*Research Brief*

# Confirming the Reliability and Validity of Others' Evaluation Tools Before Adopting for Your Programs

Garry Auld, PhD[1]; Susan Baker, EdD[1]; Kathryn McGirr, MS, RD[1];
Kristen Suzanna Osborn, MS, RD[2]; Philip Skaff, MS, RD[3]

## ABSTRACT

**Objective:** To confirm the reliability and validity of a previously validated evaluation instrument in a new context.

**Methods:** In a cross–sectional study, the processes and results of testing *Cooking Matters'* (CM) use of the *Expanded Food and Nutrition Education Program's* Behavior Checklist as a retrospective pretest/posttest were described. The researchers determined reliability, face and content validity, and response–shift bias with 95 CM participants.

**Results:** Most items had acceptable face validity and moderate reliability; other items lacked reliability, or face or content validity (were unrelated to the CM curriculum).

**Conclusions and Implications:** Proper match between evaluation tools and curricula is needed for appropriate program assessment without which outcome data can be misleading or potentially invalid. Confirmation of validity is essential when adopting others' evaluation tools in new contexts, particularly for programs with widespread use such as federally funded programs and national nonprofit organizations.

**Key Words:** validity, *Cooking Matters*, *Expanded Food and Nutrition Education Program*, Share Our Strength, nutrition behavior checklist, low income (*J Nutr Educ Behav*. 2017;49:441–450.)

## INTRODUCTION

Program evaluation and monitoring are critical processes for determining success and justifying support of nutrition education programs.[1] The federal government has stressed the need for evaluation of health and social programs since the 1971 Report of the President's Committee on Health Education.[2] It is crucial that researchers and/or program leaders use appropriate evaluation methods that provide a true assessment of a program's effectiveness.

Numerous authors recommended multistep processes to establish the validity of evaluation tools.[3-6] For example, Radhakrishna[6] recommended a 5-step process for developing and testing questionnaires, including reviewing the literature, establishing validity and reliability, testing the format, and analyzing the data. Conserving resources by not reinventing the wheel was recommended by others[4,5] who often stressed the importance of revalidating the tools in a new setting. However, many programs adopted evaluation measures that were validated for other programs. This report describes the process of determining the appropriateness of an *Expanded Food and Nutrition Education Program* (EFNEP) English language evaluation tool for use in a nonprofit organization's nutrition program that may serve as a model approach for others.

*Cooking Matters* (CM) is the nutrition education branch of the nonprofit anti-hunger organization Share Our Strength. *Cooking Matters for Adults* is a 6-week series of classes that teach cooking, food safety, nutrition, and food resource management skills with the goal of improving food security of low-income audiences.[7] The audience and general approach of CM are similar to EFNEP's and contributed to their decision to use EFNEP's validated Behavior Checklist (BCL)[8] to assess CM outcomes. The original BCL underwent extensive testing in the early 1990s and again in 1997 before the national adoption of the core 10 questions that are still used.[9] Testing included establishing content validity (through expert panels and input from EFNEP coordinators), face validity (focus groups with

[1]Department of Food Science and Human Nutrition, Colorado State University, Fort Collins, CO
[2]Department of Nutrition Services, Gunnison Watershed School District, Gunnison, CO
[3]Nutrition Department, Metropolitan State College of Denver, Denver, CO

participants), discriminant validity (comparison of treatment and comparison groups), and criterion validity (using 24-hour dietary recalls). Cronbach $\alpha$ was .71 for the 10 items whereas factor analyses established 3 subscales.[9]

A previous CM evaluation demonstrated the program's effectiveness at conferring self-reported behavior change in participants.[10] The focuses of that study, by Swindle et al,[10] were to compare the traditional pre/posttest protocol with a retrospective pretest, assess the stability of reported behavior changes at 3 or 6 months postintervention, and determine response rates using 3 different follow-up approaches. In the current study, further assessment of CM evaluation was undertaken specifically to determine the face validity, content validity, and reliability of EFNEP's 10-question BCL, which measured food safety, diet quality, and food shopping-related behaviors, in the context of the CM intervention. Face validity measured how consistently survey items were interpreted by nonexperts in the field; content validity measured the relevance of the program evaluation to the program with which it was used; and reliability measured a survey's ability to produce consistent results.[5]

A response-shift bias occurs in testing when respondents have a different frame of reference or perception of a construct at 2 different time points, eg, they interpret questions differently before and after an intervention owing to a change in understanding brought about by the educational intervention.[11,12] In theory, the subject may respond 1 way to a question about his behaviors if asked before the start of the intervention and give a different answer to the same question after completing the intervention, not because the behavior has changed but because the question is interpreted differently. If this occurs, the program's outcome measures will be skewed, resulting in an overestimation or underestimation of the program's impact.[12-16] Unlike EFNEP's traditional pretest/posttest format, at the time of this study, CM used a retrospective pretest/posttest in which both the pretest and posttest responses were collected at the end of the final class.

The retrospective pretest/posttest format was thought to eliminate a possible response-shift bias by asking about postintervention and preintervention behaviors at the same time (posttest), ensuring an individual was interpreting the construct in the same way.[11,12,15] Swindle et al[10] compared the results from a traditional pretest/posttest with a retrospective pretest/posttest among CM participants and found no evidence of a response-shift bias; this suggested that perceptions of the construct did not change because of the intervention. However, that study included only a small sample (n = 12) and should be replicated with a larger sample. In contrast, the retrospective pretest might introduce other problems because of its greater reliance on memory[17,18] and potential social desirability bias,[18,19] defined as "the tendency to transmit a culturally accepted image, according to social norms."[20] Betz and Hill[19] found greater pre-post differences in the retrospective tests for socially desirable items.

The purposes of this study were to determine whether EFNEP's English language BCL was an appropriate evaluation tool for use by CM and secondarily to describe a process that others could use to confirm that the use of others' tools were appropriate in their settings. In this case, the CM adoption of EFNEP's 10-item BCL tool (and a few additional questions from EFNEP's test bank of optional items) raised concerns because CM used a different testing protocol (retrospective pretest/posttest vs EFNEP's pretest/posttest) and a different curriculum with fewer lessons, ie, a lower dose.

## METHODS
### Study Design

A total of 95 participants took part in various parts of this mixed-methods cross-sectional study that was approved by the Colorado State University Institutional Review Board. Primary outcomes were measured as follows: (1) Face validity was determined by whether participants had difficulty understanding the item (from their perspective) and the consistency of interpretations of survey items before or after intervention; (2) content validity was assessed by expert panel consensus when comparing curricula content and objectives with evaluation items; (3) the presence (or absence) of response-shift bias was determined by comparing a participant's traditional pretest score with her retrospective pretest score; and (4) reliability was determined by using a test/retest procedure of the retrospective pretest, ie, the retrospective pretest was given at the last (sixth) class as usual (the test) and then again during a reunion class 1 week later (the retest).

### Participants

A convenience sample of participants was recruited from CM classes in Colorado. They received a $20 grocer voucher for participating in any 1 of the following: (1) face validity cognitive interview, (2) completion of a traditional pretest earlier in the first class and a retrospective pretest/posttest after the final (sixth) class, or (3) completion of the retrospective pretest a second time, 1 week after the retrospective pretest/posttest. Because data collection occurred over several months, different participants were involved in each data collection step.

### Face Validity Assessment

To determine face validity for each survey item, the researchers used cognitive interviews and preselected probes to elucidate CM participants' interpretation of survey items.[21,22] After conducting supervised one-on-one practice interviews, 3 trained interviewers, met with participants individually to review the items. Participants were asked to read each item aloud (to confirm literacy) and then were asked to explain the question in their own words, and also whether anything was confusing about the question. They were then asked, What does X mean to you? where X would be the key concept(s) for each item, such as planning meals ahead of time or run out of food. Interviewers were trained to remain neutral, both verbally and nonverbally, to try to minimize social desirability bias in responses. All interviews were audio-recorded. The resulting data revealed the percentage of participants who had difficulty understanding the items from their perspective, and whether the items were