Journal of Human Sciences and Extension

Volume 8 | Number 2 Article 9

6-30-2020

Procedures for Improving Self-report Measurements to Capture Behavior Change: An Illustration

Glenn D. Israel
University of Florida, gdisrael@ufl.edu

Halil I. Sari
Kilis 7 Aralık Üniversitesi, hisari87@gmail.com

Nicole Owens Duffy
University of Florida, nicoleowens@ufl.edu

Sebastian Galindo-Gonzalez
University of Florida, sgalindo@ufl.edu

David C. Diehl
University of Florida, dcdiehl@ufl.edu

See next page for additional authors

Recommended Citation

Israel, G. D., Sari, H. I., Owens Duffy, N., Galindo-Gonzalez, S., Diehl, D. C., Abarca Orozco, S. J., Garcia Varela, E., & Sweeney, L. (2020). Procedures for Improving Self-report Measurements to Capture Behavior Change: An Illustration. *Journal of Human Sciences and Extension, 8*(2), 9. https://doi.org/10.54718/SJGB4387

Procedures for Improving Self-report Measurements to Capture Behavior Change: An Illustration

Acknowledgments

The authors wish to thank Kim Griffin for help on the initial draft of the paper and Karla Shelnutt and anonymous reviewers for helpful comments on earlier versions of the paper.

Authors

Glenn D. Israel, Halil I. Sari, Nicole Owens Duffy, Sebastian Galindo-Gonzalez, David C. Diehl, Saul J. Abarca Orozco, Elder Garcia Varela, and Lauren Sweeney

Procedures for Improving Self-report Measurements to Capture Behavior Change 165

Procedures for Improving Self-report Measurements to Capture Behavior Change: An Illustration

Glenn D. Israel
University of Florida

Halil I. Sari
Kilis 7 Aralik University

Nicole Owens Duffy
Sebastian Galindo-Gonzalez
David C. Diehl
University of Florida

Saul J. Abarca Orozco
Universidad Cristóbal Colón

Elder Garcia Varela
Lauren H. Sweeney
University of Florida

Programs utilizing research-tested evaluation tools can help identify effective educational strategies and document program effectiveness and impact. Using the case of the UF/IFAS Extension Family Nutrition Program (FNP), this article illustrates steps for conducting a rigorous assessment of the measurement properties of evaluation instruments. The Youth Behavior Survey (YBS) was originally developed to measure students' nutrition and physical activity behaviors before and after an educational intervention. To report FNP behavior change data under indicators for the national evaluation framework, the evaluation instrument was revised. The revision included modifying item wording to reflect national indicators and changing response options. The psychometric characteristics of the revised instrument were assessed in comparison to those of the original instrument. The main objective was to examine aspects of content and construct validity for the scores produced by the instruments. The assessment included content validity of the instrument, item discrimination, consistency of relationships in item response patterns, and change between pre-test and post-test scores. We concluded that the scores produced by the revised instrument were modestly more accurate than the original. This research suggests procedures that can be applied widely to evaluating instruments for other educational interventions.

Direct correspondence to Glenn D. Israel at gdisrael@ufl.edu

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Procedures for Improving Self-report Measurements to Capture Behavior Change 166

Keywords: measurement, reliability, validity, behavior change, nutrition, physical activity, healthy lifestyle construct, young children

Introduction

Programs utilizing research-tested evaluation tools can help identify effective educational strategies, and in turn, these strategies can be used to improve program delivery and document effectiveness. It is critical that such tools provide high-quality data for decision-making. Due to the variable context of program interventions, evaluation tools should be easy to administer, valid, and reliable, so that the final analysis and conclusions are accurate (Murphy et al., 2001). This article illustrates steps for conducting a rigorous assessment of the measurement properties of evaluation instruments that can be applied widely to other educational interventions. Because so many evaluations of educational programs use clients as the source of the data, the program selected for this illustration also relies primarily on clients' self-reported behaviors.

Using self-reports to measure outcomes of educational interventions is, however, challenging in any context. For programs targeting low-income children, such as the UF/IFAS Extension Family Nutrition Program (FNP), these challenges can be especially daunting. Collecting valid and reliable data on nutrition and health behaviors depends on developing instruments with items that are clear, well-understood, and minimize bias from acquiescence and social desirability. Because FNP delivers a sizable portion of its programming to young children in schools and it is mandated to measure behavior change with a very limited budget, FNP evaluators have relied on using self-reports of behaviors collected through group administration. This data collection method is cost-effective because it is integrated into the delivery of FNP's curricula.

In order to report FNP behavior change data under indicators for the new national *SNAP-Education Evaluation Framework*, as well as address questions about the measurement properties of the original instrument, a process to revise the instrument was undertaken. The process of revising and testing the instrument is detailed in this paper. Through this process, we have identified strengths and weaknesses of the approach used and offer several recommendations for interested readers.

Program Context

The United States Department of Agriculture (USDA) Supplemental Nutrition Assistance Program (SNAP) provides nutrition assistance to low-income families. SNAP improves the food security status of families, improves health outcomes, and decreases health care costs (Carlson & Keith-Jennings, 2018). The nutrition education component, SNAP-Education (SNAP-Ed), complements SNAP services and has the goal of increasing the likelihood that SNAP-eligible persons, including youth, will make healthy food choices and adopt physically active lifestyles, consistent with the current Dietary Guidelines for Americans (U.S. Department of Health and Human Services and U.S. Department of Agriculture, 2015) and the USDA food guidance

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Procedures for Improving Self-report Measurements to Capture Behavior Change 167

(USDA FNS, 2017). The University of Florida/Institute of Food and Agricultural Sciences Extension Family Nutrition Program (FNP) implements SNAP-Ed in Florida and conducts educational programs for low-income adults and youth in 40 of the state's 67 counties (FY17). FNP emphasizes the adoption of specific behaviors, such as eating more fruits and vegetables and increasing physical activity. FNP specifically focuses on youth in group settings, such as schools and community centers. Children continue to represent the largest proportion of individuals eligible for SNAP benefits (23% of Floridians under 18 years of age live below the federal poverty level compared to 15% of Floridians between 18 and 64 years of age) (U.S. Census Bureau, 2017). Thus, obesity prevention initiatives and nutrition education programs in the school setting have focused on the dual goals of improving health and academic outcomes.

Although nutrition education and physical activity programs in schools have been important venues for SNAP-Ed nationally, obesity rates remain high (Nanney et al., 2010). As the Healthy People 2020 objective to reduce the proportion of children aged 2-19 with obesity remains a focus, nutrition and physical activity interventions warrant serious re-assessment (Wang et al., 2012). Findings from evaluations can, in turn, be used to guide educational programs designed to improve physical activity and eating habits in young children (Branscum et al., 2010).

Evaluation Context

Programs such as SNAP-Ed depend on accurate data collection to showcase impacts and outcomes to legislators, stakeholders, and consumers. Program officers for federally-funded, multi-million dollar programs, including SNAP-Ed, have mandated assessments of outcomes and impacts (Murray et al., 2017; Wyker et al., 2012). To validate the extent of the outcomes, the measurement properties of instruments should be rigorously assessed to achieve the most reliable results and to ensure the appropriateness of evaluation conclusions (Lohr et al., 1996; Mokkink et al., 2010). A number of studies have addressed the validity of instruments measuring aspects of nutritional behaviors for adult and youth audiences (Barton et al., 2011; Edmunds & Ziebland, 2002; Hall et al., 2015; Koleilat & Whaley, 2016; Magarey et al., 2009; Murphy et al., 2001; Wilson et al., 2008). In addition, Mijnarends and colleagues (2013) described several methods for testing the reliability and validity of questionnaires.

Although numerous validity and reliability studies have been conducted with evaluation tools for adult populations, few have been conducted on evaluation tools that are used with young children. The available studies examined nutrient intake but not the variety of nutrition-related behaviors taught in nutrition education programs (Koleilat & Whaley, 2016). Moreover, Livingstone et al. (2004) observed that youths' cognitive capability is constrained and time to implement evaluation surveys is limited. Thus, it is crucial to have a short evaluation instrument with key questions (allowing self-completion) to streamline the data collection process. In addition, "each questionnaire should be tested in a group similar to that for which it has been

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Procedures for Improving Self-report Measurements to Capture Behavior Change 168

designed" (Litwin, 1995, cited in Barton et al., 2011, p. 589). The latter is particularly important because youth have different knowledge levels of nutrition, healthy eating, and physical activity.

In 2016, the final version of the *SNAP-Ed Evaluation Framework* was released to better inform and evaluate multi-year interventions through short-term, medium-term, long-term, and population-results indicators (USDA FNS, 2016) at the individual, environmental/settings, and sectors of influence levels. To align FNP's evaluation and reporting with this framework, the Youth Behavior Survey (YBS) was revised. The purpose of this study was to evaluate the measurement properties of the revised YBS, specifically content validity, construct validity, and reliability, in comparison to the original version. The specific research questions were:

1) Do the revised YBS items demonstrate greater content validity than the original YBS items?
2) Do the revised YBS items discriminate differences in behaviors better than the original YBS items?
3) Do the revised YBS items demonstrate greater internal consistency and dimensionality than the original YBS items?
4) Does the revised YBS measure behavior change better than the original YBS?

To the best of our knowledge, none of the articles previously published include the same target audience as this study – second- and third-grade children from low-income families.

Methods

In this section, the development of the evaluation instrument and data collection procedures are detailed to provide the context for the steps for assessing the measurement properties of the instruments. The development process involved the FNP evaluation team revising the original YBS following best practice techniques (Padilla & Benitez, 2014) and conducting a quasi-experiment to assess how well the revised YBS measured what it was designed to measure and whether it produced more reliable behavior scores. The methods section concludes with an explanation of the data analysis procedures used at each step in the assessment.

Youth Behavior Survey

The FNP YBS is an evaluation instrument developed and used in Florida to measure second- through fifth-grade students' nutrition- and physical activity-related behaviors via self-report. The YBS is administered as a pre-test and post-test to measure behavior change resulting from the delivered nutrition education program. The original YBS was used for several years, but it had never been psychometrically tested. The instrument's ability to measure behavior change accurately was unknown. Thus, FNP, which was reaching more than 4,000 low-income children in Florida, needed evidence-based evaluation instruments to capture behavior changes resulting from the program's interventions.

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Sample and Data Collection

Approval from the University of Florida's Institutional Review Board was obtained before recruiting participants or collecting data. All students who participated in the study were from twelve Title I elementary schools in five school districts in Florida. Data were collected from second and third-grade students during the 2015-2016 academic year. Scheduled program groups at participating schools were randomized into treatment groups (either the original or the revised YBS), and all students within a group were administered the same instrument. One school had a group of students that received the revised YBS and another group that received the original YBS. Of the other eleven schools, five had the revised version of the YBS, and six had the original version. The pre- and post-test instruments were administered to a total of 422 and 261 students for the original and revised instruments, respectively. After data cleaning and removing outliers, a total of 366 and 231 students with complete data for the original and revised instruments, respectively, were used in the analysis.

Analysis Procedures

Step 1. Assess the content validity of the instrument. Content validity ultimately rests on the judgment of those participating in the instrument development process (Selltiz et al., 1976; Vaske, 2008). A number of individuals were involved in assessing the content of the original instrument and proposing changes for the revised instrument. First, two nutrition specialists reviewed the existing items and wrote candidate items for the revised instrument. The items were also reviewed by survey experts. Finally, the set of revised items was reviewed by two additional experts, including a childhood education expert and a teacher specialist in curriculum and inclusion to evaluate whether the items were appropriate for the intended grade levels. This process of expert review established the content validity of the revised instrument relative to the original instrument.

Step 2. Examine item-level statistics. Students' answers produced by both instruments were examined for extremes in item means and poor discrimination. The item discrimination index is calculated as the correlation between given responses to an item and total scale scores after excluding that item (e.g., also known as corrected item-total correlation), and it ranges from -1 to +1. The discrimination measure indicates an item's ability to distinguish between students who perform a behavior more frequently from those who do it less frequently, and therefore, large, positive values suggest greater validity. The item-level analyses were completed with "ltm: An R Package for Latent Variable Modeling and Item Response Analysis" in *R* software Version 2.1.1 (Rizopoulos, 2006).

Step 3. Assess the consistency of relationships in the item response patterns. Next, Pearson's correlations for pairs of items and Cronbach's alpha for the set of items were calculated to examine relationships among the items for both the original and revised instrument data. Evidence of construct validity would be indicated by positive pairwise correlations and

higher values for Cronbach's alpha (Carmines & Zeller, 1979). The *R* software Version 2.1.1, specifically the "R Development Core Team, 2009-2015" was used for calculating Cronbach's alpha.

Further evidence of construct validity is obtained when dimensionality analysis conforms to the expected structure (Vaske, 2008). The dimensionality of the instruments also was tested, because both nutrition and physical activity behaviors were included in the items. Thus, a multidimensional Confirmatory Factor Analysis (CFA) model with two latent factors of nutritional behavior and physical activity was tested for the revised instrument using pre-test data. For the original instrument, a CFA model with one general factor was tested using pre-test data and is named healthy lifestyle index [Note: a two-factor model was not feasible for the original instrument because only one item was used to measure physical activity]. The CFA was completed using Mplus (Muthén & Muthén, 2012).

Step 4. Assess change between pre-test and post-test scores. The paired *t*-test was used to determine whether there was a statistically significant behavior change from pre-test to post-test for the original and revised instruments, respectively. This parametric test was used because the data met the normality assumption of paired *t*-tests. Because the paired *t*-test measures can be affected by the number of items included in the analysis, these measures were also calculated across the five common items that were described previously to allow for a fair comparison and more meaningful interpretations. The paired *t*-test was conducted in SPSS® version 23 (IBM Corp., 2015).

Finally, models predicting gain scores for the sum of the five common behaviors were analyzed to further address the question of whether the revised instrument represents an improvement in measuring behavior changes in nutrition-related behaviors. Our expectation was the revised instrument should have a modest, positive effect on the resulting gain score while controlling for the pre-test score, as well as programmatic and student factors. SAS's Proc Mixed (SAS Institute, n.d.) was used to estimate hierarchal linear models, since the student-level data were nested within groups taught by a given FNP nutrition educator (Raudenbush & Bryk, 2002).

Results

Step 1. Results of the Content Assessment

As a result of the content validity assessment, changes were made to the set of items in the instrument. The original version of the YBS consisted of seven items, and the revised version consists of nine items (see Appendices A-B). The breakfast item (i.e., item 6) from the original instrument was removed because it did not align with *SNAP-Ed Evaluation Framework* indicators used by FNP (USDA FNS, 2016). The revised YBS added more physical activity items (see items 7 to 9). Also, there were five common items in both instruments; three of the items had slight wording differences. These were the items related to vegetable, fruit, whole

Procedures for Improving Self-report Measurements to Capture Behavior Change 171

grain, and dairy food consumption, and to physical activity (i.e., items 1, 2, 4, 5, and 7 in the original instrument and items 1, 2, 3, 4, and 7 in the revised instrument, see appendix).

Changes to the revised instrument also included modifying the response options. In the original YBS, each student was asked to "Circle the answer that best applies to you" by rating how frequently they do eating or physical activities during the week. In the original YBS, the response options were "Never or almost never," "Some days," "Most days," and "Every day." In contrast, the revised YBS asks each respondent to "Think about how often you do things during the week. Then circle the answer that best applies to you." The response options in the revised instrument were "0 days," "1-3 days," "4-6 days," and "7 days." As the text above shows, the wording in the revised instrument was designed to provide more concrete response categories than the original version, and this should facilitate the response process (Dillman et al., 2014).

In summary, the evaluation team judged the content of the revised instrument as better aligned with the intended outcome measures for the *SNAP-Ed Evaluation Framework* than was the original instrument.

Step 2. Examine Item-level Statistics

The item-level statistics (mean score and item discrimination) for the pre-test scores of the original and revised scales are presented in Table 1. The mean score for an item represents the average of the item-level responses in the scale (i.e., the average frequency of the behavior). Item discrimination refers to an item's ability to differentiate students who receive a low score from those who receive a high score. A high value of discrimination shows that an item was more effective in discriminating between students who performed the behavior less frequently from those who did the behavior more frequently. In the original scale, the breakfast item had the highest mean score, and the vegetable item had the lowest mean score. For the revised scale, the item, "I am physically active," had the highest mean score, and the drinking sugary beverages item had the lowest. In the original scale, the breakfast item was the poorest discriminating item, and in the revised scale, the dairy item was the poorest discriminating item. The item-level statistics for the post-test across both instruments were given in Table 2, and the findings and interpretations were similar to the pre-test data.

In terms of the comparisons of the common five items in both scales across the pre- and post-test administrations, mean scores for items related to vegetable and fruit consumption and physical activity (physically active in the revised scale) were higher for students who completed the revised instrument. It is important to note that the item structure (e.g., item wording) for the vegetable and fruit items were the same in both scales, and the only difference was the response options on the two scales. In other words, having numerical response options resulted in higher mean scores. The mean scores for the other items in the original scale were higher. Item discriminations were higher in the revised scale with a few exceptions. The increase in

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

discrimination from the original scale to the revised scale was most obvious for the whole grain item (see Tables 1 and 2). Besides the change in response options, this was also likely due to the change in the wording (e.g., providing examples of whole grain foods). The vegetable item also showed improved discrimination on the pre-test and post-test.

Table 1. Item-level Statistics for the Pre-test Scores of the Original (n = 366) and Revised (n = 231) Instruments

	-- Item Number --		--- Mean Score ---		--- Discrimination ---	
Item	**Origina**	**Revise**	**Original**	**Revised**	**Original**	**Revised**
Vegetables	1	1	2.54	2.74	.23	.31
Fruits	2	2	3.13	3.23	.39	.40
Healthy snacks	3		2.74		.32	
Whole grains	4	3	2.62	2.41	.26	.48
Dairy	5	4	2.92	2.74	.21	.17
Breakfast	6		3.66		.12	
Plain water		5		3.41		.38
Sugary beverages*		6		2.40		.32
Physically active	7	7	3.52	3.57	.25	.25
Video games*		8		2.46		.33
Play outside		9		3.48		.31

*reverse-coded item

Table 2. Item-level Statistics for the Post-test Scores of the Original (n = 366) and Revised (n = 231) Instruments

	-- Item Number --		--- Mean Score ---		--- Discrimination ---	
Item	**Original**	**Revised**	**Original**	**Revised**	**Original**	**Revised**
Vegetables	1	1	2.67	2.84	.32	.42
Fruits	2	2	3.16	3.27	.41	.39
Healthy snacks	3		2.80		.42	
Whole grains	4	3	2.82	2.74	.34	.50
Dairy foods	5	4	3.10	3.05	.19	.19
Breakfast	6		3.69		.14	
Plain water		5		3.50		.32
Sugary beverages*		6		2.53		.33
Physically active	7	7	3.50	3.64	.27	.34
Video games*		8		2.54		.35
Play outside		9		3.52		.43

*reverse-coded item

Step 3. Assess the Consistency of Relationships in the Item Response Patterns

First, pairwise relationships between the items were examined for the original and revised instruments. As shown in Table 3, the correlations for the behaviors in the original instruments were generally very weak or weak, and one correlation was slightly negative (and substantively

Procedures for Improving Self-report Measurements to Capture Behavior Change 173

zero). Only the correlation between fruits and healthy snacks behaviors had a medium strength (Cohen, 1992).

Table 3. Bivariate Correlations Between the Items in the Original Scale for the Pre-test Data (n = 366)

Item	Vegetables	Fruits	Healthy snacks	Whole grains	Dairy foods	Breakfast	Physical activity
Vegetables	-						
Fruits	.24*	-					
Healthy snacks	.17*	.36*	-				
Whole grains	.11*	.25*	.08*	-			
Dairy foods	.05	.09*	.12*	.18*	-		
Breakfast	.04	-.00	.10*	.03	.15*	-	
Physical activity	.13*	.22*	.15*	.13*	.08	.07	-

Note: *$p \leq .05$

The correlations among items in the revised instrument showed a similar pattern of very weak or weak correlations (see Table 4). At first glance, several correlations would suggest that items are related more strongly than would be expected and others less so. A nutritional behavior, avoiding sugary beverages, is moderately correlated with a physical activity behavior, avoiding video games, while the latter has a nonsignificant correlation with being physically active and a weak correlation with playing outside. As shown below, the dimensionality analysis uncovers the nuances among the relationships in a manner consistent with evaluators' expectations.

Table 4. Bivariate Correlations Between the Items in the Revised Scale for the Pre-test Data (n = 231)

Item	Vegetables	Fruits	Whole grains	Dairy foods	Plain water	Sugary beverages	Physically active	Video games	Play outside
Vegetables	-								
Fruits	.25*	-							
Whole grains	.39*	.27*	-						
Dairy foods	.11*	.14*	.23*	-					
Plain water	.20*	.21*	.27*	.08	-				
Drink sugary beverages	.08	.18*	.22*	-.01	.22*	-			
Physically active	.20*	.20*	.11*	.08	.10	.05	-		
Play video games	.06	.17*	.23*	.02	.22*	.46*	.06	-	
Play outside	.03	.25*	.17*	.12*	.26*	.06	.30*	.18*	-

Note: Drink sugary beverages and Play video games were reverse coded; *$p \leq .05$

Next, the internal consistency of the items was examined for the original and the revised instruments. Cronbach's alpha values across all items in the original instrument were .52 and .58 for the pre-test and post-test data, respectively. For the revised instrument, the Cronbach's alpha values on the pre-test were .58 for the nutrition items, .35 for the physical activity items, and .64

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Procedures for Improving Self-report Measurements to Capture Behavior Change 174

for the whole scale. The values on the post-test were .58 for the nutrition items, .44 for the physical activity items, and .68 for the entire scale. The Cronbach's alpha values across the five common items in the original instrument were .46 and .51 for the pre- and post-test data, respectively. In the revised instrument, alpha values were .55 and .57. Overall, the revised instrument achieved a higher overall alpha than did the original instrument.

There are several reasons why Cronbach's alpha values were lower than historical cut-offs (e.g., .70 or .80; Lance et al., 2006). It is likely that the shortness of the scale contributed to lower values. Furthermore, as Wells and Wollack (2003) discussed, the criteria for the Cronbach's alpha also can depend on the importance and consequences of the test. It is acceptable to have a lower Cronbach's alpha on low-stakes and/or classroom tests because, as in this study, YBS test scores in such cases did not account for any of the students' grades (Wells & Wollack, 2003).

Finally, dimensionality analysis was used to assess the consistency of the model data with designers' expectations. In this analysis, the overall model fit is compared to established benchmarks (Chi-square $p < .05$; CFI > .90; TLI > .90; RMSEA < .08; WRMR < .08), followed by a review of the model factor loadings (Hooper et al., 2008; Kline, 2005). The confirmatory factor model fit statistics with the pre-test scores produced by all items across the two scales are shown in Table 5.

Table 5. Pre-test Model Fit Statistics for Both Original and Revised Instruments

	N	Chi-Square (*df*, *p*)	CFI	TLI	RMSEA	WRMR
Original	366	26.11 ($df = 14$, $p < .05$)	0.90	0.93	0.05	0.70
Revised	231	49.35 ($df = 25$, $p < .05$)	0.93	0.90	0.06	0.79

The data for the original scale fit the single factor model that can be named "healthy lifestyle" and indicated unidimensionality (see Figure 1). On the other hand, the data for the revised scale fit a multidimensional model, where the nutrition-related items (i.e., items 1 to 6) formed a nutrition behavior construct, and the physical activity items (i.e., items 7 to 9) formed the physical activity behavior construct (see Figure 2). Figures 1 and 2 also include the estimated factor loadings, which range from .18 to .74 and .27 to .73, respectively, for the items in the original and revised instruments. The model for the original instrument provides further evidence that the breakfast item was problematic (recall this item had very low discrimination and very weak correlations). For the revised data, the multidimensional structure supported expectations of a nutritional behavior dimension and a physical activity dimension. It is also noteworthy that the two items focused on avoiding negative behaviors were correlated (even though these were reverse coded). This represents an artifact of having both positive and negative wording for the set of items and was anticipated (see Carmines & Zeller, 1979).

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Figure 1. Confirmatory Factor Analysis of Single-factor Model for the Original Instrument Based on Pre-test Data

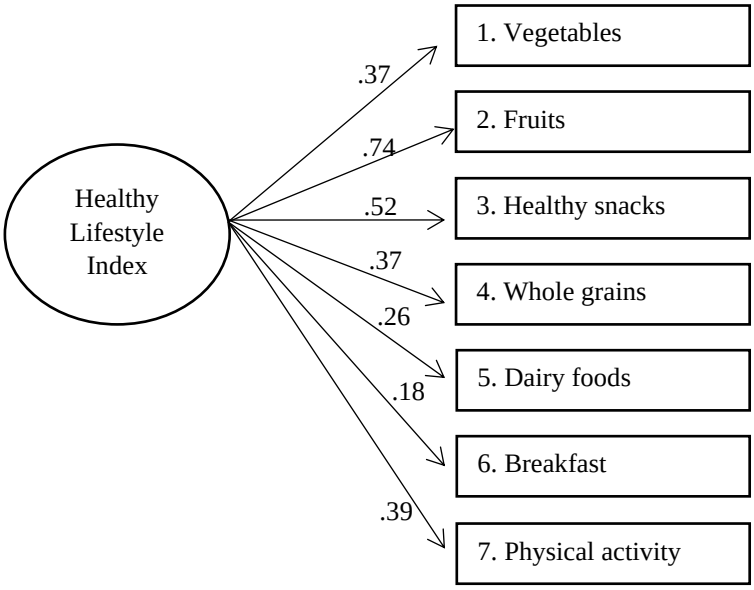
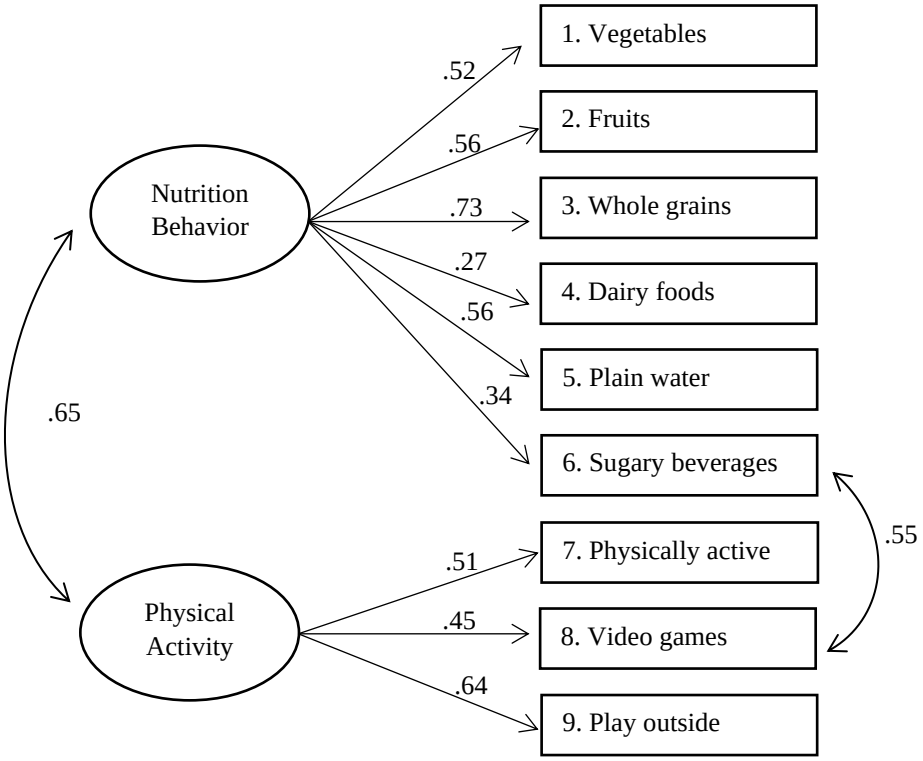


Figure 2. Confirmatory Factor Analysis of Multidimensional Model for the Revised Instrument Pre-test Data



Procedures for Improving Self-report Measurements to Capture Behavior Change 176

Step 4. Assess Change Between Pre-test and Post-test Scores

The mean and standard deviation of pre-test and post-test scores summed across all items for the students who completed the original instrument or the revised instrument are presented in Table 6, as well as results of the paired *t*-test and associated effect size. Based on the paired *t*-test, there was a positive behavior change from pre-test to post-test for both groups of students (i.e., they completed either the original instrument or the revised instrument). Even though the sample size was larger for the group that completed the original instrument, the results derived from the revised scale suggested a medium effect size as opposed to the small effect size in the original scale.

Table 6. Results of Paired t-test Across All Items

Original Scale	N	Mean	SD	t	p	Effect Size
Pre-Test	366	21.16	3.06	4.70	< .001	.24
Post-Test	366	21.77	3.11			
Revised Scale	**N**	**Mean**	**SD**	**T**	**p**	**Effect Size**
Pre-Test	231	26.46	4.63	5.64	< .001	.37
Post-Test	231	27.65	4.67			

The results of the same calculations across the five common items for both scales are presented in Table 7. We found similar results except for the effect size comparisons, in which the revised instrument produced a larger effect size than the original one. These findings suggest that the revised instrument was more effective in measuring behavior change, especially when additional items were added.

Table 7. Results of Paired t-test Across the Five Common Items

Original Scale	N	Mean	SD	t	p	Effect Size
Pre-Test	366	14.75	2.50	4.74	< .001	.24
Post-Test	366	15.28	2.51			
Revised Scale	**N**	**Mean**	**SD**	**T**	**p**	**Effect Size**
Pre-Test	231	14.70	3.02	4.77	< .001	.30
Post-Test	231	15.55	2.94			

Finally, differences in behavior changes between pre- and post-test administration was examined using gain scores (i.e., the difference between the pre- and post-test scores) on the five common items for both the original and revised instruments. The change model focused on the effect of the dummy variable "Revised instrumentation," which was coded "1" for students receiving the revised instrument and "0" for students having the original instrument. The model also included the pre-test score to control for a student's initial position, as well as attributes of the program, nutrition educator, and the student. SAS's Proc Mixed was used to estimate a hierarchal linear model, since the student-level data were nested within groups taught by a given FNP nutrition

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Procedures for Improving Self-report Measurements to Capture Behavior Change 177

educator. Of the total variation in gain scores, 8.1% was between groups taught by different educators, and 91.9% was between students within the groups.

The results in Table 8 show that the parameter estimate for the revised instrument was positive (as expected) but not statistically significant (p = .167) after controlling for the pre-test score, students' sex, program duration, and nutrition educator attributes. The results reported in the analyses above indicated that the revised instrument had slightly better measurement properties, and this should be reflected in reduced attenuation. However, the effect of the revised instrument was weak and nonsignificant (the small sample size in the treatment group contributed to this finding).

Regarding the control variables, the pre-test score had a significant negative effect on the gain score, which is a common and expected result in change models known as regression to the mean (Barnett et al., 2005). Also, at the student level (i.e., level 1), female students made slightly more behavioral changes than male students (p = .069). At the school/nutrition educator level (i.e., level 2), the education level of the nutrition educators had a large effect on the students' gain scores, with those having a nutrition educator who had an Associate's degree scoring 1.7 behavior units higher than students who had a nutrition educator with a Bachelor's degree and 1.8 units higher than those who had a nutrition educator with a high school diploma or GED. After accounting for other variables, nutrition educator years of experience and program duration had little effect on gain scores.

Table 8. Regression of Behavior Gain Score on Pre-program Behavior Score, Student Attributes, Program Attributes, and Instrumentation (n = 597)

Variable	Estimate	Standard Error	DF	*t*	*p*
Intercept	6.874				
Pre-test score	-.401	.032	547	-12.69	< .001
Revised instrumentation	.320	.228	42	1.41	.167
Female student	.319	.171	47	1.86	.069
Program duration (weeks)	-.070	.058	42	-1.21	.233
Educator experience (years)	-.155	.093	42	-1.66	.105
Highest educator degree					
High school diploma or GED	-.113	.433	42	-.26	.795
Associate's degree	1.736	.593	42	2.93	.001
Bachelor's degree	.000	--			

Overall, the hierarchal linear model was effective in accounting for a sizable amount of the variance in test scores. Of the variance between nutrition educators (i.e., level 2), the model accounted for 78.3% compared to the null model with no predictors. A smaller amount, 20.5%, of the variance was accounted for within classrooms (i.e., between students at level 1). The overall model -2 Log Likelihood decreased from 2722.2 in the null model to 2563.6 in the

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

fitted model in Table 8; likewise, the Akaike Information Criteria (AIC) decreased from 2728.2 in the null model to 2583.6 in the fitted model.

Discussion and Conclusions

This study illustrates steps for conducting a rigorous assessment of the measurement properties of evaluation instruments that can be applied widely to other educational interventions. Programs utilizing research-tested evaluation tools can better measure outcomes, and in turn, identify effective educational strategies. Such tools can also provide high-quality data for documenting program impact. In addition, federal and state agencies have been requiring more rigorous measurement of outcomes to ensure that programs, such as FNP, are effective and efficient. In the case of FNP, evaluators were charged with updating evaluation instruments to meet program accountability expectations and capture data for national indicators. The present study contributed to the development of more rigorous measurement of FNP program outcomes. Additionally, the findings provide guidance on instrument development and the impact of response categories on validity. This study also demonstrates a method for developing an evidence-based tool for youth-specific nutrition education programs beyond SNAP-Ed.

This study, which attempted to improve the response accuracy and align the scale with the *SNAP-Ed Evaluation Framework* indicators, was largely successful based on comparisons of psychometric characteristics of the original and revised instruments. In addition, quantitative measures indicated that both instruments measured the construct of interest that can be named "healthy lifestyle," but the revised instrument did so in a way that is conceptually consistent, with nutrition items and physical activity items loading on separate factors that together form the broader construct. In addition, most items in the revised scale were more effective in distinguishing the students with higher total scores from students with lower total scores. It can also be concluded that having numerical response options provided more precise total scores than response options using vague qualifiers (Dillman et al., 2014), while acknowledging changing the response option was not the only revision made for several of the items in the original scale. However, when the identical items in both instruments (e.g., vegetable and fruit items) were closely examined, the same conclusions were reached.

Furthermore, regardless of the number of items included in the analysis (all items vs. five common items) and the instrument used to collect the data, the mean scores always increased from pre- to post-test. Thus, it is safe to conclude that there is evidence of behavior change due to the intervention. Additionally, the revised instrument always produced the larger effect size than the original one, and this was more apparent with the addition of items into the revised instrument. It should be noted that the breakfast and dairy food items in the original scale and the dairy item in the revised scale had poor discrimination power. So, consideration should be given to revising or removing these items from the instrument. On the other hand, reducing the number of items can adversely affect the overall reliability of the instrument, as measured by

Procedures for Improving Self-report Measurements to Capture Behavior Change 179

Cronbach's alpha, and adding or revising items further could increase reliability. Expanding the instrument is likely to meet resistance from nutrition educators and schoolteachers because this can increase the time needed for administering the evaluation in already-tight schedules.

The analysis of change scores failed to show an improved performance for the revised instrument. The results for the HLM analysis suggested the revised instrument with a larger sample size might capture more behavior change. This is because a more valid (and accurate) instrument will have less attenuation in relationships being examined (Bohrnstedt, 1969). Stronger evidence of construct validity was shown by the dimensionality analysis, where the two-factor model for the revised instrument aligned with evaluators' expectations (Carmines & Zeller, 1979). Consequently, the cumulative evidence supports the view that the revised instrument provides modestly higher quality data for evaluating SNAP-Ed youth programming, but there is still room for improvement.

Finally, this study illustrates a number of steps and substeps for evaluating the measurement properties of instruments that can be applied widely by evaluators. Depending on the scale and scope of the educational program, as well as the expertise of the program team, the full set or a subset may be more practical. Whatever the situation, it is important to conduct some analysis of the evaluation instruments' ability to capture program outcomes.

Limitations

Developing tools that measure nutrition and physical activity behaviors for low-income youth is a challenging area, and additional rigorous research is needed. Effective evaluation tools with this target population are difficult to construct. Although the YBS tested in this study provided some evidence that it was a psychometrically robust tool that measures dietary behavior in low-income youth, the sample size for the group with the revised tool was smaller than intended. A larger sample size would have increased statistical power for the regression model used in the analysis. In addition, researchers should conduct cognitive interviews and focus group discussions with children to improve the items and the measurement scales in this instrument in the future. As mentioned by Branscum et al. (2010), the use of both qualitative and quantitative methods by researchers can help improve the evaluation tools. Multi-method or mixed-method studies of children's dietary behaviors will help create more robust evaluation tools for low-income audiences.

References

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Barton, K. L., Wrieden, W. L., & Anderson, A. S. (2011). Validity and reliability of a short questionnaire for assessing the impact of cooking skills interventions. *Journal of Human*

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Nutrition and Dietetics, *24*(6), 588–595. https://doi.org/10.1111/j.1365-277X.2011.01180.x

Bohrnstedt, G. W. (1969). Observations in the measurement of change. *Sociological Methodology*, *1*, 113–133. doi:10.2307/270882

Branscum, P., Sharma, M, Kaye, G., & Succop, P. (2010). An evaluation of the validity and reliability of a food behavior checklist modified for children. *Journal of Nutrition Education and Behavior*, *42*(5), 349–352. https://doi.org/10.1016/j.jneb.2009.12.005

Carlson, S., & Keith-Jennings, B. (2018). *SNAP is linked with improved nutritional outcomes and lower health care costs.* Center on Budget and Policy Priorities. https://www.cbpp.org/sites/default/files/atoms/files/1-17-18fa.pdf

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage. https://dx.doi.org/10.4135/9781412985642

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed mode surveys: The tailored design method* (4th ed.). John Wiley and Sons.

Edmunds, L. D., & Ziebland, S. (2002). Development and validation of the Day in the Life Questionnaire (DILQ) as a measure of fruit and vegetable questionnaire for 7-9 year olds. *Health Education Research*, *17*(2), 211–220. https://doi.org/10.1093/her/17.2.211

Hall, E., Chai, W., Koszewski, W., & Albrecht, J. A. (2015). Development and validation of a social cognitive theory-based survey for elementary nutrition education program. *International Journal of Behavioral Nutrition and Physical Activity*, *12*(1), 47. https://doi.org/10.1186/s12966-015-0206-4

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60. http://www.ejbrm.com/vol6/v6-i1/v6-i1-papers.htm

IBM Corp. (2015). *IBM SPSS Statistics for Windows, Version 23.0*.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. Guilford Press.

Koleilat, M., & Whaley, S. E. (2016). Reliability and validity of food frequency questions to assess beverage and food group intakes among low-income 2- to 4-year-old children. *Journal of Academy of Nutrition and Dietetics*, *116*(6), 931–939. https://doi.org/10.1016/j.jand.2016.02.014

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: what did they really say? *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Litwin, M. S. (1995). *How to measure survey reliability and validity*. Sage. https://dx.doi.org/10.4135/9781483348957

Livingstone, M. B. E., Robson, P. J., & Wallace, J. M. W. (2004). Issues in dietary intake assessment of children and adolescents. *British Journal of Nutrition*, *92*, S213–S222. https://doi.org/10.1079/bjn20041169

Procedures for Improving Self-report Measurements to Capture Behavior Change 181

Lohr, K. N., Aaronson, N. K., Alonso, J., Burnam, M. A., Patrick, D. L., Perrin, E. B., & Roberts, J. S. (1996). Evaluating quality-of-life and health status instruments: Development of scientific review criteria. *Clinical Therapeutics, 18*(5), 979–992. https://doi.org/10.1016/s0149-2918(96)80054-3

Magarey, A., Golley, R., Spurrier, N., Goodwin, E., & Ong, F. (2009). Reliability and validity of the Children's Dietary Questionnaire; A new tool to measure children's dietary patterns. *International Journal of Pediatric Obesity*, *4*(4), 257–265. https://doi.org/10.3109/17477160902846161

Mijnarends, D. M., Meijers, J. M. M, Halfens, R. J. G., ter Borg, S., Luiking, Y. C., Verlaan, S., Schoberer, D., Cruz Jentoft, A. J., van Loon, L. J., & Schols, J.M.G.A. (2013). Validity and reliability of tools to measure muscle mass, strength, and physical performance in community-dwelling older people: A systematic review. *Journal of the American Medical Directors Association*, *14*(3), 170–178. https://doi.org/10.1016/j.jamda.2012.10.009

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

Murphy, S. P., Kaiser, L. L., Townsend, M. S., & Allen, L. H. (2001). Evaluation of validity of items for a food behavior checklist. *Journal of the American Dietetic Association*, *101*(7), 751–761. https://doi.org/10.1016/S0002-8223(01)00189-4

Murray, E.K., Auld, G., Baker, S.S., Barale, K., Franck, K, Khan, T., Palmer-Keenan, D., & Walsh, J. (2017). Methodology for developing a new EFNEP food and physical activity behaviors questionnaire. *Journal of Nutrition Education and Behavior*, *49*(9), 777–783.E1. https://doi.org/10.1016/j.jneb.2017.05.341

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Nanney, M. S., Nelson, T., Wall, M., Haddad, T., Kubik, M., Laska, M.N., & Story, M. (2010). State school nutrition and physical activity policy environments and youth obesity. *American Journal of Preventive Medicine*, *38*(1), 9–16. https://doi.org/10.1016/j.amepre.2009.08.031

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

R Development Core Team. (2009-2015). *R: A language and environment for statistical computing, reference index* (Version 2.2.1). Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software, 17*(5), 1–25. http://hdl.handle.net/10.18637/jss.v017.i05

SAS Institute Inc. (n.d.). *Software, Version 9.4*. SAS Institute Inc.

Selltiz, C., Wrightsman, L. S., & Cook, S. W. (1978). *Research methods in social relations* (3rd ed.). Holt, Rinehart, and Winston.

U.S. Census Bureau. (2017). *2012-2016 American Community Survey 5-year estimates*. https://www.census.gov/newsroom/press-kits/2017/acs-5-year.html

U.S. Department of Agriculture Food and Nutrition Service [USDA FNS]. (2016). *SNAP-ED Evaluation framework and interpretive guide*. https://snaped.fns.usda.gov/program-administration/snap-ed-evaluation-framework

U.S. Department of Agriculture Food and Nutrition Service [USDA FNS]. (2017). *Supplemental nutrition assistance program education. Plan guidance FY 2018. Nutrition education and obesity prevention program*. https://snaped.fns.usda.gov/snap/Guidance/FY2018SNAP-EdPlanGuidance.pdf

U.S. Department of Health and Human Services and U.S. Department of Agriculture. (2015). *2015 – 2020 Dietary guidelines for Americans* (8th ed.). https://health.gov/dietaryguidelines/2015/guidelines/

Vaske, J. J. (2008). *Survey research and analysis: Applications in parks, recreation and human dimensions*. Venture Publishing, Inc.

Wang, Y. C., Orleans, C. T., & Gortmaker, S. L. (2012). Reaching the healthy people goals for reducing childhood obesity: Closing the energy gap. *American Journal of Preventive Medicine, 42*(5), 437–444. https://doi.org/10.1016/j.amepre.2012.01.018

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. University of Wisconsin, Testing & Evaluation Services. https://testing.wisc.edu/Reliability.pdf

Wilson, A. M., Magarey, A. A., & Mastersson, N. (2008). Reliability and relative validity of a child nutrition questionnaire to simultaneously assess dietary patterns associated with positive energy balance and food behaviours, attitudes, knowledge and environments associated with healthy eating. *International Journal of Behavioral Nutrition and Physical Activity*, *5*, 5. https://doi.org/10.1186/1479-5868-5-5

Wyker, B. A., Jordan, P., & Quigley, D. L. (2012). Evaluation of Supplemental Nutrition Assistance Program Education: Application of behavioral theory and survey validation. *Journal of Nutrition Education and Behavior*, *44*(4), 360–364. https://doi.org/10.1016/j.jneb.2011.11.004

Glenn Israel is Professor and Graduate Coordinator, Department of Agricultural Education and Communication, and Evaluation Specialist, Program Development and Evaluation Center at the University of Florida.

Procedures for Improving Self-report Measurements to Capture Behavior Change 183

Halil Sari is Assistant Professor in the Department of Educational Sciences, Muallim Rifat College of Education at the Kilis 7 Aralik University, Turkey.

Nicole Duffy is State Specialized Agent, Department of Family, Youth and Community Sciences, and Program Coordinator, EFNEP, at the University of Florida.

Sebastian Galindo is Research Assistant Professor, Department of Agricultural Education and Communication, and Co-director of the Family Nutrition Program Evaluation Team, University of Florida.

David Diehl is Associate Professor, Department of Family, Youth and Community Sciences, and Co-director of the Family Nutrition Program Evaluation Team, University of Florida.

Saul Abarca is a former Family Nutrition Program Post Doc on the Family Nutrition Program Evaluation Team.

Elder Garcia is a Doctoral Student and Graduate School Fellow, Nutrition Education & Behavioral Science Laboratory, Department of Health Education and Behavior at the University of Florida.

Lauren Sweeney is a former Education and Training Specialist, Department of Family, Youth and Community Sciences at the University of Florida.

Acknowledgements

The authors wish to thank Kim Griffin for help on the initial draft of the paper and Karla Shelnutt and anonymous reviewers for helpful comments on earlier versions of the paper.

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Procedures for Improving Self-report Measurements to Capture Behavior Change 184

Appendix A

Original Instrument

ID_____

Are you a boy or a girl? (circle) Boy Girl

Circle the answer that best applies to you.

1. I eat vegetables…	Never or almost never	Some days	Most days	Every day
2. I eat fruit…	Never or almost never	Some days	Most days	Every day
3. I choose healthy snacks…	Never or almost never	Some days	Most days	Every day
4. I eat whole grain foods…	Never or almost never	Some days	Most days	Every day
5. I eat or drink low-fat or fat-free dairy foods…	Never or almost never	Some days	Most days	Every day
6. I eat breakfast…	Never or almost never	Some days	Most days	Every day
7. I do physical activities…	Never or almost never	Some days	Most days	Every day

☐ PRE-Survey ☐ POST-survey

Journal of Human Sciences and Extension Volume 8, Number 2, 2020

Procedures for Improving Self-report Measurements to Capture Behavior Change 185

Appendix B

Revised Instrument

ID_____ ☐ PRE-Survey ☐ POST-survey

Are you a boy or a girl? (circle) Boy Girl

Think about how often you do things during the week. Then circle the answer that best applies to you.

1. I eat vegetables…	0 days	1-3 days	4-6 days	7 days
2. I eat fruit…	0 days	1-3 days	4-6 days	7 days
3. I eat whole grain foods… (like whole wheat bread, oatmeal, brown rice)	0 days	1-3 days	4-6 days	7 days
4. I drink low-fat (1%) or skim milk…	0 days	1-3 days	4-6 days	7 days
5. I drink plain water…	0 days	1-3 days	4-6 days	7 days
6. I drink sugary beverages… (like soda, fruit drinks, or sports drinks)	0 days	1-3 days	4-6 days	7 days
7. I am physically active…	0 days	1-3 days	4-6 days	7 days
8. I play video games…	0 days	1-3 days	4-6 days	7 days
9. I play outside…	0 days	1-3 days	4-6 days	7 days

Journal of Human Sciences and Extension Volume 8, Number 2, 2020