

Evaluating Food Stamp Nutrition Education: Process for Development and Validation of Evaluation Measures

Marilyn S. Townsend, PhD, RD

ABSTRACT

The purpose of this paper is to describe a process for developing and validating outcome measures relevant to dietary quality behaviors targeted by Food Stamp Nutrition Education (FSNE). The ultimate goal is a measure that is valid, reliable, sensitive to change, and practical for use for a wide variety of FSNE evaluation purposes. The development process has incorporated input from FSNE stakeholders at the federal and state level and follows a systematic, research-driven approach that incorporates both qualitative and quantitative research and includes methods for identification of subject domains, selection of evaluation items, initial pretesting, and reduction of items. This type of research establishes the trustworthiness of new evaluation measures.

Key Words: evaluation, measure, validity, reliability, FSNE

(*J Nutr Educ Behav.* 2006;38:18-24.)

INTRODUCTION

As has been discussed elsewhere in this special section,^{1,2} Food Stamp Nutrition Education (FSNE) now supports a considerable share of nutrition education activities intended to improve the quality of the diets of Food Stamp Program participants.³ Since its inception, FSNE has operated with great diversity, both across states and within them. This diversity has allowed states and localities to tailor programs to local interests and needs, but has complicated evaluation.^{1,2} Currently, the ability of program and policy officials to assess the effectiveness of these activities is limited, with the lack of commonly agreed-upon useful outcome measures a primary stumbling block to improving evaluation.¹

Ideally, the nutrition and food resource management information provided through FSNE should assist low-income Americans in maintaining household food security and improving the quality of their diets. In the area of household food security assessment, a common measure—the U.S. Household Food Security Survey Module—has been developed and widely adopted. Its use in national tracking surveys such as the Current Population Survey and the National Health and Nutrition Examination Survey allows ongoing monitoring of trends in food security at the national and state levels. In addition, methods of administering the survey at state and local levels have been developed and disseminated.⁴ A similar measure for assessing diet quality outcomes targeted by FSNE is lacking, and has been identified as a basic evaluation need.^{1,2}

There is a paucity of validation research on evaluation measures for all programs, but particularly those serving low-income communities.⁵⁻⁷ The process for developing such measures is complex and expensive. First, the content should reflect the behaviors targeted by FSNE for its primary prevention interventions and should be consistent with federal dietary guidance policy, as stated in the 2005 Dietary Guidelines for Americans⁸ and interpreted in annual guidance provided by the Food and Nutrition Service to state partners.⁹ Since the recommendations of the Dietary Guidelines for Americans are multifactorial, encompassing multiple aspects of dietary behavior (for example, consumption of fruits and vegetables, whole grains, etc.), multiple scales or measures may be required to assess all important aspects of diet quality. For credibility, the measures must meet generally accepted standards for validity, reliability, sensitivity, and internal consistency.⁵⁻⁷ For practicality, the measures should be easy to administer and have a low respondent burden, meaning they should be sufficiently brief and understandable to Food Stamp Program participants.^{5-7,10,11}

The purpose of this paper is to describe a proposed process for developing and validating such measures that can be useful for a variety of FSNE evaluation purposes, satisfying the criteria mentioned above. This overall process includes methods for identification of subject domains with their corresponding behaviors, selection of evaluation items, initial pretesting, and reduction of items, as well as estimates of validity and sensitivity in a longitudinal study. This type of research establishes the trustworthiness of new evaluation tools, and it provides information to program and policy officials on their value in assessing program recipient needs and monitoring progress.⁵ Through a col-

Cooperative Extension Nutrition Specialist, University of California, Davis; 1 Shields Ave., Nutrition Dept., Davis, CA 95616-8669

©2006 SOCIETY FOR NUTRITION EDUCATION
doi: 10.1016/j.jneb.2005.11.008

laborative process between the Society for Nutrition Education and federal and state stakeholders,² this process has already been initiated and is progressing with continued support from the United States Department of Agriculture's Economic Research Service and Food and Nutrition Service.

OVERVIEW OF THE EVALUATION PROCESS

The proposed research process takes place in a series of stages, moving from generation of potential measurement items to establishment of reliability, validity, and sensitivity (Tables 1 and 2). Overall, the measure should reflect the goals of the FSNE interventions, which are derived from federal dietary guidance policy. Ideally, the measure should offer the flexibility of being decomposable into scales reflecting important subcomponents of diet quality, such as fruit and vegetable consumption. This would allow state

programs or others conducting FSNE-related research and evaluation the option of using only selected scales of the measure, as indicated by program emphasis, client needs, evaluation budget, or time constraints.

Stage 1: Domain selections. The purpose of this stage is identification of subject matter areas or content domains related to diet quality that have content validity, as judged by experts. Content domains must be consistent with FSNE goals for improvement in diet quality, which, as previously stated, are based on the Dietary Guidelines for Americans.⁸ To facilitate this first stage and serve as a starting point, an extensive and thorough review of the relevant nutrition science and medical literature guides the process. For this stage, the Economic Research Service (ERS) convened a conference in 2004 of selected FSNE program professionals, researchers, and relevant agency staff to serve as experts

Table 1. Methodological considerations for assessing psychometric characteristics for a proposed FSNE diet quality measure

	Who is involved?	What?	When? (by stage)	Cost*
Validity – Development of items				
Content	Experts	Selects relevant content domains from the nutrition and medical literature. For each domain, identifies the corresponding behaviors with test items appropriate for FSNE target audiences.	Initially Stage 1	\$
Face	Clients	Matches wording of test items to vocabulary of client.	Initially Stage 1	\$\$
Validity – Testing of items [†]				
Construct	Clients	Reserve for those scales for which there are no objective measures (eg, attitudes, beliefs)	Stages 1-6	\$\$-\$\$\$\$
Convergent	Clients	Determines links to diet	After item pool/scales in place	\$\$\$\$
Criterion	Clients	Determines links to health	After item pool/scales in place	\$\$\$\$
Reliability[†]				
Stability (also called temporal reliability)	Clients	Does the item give same response over time for same client?	Mid	\$\$
Internal consistency (alpha & inter-item correlation)	Clients	Do the items in the scale all contribute to the construct?	Mid	\$\$
Other Tests[†]				
Sensitivity to change	Clients		Final stage following intervention	\$\$\$\$

*Cost refers to the relative cost among the various procedures in this proposed process

[†]A randomized, controlled trial could be conducted as one major study of 2-3 major ethnic/racial groups to include data (ie, multiple 24-hour dietary recalls, biomarkers, demographic information, behavioral items being considered for final version of the FSNE measure) collected at baseline and postintervention.

Table 2. Example of a development process for a diet quality measure for community nutrition education programs

Stage	Description	Importance for quality outcomes	Technical term
1 Domain selections	Using peer-reviewed, published research on chronic disease, select appropriate content/domains and their corresponding behaviors.	Essential	Content validity
2 Item generation	Generate draft of individual items and their response options for each behavior using peer-reviewed, published research wherever possible. The items should reflect objectives of FSNE as identified in the Logic Model. FSNE professionals should be satisfied with the overall emphasis of the measure.	Essential	
3 Item pre-testing	Review wording of each item with members of various FSNE audiences. Using individual interviews and standardized protocol, ask client to describe what the item means to her, using her own words. Clarify meaning of key words.	Essential	Face validity
4 Item testing & analyses	Using data from clients, examine performance of each item for an item difficulty analysis. For items not functioning optimally, revise wording and retest or eliminate item. Administer test to clients at two time points without the curriculum. We want clients to respond the same way at each time point. Examine performance of each scale for internal consistency.	Advisable; often done.	Item difficulty index
5 Convergent & criterion validity	Does the new diet quality measure correlate with established measures of diet or health status? Do the items reflect actual behavior as we are claiming? Are these behaviors related to health status?	Advisable; sometimes done. Advisable; sometimes done. More difficult and costly than other aspects of evaluation. Advisable; but rarely done.	Temporal reliability (stability) Internal consistency Convergent validity if we use 24-hr recall as a surrogate for actual diet. Criterion validity if we use an external measure for health, such as a biomarker (eg, a serum level that indicates nutrient intake).
6 Sensitivity	We want an instrument to reflect <i>change</i> on the posttest, so we would test for sensitivity. Remove insensitive items, as they detract from impact. Need a longitudinal design.	Advisable; but rarely done.	Sensitivity can be part of above grant proposal.

about the selection of content for FSNE evaluation measures. Using a systematic and qualitative approach, the product of this stage is the selection of content domains or areas for each evaluation measure (ie, content validity demonstrated).^{12,13}

Stage 2: Item generation. Each content domain is represented by a group of behaviors used by FSNE clients. Each behavior is represented by one or more items with response options. This review generates a pool of items compiled from previously conducted research with low- and middle-income consumers, focusing on existing tools with reported psychometric properties. The item pool should contain as many items as possible, many more than desired for the final measure. As a guideline, twice as many items are needed for this stage as for the final measure, so that inadequate items can be eliminated.¹⁴ In the case of FSNE, a pool of items was prepared by Mathematica Policy Research Incorporated,¹⁵ under contract to the Economic Research Service, in preparation for item selection by FSNE content experts. Items should have correct responses that are unequivocal among both the nutrition academic community and knowledgeable consumers.¹⁴

To identify items for the behaviors associated with FSNE clients for each content domain, ERS, in conjunction with the Society for Nutrition Education (SNE), Cooperative State Research, Education, and Extension Service (CSREES), and the FNS, identified FSNE experts representing a diversity of state, regional, and program delivery perspectives, and invited them to a workshop held April 15-16, 2004, at ERS. At the workshop, experts selected items from the pool that they deemed to have content validity and made suggestions for modification or additions, as necessary. This process resulted in a set of candidate items for inclusion in diet quality measures. The selection of items should reflect all behaviors associated with each domain. For example, for the calcium domain, the items should reflect a multitude of behaviors used by FSNE clients to eat calcium-rich foods. These items were estimated by FSNE experts to be relevant to the content and intended outcomes of FSNE interventions as they are implemented in states and territories across the United States. Although item selection was a collective decision by experts very familiar with FSNE content, this was nonetheless a subjective judgment and was only the first of multiple steps in establishing validity.^{12,13,14}

Stage 3: Item pretesting. While items representing the selected behaviors may have content validity, they may not be understandable and answerable by the target audience. In the case of the pool of candidate items compiled from a literature review for the April 15-16 FSNE Workshop, for example, most were not developed with the FSNE audience in mind. Instead, we know from the literature review that most validation studies are conducted with middle-income, educated subjects. The FSNE experts at the

Workshop recommended that cognitive testing be employed to assess how well the items work with the many FSNE target audiences. Cognitive testing is a form of structured interviewing that provides insights into how survey items are understood by respondents typical of the target audience. This interview method examines respondent strategies for developing responses to those items.¹⁶ Willis identified three qualitative cognitive testing strategies for questionnaire development.¹⁷ The first is the *think aloud* technique, where participants respond to a questionnaire item and then are asked to retrospectively describe the meaning of the item using her own words and, step-by-step, elucidate how she decided upon her selected response option. The second is the use of *probing* to encourage the respondent to elucidate further these meanings. And the third is the *paraphrasing* technique, where the respondent is asked to restate the item using her own words. Respondents are also asked how to make each item more understandable to another reader. Modifications are made to the wording of items that are not well understood, and revisions are further tested with participants.¹⁸ The process is iterative. The process can also provide information on the relevance of the item to the diet quality of the respondent. This stage results in wording of the item and its response option in a near final form. The wording should now be meaningful to the FSNE audience (ie, face validity demonstrated).^{12,13}

To provide the most useful information, the candidate items selected by experts attending the FSNE Workshop should be cognitively tested with individuals representative of the FSNE target audience. Since this FSNE audience is extremely diverse, including individuals of all ages and ethnicities—among them, a considerable subset who does not speak English as their first language—it may not be practical to conduct cognitive testing with all subgroups at the initial stage of development of the measure. Therefore, it would be important to select initial subgroups for testing that represent a major subset of the overall target population. If results are promising, further testing could be done with other important subgroups, as priorities and budgets allow.

Stage 4: Item testing and analyses. Once cognitive testing has produced a set of items that are understandable and answerable by the target population, a larger, more structured data collection can be conducted and results used for further analysis. In this stage, items are assessed for suitability using analytical tests for content, temporal reliability, clustering of items into meaningful scales, and internal consistency of scales. Results can be used to decide which items should be retained.¹⁹ Analyses that provide useful information include the following:

Item analysis. Item analysis generating an item difficulty index indicates the extent to which participants answer an item in the same way. Items reflecting behaviors that almost all FSNE clients already perform at baseline on the pretest would not be useful in the measure. For example, items with

mean values over 0.8 would have little room for change in the expected direction. Items with means below 0.2 may not be relevant to a sufficient number of food stamp recipients and should be considered for deletion.

Temporal reliability. Reliability, also known as stability, represents the consistency of the item responses on the evaluation instrument.^{12,20} If item responses are consistent (ie, reliable), then respondents reply to the item with the same answer during this period, with no intervening education intervention. A measure that is not stable would be considered unreliable. Because the first testing influences responses on the second testing as clients are now familiar with the measure, reliability coefficients have been shown to sometimes be inflated. In the case of the proposed FSNE measure, then, it would be appropriate to have individuals typical of the FSNE population complete the evaluation tool on two occasions 3 to 6 weeks apart with no intervention.

Factor analysis. Each evaluation tool or scale is factor analyzed.²¹ The factor loadings are examined to see if the items cluster into one or more groups that are intuitively meaningful. Ideally, if the scale is intended to measure one domain (eg, fruit and vegetables), then the items should load on one factor. Generally, items should have a loading of $> +.30$ with the factor.

Internal consistency. There are several approaches to addressing the issue of internal consistency, also known as homogeneity of the scale. For a scale of items with more than 2 response options, Cronbach's coefficient alpha is used.^{13,22} (Note: For a scale of dichotomous items, the Kuder-Richardson Formula 20 is the appropriate analysis.) Items are deleted for maximum alpha level, with consideration of theoretical justification for retaining or deleting items. An alpha between .60 and .69 is minimally acceptable. An alpha greater than .70 is desirable.^{12,13} If temporal reliability is high, but the coefficient alpha is low for a scale on a measure, then the scale would be considered unreliable.

The second approach is termed median inter-item correlation and is an index of the degree to which individual items correlate with each other. The alpha and inter-item correlations range from 0 to 1. A median inter-item correlation of greater than .20 is desirable in educational measurement.¹³

Ethnic differences. Ideally, items should score similarly for the ethnic and cultural groups in the FSNE audience. However, cultural differences in food patterns or interpretation of the items exist. Identifying those items and seeking alternatives that provide more uniform results is important. If this is not possible, users should be informed of the potential variation in results.⁵

Stage 5: Assessing scales for convergent and criterion validity. For policy makers, program officials, and the general public, the most important aspect of any scale or evaluation tool is the extent to which it measures

what it is intended to measure (ie, its accuracy).^{12,13} In the case of FSNE, how well does the scale measure dietary quality? Validation of dietary quality measures employs one or both of two approaches, convergent and criterion validity. Convergent validity is assessed using multiple dietary recalls, or food records. These longer, more detailed forms of dietary assessment are commonly accepted measures of dietary status; however, because they are time consuming, costly to administer and analyze, and create high levels of respondent burden, they are not practical for ongoing monitoring of public programs such as FSNE.

The dietary quality estimates obtained from these longer forms of self-report, whether expressed as nutrient intakes, food groups servings, or a summary dietary quality score such as USDA's Healthy Eating Index, would be expected to correlate with relevant responses to the short items included in the proposed FSNE measure.^{5,23} If this correlation is strong, the short measure could be considered an adequate substitute for the more burdensome, longer forms of dietary assessment. However, since both types of data are reported by the FSNE clients, both may suffer from measurement errors inherent to self-report.²³

Criterion validity of the FSNE measure would be assessed using a biological measure (ie, biomarker) of dietary status. Townsend et al, for example, used serum carotenoids as a biomarker to assess a short client-friendly measure of fruit and vegetable behaviors.⁵ Although biomarkers do have limitations associated with their use, their strength is that they are not self-report measures, and therefore, they provide an independent assessment of diet quality. Ideally, FSNE researchers can combine convergent and criterion validity, using both long-form dietary assessment methods and biomarkers to assess the validity of a FSNE diet quality measure.

Stage 6: Sensitivity assessment (Table 1). Experts have considered items for the measures that research has indicated offers potential for capturing existing change from the FSNE education intervention. This stage collects the evidence with FSNE clients. This aspect of assessment investigates to what extent a measure is sensitive or responsive to changes in behaviors of a target population, such as FSNE participants.²² For example, we might assess those changes in fruit and vegetable behaviors using an indicator of dietary quality, such as serum carotenoids, as a biomarker and/or Vitamins A, C, and folate from multiple 24-hour dietary recalls. Using a longitudinal research design, items are considered appropriate for the evaluation tool if they correspond to changes in an indicator of dietary quality (eg, the multiple 24-hour dietary recalls and/or the serum carotenoid or other biomarker).^{24,25}

Limitations. A comprehensive FSNE evaluation presents major challenges. The focus of this effort is dietary quality, not other FSNE emphases such as Food Resource Management or Food Safety. Moreover, all age, sex, ethnic, racial, cultural,

and language groups participate in the Food Stamp Program and would, therefore, ideally be positively influenced by the nutrition information provided through FSNE. However, for reasons of cost and practicality, current measurement development efforts being supported by USDA's ERS and FNS are focusing on a subset of the population defined as English-speaking non-Hispanic White, non-Hispanic African-American, and Hispanic women who are the primary person responsible for food purchasing, meal planning, and food preparation in their households.² Additional measures may need to be developed to assess aspects of FSNE education other than diet quality, or for use with some subpopulation groups (eg, school-age children). Other types of data, such as that from food purchase receipts, may be useful as supplemental sources of information. It may be important to continue to periodically collect more detailed data on nutrition knowledge, attitudes, and behaviors of the Food Stamp Program population, as was done in the 1996 National Food Stamp Program Survey conducted by USDA's Food and Nutrition Service.²⁶

CONCLUSIONS

This paper provides an overview of important research steps in developing and validating short measures related to dietary quality that would be practical for use in a large, community-based nutrition education program, such as those currently funded through FSNE. This information should assist nutrition education researchers and program managers working together to improve evaluation of publicly funded nutrition education programs.

No evaluation tool is a perfect measure of what it is intended to measure, and all have sources of error. Lengthy dietary assessment for measurement of program outcomes is extremely difficult to conduct. These measures are often very costly and out of the reach of community programs such as FSNE. However, stakeholders increasingly demand accountability, and FSNE program managers are interested in providing it. Therefore, this paper outlined a systematic, research-driven process for developing and validating short measures that is based on the identified priorities of FSNE experts. If this 6-stage research process described herein results in a useful common core set of evaluation tools that can be used at state and local levels, pooled for regional and national reporting use, and measured over time, then this process would be a cost-effective way of addressing stakeholders' information needs.

IMPLICATIONS FOR RESEARCH AND PRACTICE

Accurate assessment of the behavioral impacts of community nutrition education programs requires valid and reliable evaluation tools or measures. For FSNE program managers, development of common measures of dietary quality with known reliability, validity, and sensitivity will provide a basis for improving program evaluation and a method for reporting

outcomes collectively for states. Results from the evaluation process will assist program managers with making targeted improvements in FSNE program quality. Documenting the process of development of this proposed measure will also provide information for policy makers and researchers seeking to improve FSNE dietary assessment methodology.

REFERENCES

1. Taylor-Powell E. Evaluating Food Stamp Nutrition Education: A view from the field of program evaluation. *J Nutr Educ Behav.* 2006;38:12-17.
2. Guthrie J, Stommes E, Voichek J. Evaluating Food Stamp Nutrition Education: Issues and opportunities. *J Nutr Educ Behav.* 2006;38:6-11.
3. Little DM, Newman ME. *Food Stamp Nutrition Education within the Cooperative Extension/Land-Grant University System: National Report FY 2002.* Washington DC: USDA, Cooperative State Research, Education and Extension Service (CREES); 2003.
4. Bickel G, Nord M, Price C, Hamilton WL, Cook JT. *Guide to Measuring Household Food Security.* Alexandria, VA: USDA, Food and Nutrition Service; Revised 2000.
5. Townsend MS, Kaiser LL, Allen LH, Joy AB, Murphy SP. Selecting items for a food behavior checklist for a limited resource audience. *J Nutr Educ Behav.* 2003;35:69-82.
6. McClelland JW, Keenan DP, Lewis J, et al. Review of evaluation tools used to assess the impact of nutrition education on dietary intake and quality, weight management practices, and physical activity of low-income audiences. *J Nutr Educ.* 2001;33:S35-S48.
7. Contento IR, Randell JS, Basch CE. Review and analysis of evaluation measures used in nutrition education intervention research. *J Nutr Educ Behav.* 2002;34:2-25.
8. US Department of Health and Human Services and US Department of Agriculture. *Dietary Guidelines for Americans, 2005.* Available at: www.healthierus.gov/dietaryguidelines. Publication # HHS-ODPHP-2005-01-DGA-A. Accessed May 9, 2005.
9. U.S. Department of Agriculture, Food and Nutrition Service. *Food Stamp Program. Food Stamp Nutrition Education Plan Guidance Federal Fiscal Year 2006.* March 2005. Accessible at: http://www.nal.usda.gov/foodstamp/programplan/FSNE_Plan_Guidance_06.pdf. Accessed May 9, 2005.
10. Murphy SP, Kaiser LL, Townsend MS, Allen LH. Evaluation of validity of items for a health beliefs checklist. *J Am Diet Assoc.* 2001;101:751-761.
11. Glasgow RE, Perry JD, Toobert DJ, Hollis JF. Brief assessments of dietary behavior in field settings. *Addictive Behav.* 1996; 21(2):239-247.
12. Litwin MS. *How to Measure Survey Reliability and Validity.* Thousand Oaks, CA: Sage Publications; 1995.
13. Nunnally JC, Bernstein IH. *Psychometric Theory*, 3rd ed., New York: McGraw-Hill, Inc.; 1994;99-100.
14. Parmenter K, Wardle J. Evaluation and design of nutrition knowledge measures. *J Nutr Educ.* 2000;32:269-277.
15. Hartline-Grafton H, Nyman R, Briefel R, Cohen R. Developing Common Core Survey Questions to Assess Key Dietary Behavioral Outcomes of FSNE: Launching the Research Process. *Prototype Notebook: Short Questions on Dietary Intake, Knowledge, Attitudes and Behaviors.* Submitted by Mathematica Policy Research, Inc. to US Department of Agriculture, Economic Research Service, Food and Rural Economics Division. E-FAN No. (04010) 171 pp, September 2004. Available at: <http://www.ers.usda.gov/publications/efan04010/>.
16. Willis GB. *Cognitive Interviewing and Questionnaire Design: A Training Manual (Working Paper Series No. 7).* Hyattsville, MD: Centers for Disease Control and Prevention, National Center for Health Statistics; 1994.
17. Willis GB, Royston P, Bercini D. The use of verbal report methods in the development and testing of survey questionnaires. *Appl Cogn Psychol.* 1991;5:251-267.

18. Godin G, Kok G. The Theory of Planned Behavior: Review of its applications to health-related behaviors. *Am J Health Promot.* 1996; 11(2):87-98.
19. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chron Dis.* 1985;38(1):27-36.
20. Carmines EG, Zeller RA. *Reliability and Validity Assessment.* Newbury Park, CA: Sage Publications; 1979.
21. Pedhazur EJ, Schmelkin LP. *Measurement, Design, and Analysis: An Integrated Approach.* Mahwah, NJ: Lawrence Erlbaum Assoc.; 1991.
22. Cronbach LJ. Coefficient alpha in the internal structure of tests. *Psychometrika.* 1951;16:297-334.
23. Mertz W. Food intake measurements: Is there a "gold standard"? *J Am Diet Assoc.* 1992;92(12):1463-1465.
24. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chron Dis.* 1987; 40(2):171-178.
25. Kristal AR, Beresford SA, Lazovich D. Assessing change in diet-intervention research. *Am J Clin Nutr.* 1994;59(suppl):185S-189S.
26. Cohen B, Ohls J, Andrews M, Ponza M, Moreno L, Zambrowski A, Cohen R. Food Stamp Participants' Food Security and Nutrient Availability. Final Report. 1999. 286 pp. Available at: <http://www.fns.usda.gov/oane/MENU/Published/NutritionEducation/Files/nutrient>. Accessed 11/20/04.