

OUTCOME MEASURES RATING FORM GUIDELINES

FROM: CanChild Centre for Childhood Disability Research
Institute of Applied Health Sciences, McMaster University
1400 Main Street West. Room 408
Hamilton, Canada L8S 1C7
fax (905) 522-6095
lawm@mcmaster.ca

PREPARED BY: Mary Law, Ph.D. O.T.(C)

FOR FURTHER DISCUSSION OF ISSUES: Law, M. (1987). Measurement in occupational therapy: Scientific criteria for evaluation. *Canadian Journal of Occupational Therapy*, 54, 133-138.

GENERAL INFORMATION: Name of Measure, Authors, Source and Year.

1. FOCUS

- a. FOCUS OF MEASUREMENT. Use the ICF framework to indicate the focus of the measurement instrument that is being reviewed. The definitions are as follows: **BODY FUNCTIONS:** are the physiological functions of body systems (including psychological functions). **BODY STRUCTURES:** are anatomical parts of the body such as organs, limbs and their components. **ACTIVITIES AND PARTICIPATION:** activity is the execution of a task or action by an individual. Participation is involvement in a life situation. **ENVIRONMENTAL FACTORS:** make up the physical, social and attitudinal environment in which people live and conduct their lives.
- b. ATTRIBUTE(S) BEING MEASURED. The rating form lists attributes organized using the ICF framework. Check as many attributes as apply to indicate what is being measured by this instrument.
- c. SINGLE OR MULTIPLE ATTRIBUTE. Check the appropriate box to indicate whether this measure assesses a single attribute only or multiple attributes.
- d. List the PRIMARY PURPOSE for which the scale has been designed. Secondary purposes can also be listed but the instrument should be evaluated according to its primary purpose (i.e., discriminative, predictive, evaluative).
DISCRIMINATIVE: A discriminative index is used to distinguish between individuals or groups on an underlying dimension when no external criterion or gold standard is available for validating these measures.
PREDICTIVE: A predictive index is used to classify individuals into a set of predefined measurement categories... either concurrently or prospectively, to determine whether individuals have been classified correctly.

EVALUATIVE: An evaluative index is used to measure the magnitude of longitudinal change in an individual or group on the dimension of interest.

(Kirshner, B. & Guyatt G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases*, 38, 27-36.)

- e. PERSPECTIVE. Indicate the possible respondents.
- f. POPULATION for which it is designed (AGE). If no age is stated, mark as age unspecified. List the diagnostic groups for which the measure is used.
- g. EVALUATION CONTEXT refers to the environment in which the assessment is completed. Check all possible environments in which this assessment can be completed.

2. CLINICAL UTILITY

- a. CLARITY OF INSTRUCTIONS. Check one of the ratings. Excellent: clear, comprehensive, concise and available; Adequate: clear, concise but lacks some information; Poor: not clear and concise or not available.
- b. FORMAT. Check all applicable items to indicate the format of data collection for the instrument. Possible items include naturalistic observation, interview, a questionnaire (self-completed, interview administered or caregiver-completed) and task performance.
PHYSICALLY INVASIVE indicates whether administration of the measure requires procedures which may be perceived as invasive by the client. Examples of invasiveness include any procedure which requires insertion of needles or taping of electrodes, or procedures which require clients to take clothing on or off.
ACTIVE PARTICIPATION OF CLIENT. Indicate whether completion of the measure requires the client to participate verbally or physically.
SPECIAL EQUIPMENT REQUIRED. Indicate whether the measurement process requires objects which are not part of the test kit and are not everyday objects. Examples of this include stopwatches, a balance board or other special equipment.
- c. TIME TO COMPLETE THE ASSESSMENT. Record in minutes. For ADMINISTRATION, SCORING and INTERPRETATION, consider the time and the amount of training and the ease with which a test is administered, scored and interpreted, and indicate whether these issues are easy or more complex. For ADMINISTRATION, SCORING and INTERPRETATION to be rated as easy, each part of the task should be completed in under one hour with minimal amount of training and is easy for the average service provider to complete.
- d. EXAMINER QUALIFICATIONS. Indicate if formal training is required for administering and interpreting this measure.

- e. **COST.** In Canadian funds, indicate the cost of the measurement manual and score sheets. For **SCORE SHEETS**, indicate the number of sheets obtainable for that cost. List the **SOURCE** and the **YEAR** of the cost information so readers will know if the information is up to date.

3. SCALE CONSTRUCTION

- a. **ITEM SELECTION.** Check one of the ratings. Excellent: included all relevant characteristics of the attribute based on comprehensive literature review and survey of experts - a comprehensive review of the literature only is enough for an excellent rating, but a survey of experts alone is not enough; Adequate: included most relevant characteristics of the attribute; Poor: convenient sample of characteristics of the attribute.
- b. **WEIGHTING.** Indicate whether the items in the tool are weighted in the calculation of the total score. If items are weighted, indicate whether the authors have weighted these items implicitly or explicitly. Implicit weighting occurs when there are a number of scales and each have a different number of items and the score is obtained by simply adding the scores for each item together. Explicit weighting occurs when each item or score is multiplied by a factor to weight its importance.
- c. **LEVEL OF MEASUREMENT.** State whether the scale used is **NOMINAL** (descriptive categories), **ORDINAL** (ordered categories), or **INTERVAL** or **RATIO** (numerical) for single and for summary scores. Indicate the **SCALING METHOD** that was used and the **NUMBER OF ITEMS** in the measure. Indicate if **SUBSCALE SCORES** are obtained. Indicate whether the subscales can be administered alone and the scores interpreted alone. In some cases, the scores can be interpreted alone, but the whole measure must be administered first. List the subscales with the number of items and indicate if there is evidence of reliability and validity for the subscales so that the scores can be used on their own. *Standardization* is the process of administering a test under uniform conditions.

4. STANDARDIZATION

- a. **MANUAL.** Check one of the ratings. Excellent: published manual which outlines specific procedures for administration; scoring and interpretation; evidence of reliability and validity. Adequate: manual available and generally complete but some information is lacking or unclear regarding administration; scoring and interpretation; evidence of reliability and validity. Poor: no manual available or manual with unclear administration; scoring and interpretation; no evidence of reliability and validity.
- b. **NORMS.** Indicate whether norms are available for the instrument. Please note that instruments which are only meant to be evaluative do not require norms. Indicate all **AGES** for which norms are available, the **POPULATIONS** for which the measure has been normed (e.g., children with cerebral palsy, people with spinal cord injuries), and indicate the **SIZE OF THE SAMPLE** which was used in the normative studies.

5. RELIABILITY

Reliability is the process of determining that the test or measure is measuring something in a reproducible and consistent fashion.

- a. RIGOUR OF STANDARDIZATION STUDIES FOR RELIABILITY. Excellent: More than 2 well-designed reliability studies completed with adequate to excellent reliability values; Adequate: 1 to 2 well-designed reliability studies completed with adequate to excellent reliability values; Poor: No reliability studies or poorly completed, or reliability studies showing poor levels of reliability.
- b. RELIABILITY INFORMATION. *Internal Consistency*: the degree of homogeneity of test items to the attribute being measured. Measured at one point in time.
Observer: i) *intra-observer* - measures variation which occurs within an observer as a result of multiple exposures to the same stimulus. ii) *inter-observer* - measures variation between two or more observers. *Test-Retest*: measures variation in the test over a period of time. Complete the table and reliability information by filling in the TYPE OF RELIABILITY which was tested (internal consistency, observer, test-retest); the STATISTIC that was used (e.g., Cronbach's coefficient alpha, kappa coefficient, Pearson correlation, intra-class correlation); the VALUE of the statistic that was found in the study; and the RATING of the reliability. Guidelines for levels of the reliability coefficient indicate that it will be rated excellent if the coefficient is greater than .80, adequate if it is from .60 to .79, and poor if the coefficient is less than .60.

6. VALIDITY

- a. RIGOUR OF STANDARDIZATION STUDIES FOR VALIDITY. Excellent: More than 2 well-designed validity studies supporting the measure's validity; Adequate: 1 to 2 well designed validity studies supporting the measure's validity; Poor: No validity studies completed, studies were poorly completed or did not support the measure's validity.
- b. CONTENT VALIDITY. Check one of the ratings. *Content Validity*: the instrument is comprehensive and fully represents the domain of the characteristics it claims to measure. (Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.) Excellent: judgmental or statistical method (e.g. factor analysis) was used and the measure is comprehensive and includes items suited to the measurement purpose; Adequate: has content validity but no specific method was used; Poor: instrument is not comprehensive. METHOD. Note whether a judgmental (e.g., consensus methods) or statistical method (e.g., factor analysis) of establishing content validity was used.
- c. CONSTRUCT VALIDITY. *Construct Validity*: the measurements of the attribute conform to prior theoretical formulations or relationships among characteristics or individuals. (Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.) Excellent: More than 2 well-designed studies have shown that the instrument conforms to prior theoretical relationships among characteristics or individuals; Adequate: 1 to 2 studies

demonstrate confirmation of theoretical formulations; Poor: No construct validation completed.

Indicate the STRENGTH OF ASSOCIATION of the findings for construct validity by listing the value of the correlation coefficients found.

d. **CRITERION VALIDITY.** Check one of the ratings.

Criterion Validity: the measurements obtained by the instrument agree with another more accurate measure of the same characteristic, that is, a criterion or gold standard measure. (Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.)

Indicate whether the type of criterion validity which was investigated is CONCURRENT, PREDICTIVE, or both. Excellent: More than 2 well-designed studies have shown adequate agreement with a criterion or gold standard; Adequate: 1 to 2 studies demonstrate adequate agreement with a criterion or gold standard measure; Poor: No criterion validation completed. Indicate the STRENGTH OF ASSOCIATION of the evidence for criterion validity by listing the values of the correlation coefficients which were found in the criterion validity studies. Using the information from the assessment that has been completed on this measure, check the appropriate rating to give an overall assessment of the quality of the measure.

e. **RESPONSIVENESS.** Check one of the ratings (applicable only to evaluative measures).

Responsiveness: the ability of the measure to detect minimal clinically important change over time. (Guyatt, G., Walter, S.D. & Norman, G.R. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, 40, 171-178.) Excellent: More than 2 well-designed studies showing strong hypothesized relationships between changes on the measure and other measures of change on the same attribute; Adequate: 1 - 2 studies of responsiveness; Poor: No studies of responsiveness; N/A: Check if the measure is not designed to evaluate change over time.

7. OVERALL UTILITY

Excellent: Adequate to excellent clinical utility, easily available, excellent reliability and validity. Adequate: Adequate to excellent clinical utility, easily available, adequate to excellent reliability and adequate to excellent validity. Poor: Poor clinical utility, not easily available, poor reliability and validity.

8. MATERIALS USED

Please indicate and list the sources of information which were used for this review. By listing sources of information and attaching appropriate journal articles or correspondence with authors, it will be easier to find further information about this measure if it is required.