

## Impact Evaluation Measures: Definitions and Methodologies

### Definitions of Key Terms

**Validity:** *This is a general term meaning ‘accuracy’ of the question responses. Does the instrument measure what it is intended to measure? If it does, then we would say that the instrument is a valid measure with this audience. There are various types of validity and each type takes a somewhat different approach to assessing the accuracy of an instrument. Here are some useful types of validity for this youth evaluation project.*

1. Content validity is the extent to which the questions on the instrument cover the full range of meaning for the concept being measured. The content validity of an instrument is determined by a group of experts in the field of nutrition science, human development, EFNEP and FSNE.
2. Face validity determines “on the face of it,” This type of validity rests on the judgment of the developer and the clients (usually during a pilot test of the instrument).
3. Criterion and convergent validity both relate to the predictive ability of an instrument/measure. With criterion validity, the performance or outcome that an instrument/measure is designed to predict is called a criterion. The validity of the criterion must be established because it is the standard by which the new instrument/measure is validated. Convergent validity examines whether an instrument/measure correlates in a predicted manner with variables that theoretically it should correlate with.

**Reliability:** *This is a general term and refers to ‘consistency’ of responses to the questions.*

1. *The first type of Reliability refers to the stability of the questions. Stability focuses on repeated administration of the question with the same clients when no nutrition education experience is present. Does the question elicit the same response from youth each time it is asked? If it does, then we would say that the instrument is a reliable question with our low-income audience.*
2. *A second type of reliability, internal consistency, focuses on the extent to which clients respond the same or very similar to different items measuring the same domain (eg, fruit and vegetable behavior or goal setting knowledge or goal setting self-efficacy) on the instrument/measure.*
3. Sensitivity is the extent to which values on the instrument/measure change when there is a change or difference in what is being measured.

Note: Definitions developed by Dr. Lisa A. Guion, Associate Professor, Department of Family, Youth and Community Sciences, University of Florida; and revised by Dr. Marilyn Townsend, University of California-Davis. (2006)

## **References**

Babbie, Earl. 2001. *The Practice of Social Research, Ninth Edition*. Belmont, CA: Wadsworth Publishing Company.

Rossi, P.R.; Lipsey, M.W. & Freeman, H.E. (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage Publications.

Litwin MS. *How to Measure Survey Reliability and Validity*. Thousand Oaks, California: Sage Pub; 1995.

Pedhazur EJ, Schmelkin LP. *Measurement, Design, and Analysis: An Integrated Approach*. New Jersey: Lawrence Erlbaum Assoc., 1991.

Nunnally JC, Bernstein IH. *Psychometric Theory*, 3rd ed., New York: McGraw-Hill, Inc. 1994.

## **Considerations when Assessing Tools**

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

	<b>Who is involved?</b>	<b>What?</b>	<b>When?</b>	<b>Cost †</b>
<b>Validity – Development of items</b>				
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.	During 1 <sup>st</sup> stage	\$
Face	Clients	Matches wording of test items to vocabulary of client.	During 1 <sup>st</sup> stage	\$\$
<b>Validity – Testing of items ‡</b>				
Construct	Clients	Reserve for those scales for which there is no objective measures (eg, attitudes, beliefs)	Throughout	\$\$-\$\$\$\$
Convergent	Clients	Determines links to diet	After item pool/scales in place	\$\$\$\$
Criterion	Clients	Determines links to health	After item pool/scales in place	\$\$\$\$
<b>Reliability ‡</b>				
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	\$\$
<b>Other Tests ‡</b>				
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 post intervention.

**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 <b>Domain selections</b>	Using peer-reviewed published research on chronic disease, select appropriate content/domains and their corresponding behaviors.	Essential	Content validity
#2 <b>Item generation</b>	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 <b>Item pre-testing</b>	Review wording of each item with members of various FSNE audiences. Using individual interviews and standardized protocol, ask client what the item means to her using her own words. Clarify meaning of key words.	Essential	Face validity
#4 <b>Item testing &amp; analyses</b>	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.	Advisable, often done.	Item difficulty index
	Administer test to clients at two time points without the curriculum. We want clients to respond the same way at each time point.	Advisable, sometimes done.	Temporal reliability (stability)
	Examine performance of each scale for internal consistency.	Advisable, sometimes done.	Internal consistency
#5 <b>Convergent &amp; criterion validity</b>	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?	More difficult and costly than other aspects of evaluation.	<u>Convergent</u> validity if use 24-hr recall as a surrogate for actual diet.
		Advisable, but rarely done.	<u>Criterion</u> validity if we use an external measure for health such as a biomarker (eg, a serum level that indicates nutrient intake.)
#6 <b>Sensitivity</b>	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.

Source: Marilyn Townsend, University of California, 2005  
 Published in Journal of Nutrition Education and Behavior, Volume 38 Number 1, Jan/Feb 2006,  
 page 18.