M. S., & Kaiser, L. L. (2007). Brief psychosocial fruit and vegetable tool is sensitive for the US Department of Agriculture's Nutrition Education Programs. *Journal of the American Dietetic Association*, *107*(12), 2120–2124.

Townsend, M. S., Kaiser, L. L., Allen, L., Joy, A., & Murphy, S. (2003). Selecting items for a food behavior checklist for a limited-resource audience. *Journal of Nutrition Education and Behavior, 35*(2), 69–82.

Townsend, M., Sylva, K., Martin, A., Metz, D., & Wooten-Swanson, P. (2008). Improving readability of an evaluation tool for low-income clients using visual information processing theories. *Journal of Nutrition Education and Behavior*, *40*(3), 181–186.

Townsend, M. S., Young, T., Ontai, L., Ritchie, L., & Williams, S. T. (2009). Guiding family-based obesity prevention efforts in children, Part 2: What behaviors do we measure? *International Journal of Child and Adolescent Health*, *2*(1), 31–48.

U.S. Department of Agriculture, Food and Nutrition Service. (2005). *Nutrition education: Principles of sound impact evaluation.* Retrieved from http://www.fns.usda.gov/nutrition-education-principles-sound-impact-evaluation

U.S. Department of Agriculture, Food and Nutrition Service. (2013). *Supplemental Nutrition Assistance Program Education guidance FY 2013.* Retrieved from http://snap.nal.usda.gov/snap/Guidance/GuidingPrinciples2014.pdf

U.S. Government Accountablity Office. (2012). *Designing evaluations: 2012 revision*. Washington, DC: GAO. Retrieved from http://gao.gov/assets/590/588146.pdf

Weidmer, B. (1994, May 11–15). *Issues and guidelines for translation in cross-cultural research*. Paper presented at the 49th Annual Conference of the American Association for Public Opinion Research, Danvers, MA, and published in *Proceedings of the Survey Research Methods Section, American Statistical Association 1994.* Retrieved from http://www.amstat.org/sections/srms/Proceedings/papers/1994_215.pdf

Willis, G. B. (2005). *Cognitive interviewing. A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications, Inc.

Winer, B. J., Brown, D. R., & Michels, K. M.. (1991). *Statistical principles in experimental design*. Boston, MA: McGraw-Hill.

World Health Organization (WHO). (no date). *Process of translation and adaptation of instruments*. Retrieved from http://www.who.int/substance_abuse/research_tools/translation/en/

Glossary

ANOVA or analysis of variance: a statistical method that determines whether a significant relationship exists between variables.

Attrition: occurs when a participant fails to complete a program and/or the post-intervention survey.

Clinical random trial: an experimental design where a group of N participants is either assigned to receive an intervention as part of the treatment condition or not receive an intervention as a comparison condition.

Cluster design: an experimental design in which research subjects are not sampled or selected independently, but in a group.

Cognitive interviewing: a pretesting approach that examines respondents' thought processes regarding their answers to survey questions.

Comparison group: the group of participants who do not receive the treatment/intervention in a quasi-experimental design.

Concurrent validity: a type of criterion validity that measures how well the item or scale correlates with the "gold standard" measure of the same variable; for example, for measuring dietary intake the "gold standard" is 24-hour dietary recalls.

Construct validity: a theoretical measure of how meaningful a survey instrument is, and is usually determined after years of experience by numerous investigators (more like hypothesis testing than calculation of correlation coefficients).

Content validity: a formal expert review of how good an item or series of items appears, usually assessed by individuals who are experts in the subject matter of interest.

Continuous measure: a measure that can have an infinite amount of values.

Control group: the group of participants who do not receive the treatment/intervention in a randomized design.

Convergent validity: a type of construct validity that proves that measures that theoretically should be related to each other are in fact observed to be related to each other.

Correlation coefficients (r-values): measures the degree of correlation between two data points; r-values are considered good if they are equal to or exceed 0.70.

Counterfactual: refers to the control/comparison condition in an evaluation design; it is the state of affairs (e.g., the measured outcomes) in the absence of the intervention.

Covariate: the independent variable that researchers have control over or can choose or manipulate.

Criterion validity: a measure of the effectiveness of a variable(s) to predict an outcome measure.

Cronbach coefficients (α-values): measures internal consistencies; α-values are considered good if they are equal to or exceed 0.70.

Degree of freedom: the number of values in the final calculation that are free to vary.

Dichotomous measure: a measure that can only have two responses (e.g., gender or yes/no).

Difference-in-difference estimation: a statistical method that measures the effect of a treatment at a given point in time.

Dosage: measures participants' exposure to the intervention, that is, the amount of the program received.

External validity: extent to which study results can be generalized to other populations or settings.

Face validity: a casual review of how good an item or group of items appears as assessed by individuals with no formal training in the subject matter of interest.

Fidelity: how closely an intervention was implemented as designed.

Formative research: an application of qualitative and quantitative methods to gather data useful for the development and implementation of intervention programs.

Group Randomized Trial (GRT): a research design where the program is randomly assigned to the school or community center with observations of key outcomes occurring at the individual level.

Impact evaluation: measures the net change in outcomes for a particular group of people that can be attributed to a specific program.

Impact measures: used to measure the overall impact of an intervention or program.

Intercept survey: face-to-face surveys that are typically carried out in a space where there is a good population of the target audience. The strategic advantages of intercept surveys are the speed in which they can be conducted, their low cost, and the ability to poll a large number of consumers.

Internal consistency reliability: indicates how well a group of items in a scale measures the same characteristic or concept.

Internal validity: refers to the strength of the conclusion that can be drawn about the relationship between the program and its effects on the outcomes of measurement.

Interobserver or inter-rater reliability: indicates how well two or more interviewers agree in their assessment of a variable; should be used whenever there is a subjective component in the measurement of an external variable.

Intraclass correlation coefficient (ICC): the proportion of variance attributable to the group over and above the variation attributable to the individuals within the group.

Logic model: a flow diagram that outlines the program's inputs, activities or processes, outputs, and outcomes.

Measurement error: the difference between the actual value and the value obtained by the measurement.

Meta-analysis: statistical technique in which the results of two or more studies are mathematically combined to improve the reliability of the results.

Nonresponse: missing or incomplete data because respondents did not answer questions within the survey or declined to complete the survey outright.

Null hypothesis: refers to a general or default position; that there is no relationship between two measured phenomena or there is no effect.

Odds ratio: a statistical method to quantify how strongly the having or not having of X is associated with having or not having of Y in that population.

Outcome assessment: an examination of the extent to which an intervention program achieves its stated goals.

Outcome measures: used to measure the success of an intervention or program; they are the benefits or changes for individuals or populations during or after participating in an intervention or program. Outcomes may relate to behavior, skills, knowledge, attitude, values, condition, or other attributes.

Power analysis: determines the minimum sample size needed for an evaluation study.

Predictive validity: a type of criterion validity that measures how well the item or scale predicts expected future observations.

Quasi-experimental design: an experimental design that lacks random assignment into a treatment or comparison group.

Random error: caused by any factors that randomly affect measurement of the variable across the sample.

Randomized experimental design: uses controlled application of an intervention or treatment and randomization to provide evidence of the missing counterfactual and support claims of causal inference.

Reach: measures the proportion of the target population that participated in the intervention.

Regression discontinuity: the assignment to the intervention is based on a cutoff score on a measured covariate. The effect of the intervention is measured as the discontinuity between treatment and control regression lines at the cutoff.

Reliability: the extent to which an instrument or survey item(s) produces the same result when applied to the same person under the same conditions.

Response bias: occurs in statistical surveys if the answers of respondents differ from the potential answers of those who did not answer.

Response rate: proportion of eligible participants in a sample who responded to the survey.

Sample size: number of observations in a statistical sample.

Selection bias: occurs when individuals are more likely to take part in a research study than others.

Sensitivity to change: capability of an instrument and/or survey item(s) to measure change statistically.

Social marketing: seeks to influence social attitudes and/or behaviors among the target audience using basic marketing principles.

Test-retest reliability: an indicator of the stability of responses over time that is measured by having the same set of respondents complete the survey (a single item or group of items or scale within an instrument) at two different time points.

Type I error: probability of rejecting the null hypothesis when it is true. A Type I error rate of $\alpha = 5\%$ is desired, but higher Type I error rates are acceptable depending on the risk associated with claiming that statistical differences are significant when they are not.

Type II error: probability of retaining the null hypothesis when it is in fact false. The Type II error rate determines statistical power and vice versa. Power $(1-\beta)$ is the probability of rejecting the null hypothesis when it is false. By convention, researchers strive for $\beta \geq 80\%$. Lower levels of power lead to Type II errors that are at an unacceptable rate.

Validity: extent to which an instrument or survey item(s) measures what it is supposed to measure.

Variance: measures how individual data points vary from the average value of the data set.

Variance inflation factor: measure of collinearity (variables are highly correlated) in multiple regression.

.

Appendix A: Dietary and Nutrition Behavior Instruments for Use With Low-Literacy and Limited-Resource Audiences

Outcome Measures	Instrument	Study Population(s)	Mode(s) of Data Collection	Reliability	Validity	Sensitivity to Change	URL Link (if available)
Cups of fruits, vegetables, and fruits and vegetables consumed by child on a typical day Child ate variety of fruits each day Child ate variety of vegetables each day	Food Stamp Program Fruit and Vegetable Checklist (Townsend et al., 2003) University of California Cooperative Extension Food Behavior Checklist (Townsend et al., 2008)	Low-income women	Self-administered, self-administered in group setting, and interviewer administered individually and in groups	The internal consistency for the 7-item fruit and vegetable subscale was high (α = 0.80)	The 7-item fruit and vegetable subscale showed a significant correlation with serum carotenoid values (r = 0.44, $p < 0.001$), indicating acceptable criterion validity, and showed significant correlation with dietary variables	Demonstrated sensitivity to change for items expected to change as a result of the study intervention	http://townsendlab.ucdavis.edu http://townsendlab.ucdavis.edu
Willingness of child to try new fruits Willingness of child to try new vegetables	Willingness to try new fruits and vegetables (Jamelske, Bica, McCarty, & Meinen, 2008)	4th, 7th, and 9th graders	Self-administered	Not reported	Not reported	Compared with controls, intervention participants reported an increased willingness to try new fruits and vegetables at school ($p < 0.01$)	
Availability of fruits and vegetables at home during past week	Fruit, juice, and vegetable availability questionnaire (Marsh, Cullen, & Baranowski, 2003; Cullen et al., 2003)	Parents of 4th and 6th graders	Self-administered and interviewer administered via telephone	The internal consistencies for the fruit and vegetable availability items were high	There was significant agreement between self-reported and observed at-home availability for all fruit juices and most fruits and vegetables	Fruit, juice, and vegetable availability was a significant predictor of child fruit, juice, and vegetable consumption ($p < 0.05$)	

Outcome Measures	Instrument	Study Population(s)	Mode(s) of Data Collection	Reliability	Validity	Sensitivity to Change	URL Link (if available)
Consumption of fruits, vegetables, and milk Amount of physical activity, sleep, and TV watching Parental behaviors	Healthy Kids Tool (Ontai et al., 2009; Townsend et al., 2009)	Low-income parents/ caregivers of children aged 3–5 years	Interviewer-administered	Not reported	Not reported	No intervention	http://townsendlab.ucdavis.edu
Mediators of fruit and vegetable behavior changes, including self-efficacy	Fruit and Vegetable Inventory (Townsend & Kaiser, 2005 & 2007)	Low-income women	Self-administered in group	The internal consistency for the 7-item self-efficacy subscale was moderately high (α = 0.77) and test-retest coefficients were all significant	Compared to serum carotenoid, the 7 self-efficacy items had a correlation equal to 0.18 ($p < 0.10$). Construct was not significant compared to dietary recall and HEI	Controlling for energy intake, the change scores for were correlated with reported changes in fruit and vegetable behaviors ($r = 0.28$, $p = 0.01$)	http://townsendlab.ucdavis.edu
Fruit and vegetable intake, physical activity, and TV viewing	Shape Up Somerville Dietary Intake Survey (Economos et al., 2008)	Low-income children aged 6–9 years	Interviewer-administered in person	Test-retest reliability was very high for fruit and vegetable recall	There was significant agreement between self-reported and observed fruit and vegetable intake.	No intervention	http://www.childreninbalance.org

Outcome Measures	Instrument	Study Population(s)	Mode(s) of Data Collection	Reliability	Validity	Sensitivity to Change	URL Link (if available)
Children's fruit and vegetable preference and parental practices to encourage fruit and vegetable intake	Home Nutrition Questionnaire (Dave et al., 2010)	Low-income Hispanic parents of children aged 5–12 years	Self-administered	Reliability coefficients for children's preference, parental practices were both moderately high	Not reported	No intervention	See published article
Consumption of fruits, vegetables, milk, and other foods	The Day in the Life Questionnaire (DILQ) (Edmunds & Ziebland, 2002; Moore et al., 2005 & 2007)	Low-income children aged 9–11 years	Self-administered, with questions read to participants	Test-retest reliability was moderate for fruit and vegetable recall	Compared to observations, reported fruit and vegetable intake approached 70% agreement. Compared to dietary recalls, the computerized questionnaire gave higher estimates of fruit intake (k = 0.29). Compared to dietary interview, Spearman rank coefficients were 0.39 for fruit and 0.41 for vegetables	Demonstrated changes in mean daily fruit consumption (p = 0.04) and mean fruit consumption at morning break ($p \leq 0.000$)	http://www.performwell.org/index.php/find-surveyassessments/outcomes/health-a-safety/good-health-habits/day-in-the-life-questionnaire-ages-9-11

Appendix B: Example of a Debriefing Guide for Cognitive Interviews

INTRODUCTION

Thank you for meeting with us. We are interviewing parents who have a child who will be in the first, second, or third grade this coming school year. We are conducting the interviews with a small number of parents in order to test a survey instrument we will use next year with a much larger number of parents. This survey is part of a study sponsored by the U.S. Department of Agriculture's Food and Nutrition Service.

Today, I will ask you to fill out a survey for your child who will be in the first, second, or third grade this coming school year. If you have more than one child who will be in one of these grades, please fill out the survey for your youngest child. When you are finished filling out the survey, we will go back through the survey together to discuss any problems you may have had answering the questions, any words you may have found difficult to understand, or any suggestions you have for improving the survey. All of the information you provide will be kept completely confidential. Your name will not be associated with your responses.

RESPONDENT COMPLETES QUESTIONNAIRE

Start time: _____ End time: _____

***Note any questions that the respondent has difficulty answering and address those questions during the debriefing.

AFTER QUESTIONNAIRE IS COMPLETED

- ***On a scale of 1 to 5, where "1" is very difficult and "5" is very easy, how would you rate this survey? Why?***
- ***On a scale of 1 to 5, where "1" is very boring and "5" is very interesting, how would you rate this survey? Why?***
- ***What did you think about the length of the survey—was it too long, too short, or about right?***
- ***Do you remember any specific words or questions that were confusing? If yes, what were they?***
- ***Any general comments on the survey before we go back through it question by question?***

FEEDBACK

GENERAL COGNITIVE TESTING PROBES

- ***In your own words, what is this question asking?***
- ***What parts of the question were difficult to answer?***
- ***How easy or difficult is it to use the response scale?***
  - ***Were you able to find a response that fit your opinion?***
  - ***What other response options should be included?***
- ***What else could we do to improve this question?***

FEEDBACK

Questions on Whether Certain Foods Are Available at Home

1. Were any of these foods available in your home during the past week? Include fresh, frozen, canned, and dried foods. *(Circle Yes or No for each food.)*

a. Bananas	Yes	No
b. Apples	Yes	No
c. Grapes	Yes	No
d. Raisins	Yes	No
e. Pears	Yes	No
f. Celery	Yes	No
g. Carrots	Yes	No
h. Cucumbers	Yes	No
i. Broccoli	Yes	No
j. Zucchini	Yes	No
k. Potato chips, nacho chips, or corn chips	Yes	No
l. Regular soft drinks or sodas	Yes	No

QUESTION 1

- ***How easy or hard was it to recall whether these foods were in your home during the past week?***

FEEDBACK

Questions on the Fruits and Vegetables Your Child Eats

For the next questions, think about what your child ate during the past week, or the past 7 days. Do not include school, day care, or before/after school care.

2. How many days during the past week did your child eat more than one kind of fruit each day? Do not include fruit juice. *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 2

- ***How did you come up with your answer?***
- ***How easy or difficult is it to use the response scale?***
- ***Were you able to find a response that fit your answer?***
- ***How easy or hard was it to remember if your child ate more than one kind of fruit each day during the past week. Probe – difficulty in thinking about was done each day over the past week.***
- ***Did you consider food that was eaten outside the home (e.g., summer camp)?***

FEEDBACK

3. During the past week, how many cups of fruit did your child eat each day? Do not include fruit juice. *(Circle one.)*

    1. None
    2. ½ cup
    3. 1 cup
    4. 1 ½ cups
    5. 2 cups
    6. 2 ½ cups
    7. 3 cups or more

None 1 cup 2 cups 3 cups

QUESTION 3

- ***How did you come up with your answer?***
- ***How easy or difficult is it to use the response scale?***
- ***Were you able to find a response that fit your answer?***
- ***How easy or hard was it to remember how many cups of fruit your child ate each day in the past week?***
- ***How did you use the pictures to answer this question?***
- ***Did you consider food that was eaten away from home (e.g., summer camp)?***

FEEDBACK

4. How many days during the past week did your child eat more than one kind of vegetable each day? Do not include vegetable juice, French fries, or white potatoes. *(Circle one.)*

   1. None
   2. 1 to 2 days
   3. 3 to 4 days
   4. 5 to 6 days
   5. Every day

QUESTION 4

- ***How did you come up with your answer?***
- ***How easy or hard was it to remember if your child ate more than one kind of vegetable each day?***

FEEDBACK

5. During the past week, how many cups of vegetables did your child eat each day? Do not include vegetable juice, French fries, or white potatoes. *(Circle one.)*

   1. None
   2. ½ cup
   3. 1 cup
   4. 1 ½ cups
   5. 2 cups
   6. 2 ½ cups
   7. 3 cups or more

None 1 cup 2 cups 3 cups

QUESTION 5

- ***How did you come up with your answer?***
- ***How easy or hard was it to remember how many cups of vegetables your child ate each day in the past week?***
- ***Did you consider food that was eaten away from home (e.g., summer camp)?***

FEEDBACK

6. During the past week, did your child eat any meals or snacks that were provided by his/her school, before school care program, after school care program, or day care? *(Circle all that apply.)*

    1. No, did not eat breakfast, lunch, or snacks provided by school, before or after school care program, or day care
    2. Yes, breakfast
    3. Yes, lunch
    4. Yes, snacks

QUESTION 6

- ***How did you come up with your answer?***
- ***If it was the school year, how would you have answered the question?***

FEEDBACK

7. Is your child willing to try a new kind of fruit? Do not include fruit juice. *(Circle one.)*

    1. No
    2. Maybe
    3. Yes

QUESTION 7

- ***How did you come up with your answer?***
- ***How easy or difficult is it to use the response scale?***
- ***Were you able to find a response that fit your answer?***

FEEDBACK

8. How many days during the past week did you give your child fruit for a snack? Do not include fruit juice. *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 8

- ***How did you come up with your answer?***
- ***What did you include as snack time?***

FEEDBACK

9. How many days during the past week did you give your child fruit at dinner? Do not include fruit juice. *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 9

- ***How did you come up with your answer?***
- ***How easy or hard was it to answer this question?***

FEEDBACK

10. Is your child willing to try a new kind of vegetable? *(Circle one.)*

    1. No
    2. Maybe
    3. Yes

QUESTION 10

- ***How did you come up with your answer?***

FEEDBACK

11. How many days during the past week did you give your child a vegetable for a snack? Do not include vegetable juice, French fries, or white potatoes. *(Circle one*.)

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 11

- ***How did you come up with your answer?***
- ***How easy or hard was it to answer this question?***

FEEDBACK

12. How many days during the past week did you give your child a vegetable at dinner? Do not include vegetable juice, French fries, or white potatoes. *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 12

- ***How did you come up with your answer?***
- ***How easy or hard was it to answer this question?***

FEEDBACK

Questions on Milk

13. Did your child drink milk or use milk on his/her cereal at home during the past week? *(Circle one.)*

    1. No **[Go to Question 16]**
    2. Yes

14. What kind of milk did your child usually drink or use on his/her cereal at home during the past week? *(Circle one.)*

    1. Whole milk
    2. 2% milk (reduced-fat milk)
    3. 1% milk (low-fat milk)
    4. Skim milk (fat-free milk)
    5. Other (for example, soy, almond or rice milk)

QUESTIONS 13-14

- ***How did you come up with your answer?***
- ***Does your child drink more than one type of milk? If so, how did you answer this question?***
- ***Does your child drink a different type of milk at school than at home? If so, how would you answer this question during the school year?***
- ***In your own words, can you please describe to me how the types of milk in this question are different?***
- ***Are there other words or phrases you use for these types of milk?***

FEEDBACK

15. How many days during the past week did you give your child milk to drink at dinner? *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 15

- ***How did you come up with your answer?***
- ***How easy or hard was it to remember how many days in the past week you gave your child milk to drink at dinner?***

FEEDBACK

16. Which one of these statements best describes how you feel about the milk you give your child? *(Circle one.)*

    1. I believe that whole milk is healthier than low-fat milk for my child.
    2. I believe that low-fat milk is healthier than whole milk for my child.
    3. I believe that whole milk and low-fat milk are equally healthy for my child.

QUESTION 16

- ***Why did you choose this answer?***
- ***How easy or difficult is it to use the response scale?***
- ***Were you able to find a response that fit your opinion?***
- ***In your own words, please explain to me what whole milk and skim milk are.***
- ***In your own words, please explain to me what "healthier" means.***

FEEDBACK

Questions on Eating and Shopping Habits

17. How strongly do you agree or disagree with each of these statements? *(Circle one for each statement.)*

a. It is easy to buy fresh fruits or vegetables where I live.	Strongly agree	Agree	Disagree	Strongly disagree
b. There is a large selection of fresh fruits or vegetables where I live.	Strongly agree	Agree	Disagree	Strongly disagree
c. I do not usually buy fresh fruits or vegetables because they spoil quickly.	Strongly agree	Agree	Disagree	Strongly disagree
d. I can afford fruits or vegetables in the store where I shop for most of my food.	Strongly agree	Agree	Disagree	Strongly disagree
e. I can encourage my child to try new fruits or vegetables.	Strongly agree	Agree	Disagree	Strongly disagree
f. I usually drink 1% milk or skim milk.	Strongly agree	Agree	Disagree	Strongly disagree

QUESTION 17

- ***For each item, how did you come up with your answer?***
- ***What does "fresh fruits or vegetables" mean to you?***
- ***What does "where I live" mean to you?***
- ***What does "large selection" mean to you?***
- ***What does "spoil quickly" mean to you?***
- ***What does "afford" mean to you?***
- ***What does "where I shop for most of my food" mean to you?***
- ***What does "I can encourage my child" mean to you?***
- ***If "strongly agree" or "agree," how often do you drink 1% or skim milk?***

FEEDBACK

18. During the past month, how often did your child ask you to buy a certain type of fruit? *(Circle one.)*

    1. Never
    2. Seldom
    3. Sometimes
    4. Often
    5. Always

QUESTION 18

- ***How did you come up with your answer?***
- ***How easy or difficult is it to use the response scale?***
- ***Were you able to find a response that fit your opinion?***
- ***When was the last time your child asked you to buy fruit for him or her?***
- ***What was the last fruit your child asked you to buy for him or her?***
- ***Where did your child ask you to buy fruit for him or her? Probe: grocery store, restaurant, take out.***

FEEDBACK

19. During the past month, how often did your child ask you to buy a certain type of vegetable? *(Circle one.)*

    1. Never
    2. Seldom
    3. Sometimes
    4. Often
    5. Always

QUESTION 19

- ***How did you come up with your answer?***
- ***When was the last time your child asked you to buy a vegetable for him or her?***
- ***What was the last vegetable your child asked you to buy for him or her?***
- ***Where did your child ask you to buy vegetables for him or her? Probe: grocery store, restaurant, take out.***

FEEDBACK

20. How many days during the past week did your child help you make or cook a meal? For example, did your child wash fruits or vegetables? *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 20

- ***How did you come up with your answer?***
- ***When was the last time your child helped you make or cook a meal? How did he or she help you?***

FEEDBACK

21. How many days during the past week did you and your child sit down to eat dinner or supper as a family? *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 21

- ***How did you come up with your answer?***
- ***What does "sit down to eat dinner or supper as a family" mean to you?***

FEEDBACK

22. How many days during the past week did your child eat dinner or supper with the TV on? *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 22

- ***How did you come up with your answer?***

FEEDBACK

23. How many days during the past week did your child help select the food your family eats at home? *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 23

- ***How did you come up with your answer?***
- ***What does "select the food your family eats" mean to you?***
- ***How does your child help you select food your family eats at home?***

FEEDBACK

24. How many days during the past week did your child ask to have fruits or vegetables to eat? *(Circle one.)*

    1. None
    2. 1 to 2 days
    3. 3 to 4 days
    4. 5 to 6 days
    5. Every day

QUESTION 24

- ***How did you come up with your answer? Did you include all snacks and meals?***

FEEDBACK

25. How many days during the past week did you eat fruit for a snack? *(Circle one.)*

1. None
2. 1 to 2 days
3. 3 to 4 days
4. 5 to 6 days
5. Every day

QUESTION 25

- ***How did you come up with your answer?***
- ***How easy or hard was it to remember how many days in the past week you ate fruit as a snack?***

FEEDBACK

26. How many days during the past week did you eat vegetables for a snack? *(Circle one.)*

1. None
2. 1 to 2 days
3. 3 to 4 days
4. 5 to 6 days
5. Every day

QUESTION 26

- ***How did you come up with your answer?***
- ***How easy or hard was it to remember how many days in the past week you ate vegetables as a snack?***
- ***How did you feel about providing information on the amount of fruits and vegetables you eat?***

FEEDBACK

Questions about You and Your Household

27. Does anyone in your household currently get food assistance, such as Supplemental Nutrition Assistance Program (SNAP) benefits or Food Stamps? *(Circle one.)*

1. No
2. Yes

QUESTION 27

- ***Are there other words or phrases you use to refer to food assistance?***

FEEDBACK

28. Do any children in your household currently get Women, Infants, and Children (WIC) benefits? *(Circle one.)*

    1. No **[Go to Question 30]**
    2. Yes

29. Are any of these children 2 to 5 years old? *(Circle one.)*

    1. No
    2. Yes

30. How many people under 18 years of age live in your household?

    _____

31. Including yourself, how many people 18 years of age or older live in your household?

    _____

32. What is your age? *(Circle one.)*

    1. 18 to 24
    2. 25 to 34
    3. 35 to 44
    4. 45 to 54
    5. 55 to 64
    6. 65 to 74
    7. Over 74

33. What is your gender? *(Circle one.)*

    1. Male
    2. Female

Please answer the next two questions about your ethnicity and race.

34. Are you Hispanic or Latino? *(Circle one.)*

    1. No
    2. Yes

35. What is your race? *(Circle all that apply.)*

    1. White
    2. Black or African American
    3. Asian
    4. Native Hawaiian or other Pacific Islander
    5. American Indian or Alaska Native
    6. Other (*Describe*): _____

36. In what month was the child who is participating in the "What Does Your Child Eat?" study born? *(Circle one.)*

    1. January
    2. February
    3. March
    4. April
    5. May
    6. June
    7. July
    8. August
    9. September
    10. October
    11. November
    12. December

37. In what year was the child who is participating in the "What Does Your Child Eat?" study born? *(Enter year; for example, 1999.)*

    ____

38. Do you have any other children attending the same school as the child who is participating in the "What Does Your Child Eat" study? *(Circle one.)*

    1. Yes
    2. No

39. Does your family speak English at home? *(Circle one.)*

    1. We speak English all of the time at home.
    2. We speak English some of the time at home and speak another language some of the time.
    3. We never speak English at home. We speak another language.

Appendix C: Guidelines and Suggestions for Training Data Collectors

- Begin with introductions, a brief overview of the training and schedule, and a review of logistical information and ground rules. End each day with a review of key material, perhaps using a question-and-answer format in which you ask the questions and trainees provide the answers. You might also offer suggestions for practice, especially if certification is the next day. Start subsequent days with a brief review of material from the previous day(s), answer any pending questions, share any overnight changes in procedures or schedule, and review the day's agenda. After training ends each day, the primary investigator (PI) or evaluation manager should gather the training staff together for a daily debriefing to address any questions and issues raised that day, data collector performance and concerns, training agenda changes, and any upcoming events or project news that need to be conveyed to the training team.
- Proceed through the agenda responding to questions if they are general and pertinent to all or most trainees. Questions about highly unlikely "what if" scenarios or other individual situations should be written down and handled independently during a break rather than being allowed to derail the flow of training and your timeline. If a question turns out to have wider applicability, you can share the information with the class later.
- It is often helpful to review surveys in a "round robin" format initially with the lead trainer playing the role of respondent and trainees taking turns reading the questions. This approach allows the trainer to still guide where the interview goes and can point out tricky questions, newly introduced response categories, and so on.
- Because a round robin exercise provides limited opportunity for each individual to practice, the group should also be broken up into pairs for mock interviews in which one plays the role of interviewer and the other the respondent. Then they reverse roles. As instruments become more numerous and/or complicated, the more rounds of paired mock interviews should be scheduled.
- Because the ability to describe a study succinctly and respond to questions in a natural and factual way is such an important requirement, special attention should be given to practicing responses to "Frequently Asked Questions" (FAQs). A simple read-through of an FAQ document is insufficient. Trainers can reword the questions or throw in related ones and solicit multiple responses so trainees can develop more spontaneity and increase their comfort.
- Exercises can also be planned around case management, data entry, handling of the various forms, or recruitment.
- For projects requiring bilingual data collectors, include a specific plan for bilingual training and hold this training immediately following regular project training. Bilingual training should be led by bilingual staff and include a general overview of cultural sensitivity and cultural differences related to food-related activities as well as a review of study-specific materials.
- Refresher training (for existing data collectors) would be advisable if there is an extended time gap between rounds of data collection (e.g., pre-intervention collection at the start of each semester) or between pre- and post-intervention data collection. This training also provides the opportunity to focus on elements of data collection that can be improved based on lessons learned from earlier data collections.

Appendix D: Results From an Attrition Analysis

Table D-1 shows the results from the attrition analysis conducted for the evaluation of the Iowa Nutrition Network's (INN's) Building and Strengthening Iowa Community Support (BASICS) for Nutrition and Physical Activity Program. The potential impact of attrition from the evaluation study on generalizability of the study findings was assessed by comparing the pre-intervention similarity of study participants who provided follow-up data and those who did not. Respondents in the oldest age group (45 years or older) were nearly 3.5 times more likely than individuals in the youngest age group (18–34) to complete the follow-up survey, and respondents between the ages of 35 and 44 were more than twice as likely as individuals in the youngest age group to complete the follow-up survey.

Table D-1.— Attrition Analysis for the Evaluation of the BASICS Program

Characteristic	Estimated Odds Ratio	95% Wald Confidence Limits[a]		*p*-value
		Lower	Upper	
Child demographics				
Sex				
Male	0.90	0.68	1.20	0.4789
Female (reference group)	1.00	—	—	—
Age	0.78	0.56	1.08	0.1330
Parent[b] / household demographics				
Respondent age				
18 to 34 (reference group)	1.00	—	—	—
35 to 44	2.15**	1.55	2.98	<0.0001
45 or older	3.60**	1.72	7.50	0.0006
Respondent sex				
Male	0.73	0.43	1.25	0.2514
Female (reference group)	1.00	—	—	—
Respondent race and ethnicity				
White, non-Hispanic (reference group)	1.00	—	—	—
Hispanic or Latino	0.70	0.48	1.03	0.0676
Black, non-Hispanic	0.91	0.61	1.37	0.6575
Other or more than one race[c]	1.78	0.88	3.61	0.1117
Size of household	1.00	0.93	1.08	0.9673

** Indicates statistical significance if the *p*-value is less than or equal to 0.01.

[a] Estimate (with 95% confidence limits) indicates the odds ratio of completers to noncompleters.

[b] Represents the parent/guardian who completed the survey.

[c] Includes respondents who selected more than one race category.

Notes: Generalized linear mixed model (SAS PROC GLIMMIX) used to evaluate program attrition while accounting for the clustering of students within schools. Dichotomous participation indicator (based on availability of post-intervention data) was regressed on child and parent demographic characteristics and household descriptors.

Source: Parent Baseline Survey, data collected September–October 2011; respondents are parents/caregivers of children participating in the evaluation study.