Criteria and Checklist for Measure Development Papers

Grayson N. Holmbeck and Katie A. Devine

Revised 4/30/09

□ 1. Establishes Scientific Need for the Instrument

- □ a. Reviews research and/or clinical practices to establish need for the instrument
- □ b. Specifies the new contribution of the measure relative to previous research

□ 2. Attends to Content Validity During Initial Measure Development (based on Clark & Watson, 1995; Haynes, Richard, & Kubany, 1995; Haynes, Nelson & Blaine, 1999)

- □ a. Defines the construct
  - □ i. Reviews theory underlying the construct
  - □ ii. Specifies what will be included and excluded in the measure
  - □ iii. Specifies facets or dimensions of construct
- □ b. Specifies contexts/situations for the measure
  - □ i. Specifies setting for completion of measure
- □ c. Specifies intended function of the measure
  - □ i. Specifies purpose of measure
  - □ ii. Specifies target population
  - □ iii. Specifies appropriate age range
  - □ iv. Determines if appropriate for multiple developmental levels and ethnic groups
- □ d. Selects and generates items based on:
  - □ i. Clinical experience
  - □ ii. Relevant theories
  - □ iii. Empirical literature
  - □ iv. Rational deduction
  - □ v. Related Instruments
  - □ vi. Consultation with experts
  - □ vii. Focus groups with target population
- □ e. During item generation, matches items to facets/dimensions
  - □ i. Includes appropriate numbers of items for each dimension
  - □ ii. Attends to test length (generates an appropriate number of items given the setting in which it will be used, generates enough items to allow for some items to be dropped

during the test refinements process, generates enough items to assess the construct)

- □ f. Conducts qualitative item analysis (relevance of each item, wording of items, check for redundancy across items)
- □ g. Addresses literacy and reading level issues for the target population
- □ h. Determines response format and scoring method
  - □ i. Selects response format (e.g., Likert, etc)
  - □ ii. Attempts to reduce impact of response sets by not wording all items in same direction
  - □ iii. Scoring method is explained
- □ i. Develops appropriate instructions for measure (including time frame; e.g., "During the past two weeks…")
- □ j. Has experts review the initial version of the instrument
- □ k. Has members of target population review initial version of the instrument
- □ l. After refinement of measure:
  - □ i. Additional item analysis
  - □ ii. Additional review by experts
  - □ iii. Additional review by members of target population
- □ m. Conducts pilot testing of measure

□ 3. <u>Evaluation of Reliability</u>

- □ a. Evaluates internal consistency (subscales, full scale)
- □ b. Evaluates temporal stability (test-retest)
- □ c. Uses generalizability theory in assessing reliability
- □ d. Cross-validates reliability estimates

□ 4. <u>Develops Norms for the Measure</u>

- □ a. Develops norms for different relevant populations

□ 5. <u>Quantitative Item Analysis</u>

- □ a. Examines whether items discriminate between relevant groups
- □ b. Includes corrected item-to-total correlations
- □ c. Includes average correlations between individual items and all other items
- □ d. Evaluates distributions of items and eliminates items with inadequate distributions
- □ e. Evaluates items using Item Response Theory (particularly if it is a measure that assesses abilities or skills)

□ i. Examines item characteristic curves (see Nunnally & Bernstein, 1994)

□ ii. Examines unidimensionality of items, the appropriateness of using a total summary score, and differential item functioning using Rasch analysis (Tennant, McKenna, & Hagell, 2004; Tesio, 2003)

□ 6. Conducts Factor Analyses

□ a. Evaluates factor structure of measure via exploratory factor analyses/principal components analyses

□ b. Confirms hypothesized factor structure of measure via confirmatory factor analyses

□ 7. Evaluation of Validity

□ a. Clearly articulates plan for assessing validity

□ b. Includes a priori hypotheses for major analyses

□ c. Evaluates overall construct validity of measure (which involves a general evaluation of all validity evidence for the measure)

□ d. Evaluates convergent validity, which is the degree of convergence between the target measure and other instruments purporting to measure the same construct

□ e. Evaluates discriminant validity, which is the degree to which the target measure is not associated with other measures that assess different constructs

□ f. Evaluates criterion-related validity, which is the degree to which scores on the target measure are associated with measures of non-test behaviors (includes concurrent and predictive validity)

□ g. Cross-validates validity estimates

□ 8. Evaluates Diagnostic Utility, Clinical Utility, and Cost-Effectiveness (based on Haynes et al., 1999)

□ a. Evaluates degree of treatment utility

□ i. Is the measure sensitive to change?

□ ii. Can it be used repeatedly over the course of treatment and does it reflect improvement or worsening of symptoms? (see Kazdin, 2005)

□ b. Evaluates degree of diagnostic utility (see Bossuyt et al., 2003)

□ i. Includes estimates of diagnostic accuracy (sensitivity, specificity, positive and negative predictive power)

- □ c. Evaluates degree of incremental validity (does the measure add value in clinical judgment above and beyond other measures?)
- □ d. Evaluates measure’s cost-effectiveness

□ 9. <u>Translates Measure into Other Languages</u>

- □ a. Semantic equivalence: Translation by a native speaker and back-translation by an independent native speaker. Region-specific language should be used when possible
- □ b. Content equivalence: Native language speaker has reviewed content of items for appropriateness and equivalence
- □ c. Technical equivalence: All language versions contain the same item and scale formatting

References

Bossuyt, P. M. et al. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Annals of Internal Medicine, 138,* W1-W12.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309-319.

Haynes, S. N., Nelson, K., & Blaine, D. D. (1999). Psychometric issues in assessment research. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed.; pp. 125-154). New York: Wiley.

Haynes, S. N., Richard, D. C. S, & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238-247.

Kazdin, A. E. (2005). Evidence-based assessment: Issues in measure development and clinical application. *Journal of Clinical Child and Adolescent Psychology, 34,* 548-558.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Tennant, A., McKenna, S. P., & Hagell, P. (2005). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7,* S22-S26.

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine, 35,* 105-115.