Evidence Summary for Pediatric Rehabilitation Professionals

Outcome Measures: A Primer

A. Before selecting a test, it is important to know:

1. The purpose of testing: (i.e. what will the results be used for?)

    a. To discriminate i.e. is child able to do age appropriate activities
    b. To predict i.e. how child will do in future
    c. To evaluate i.e. measure change over time
    d. To plan i.e. assess present skills and make plan on how to proceed and progress

2. Who the child is:

    a. Age
    b. Suspected or known diagnosis
    c. Presenting challenges

3. What content areas need to be assessed:

    a. Impairment, *for e.g.:*

        i. Vision
        ii. Strength
        iii. Sensation
        iv. Tone

    b. Activity, *for e.g.:*

        i. Fine motor skills
        ii. Gross motor skills
        iii. Visual perceptual skills

    c. Participation
    d. Environmental Factors
    e. Quality of Life

4. What are the constraints for the child, family, and/or assessor:

    a. Time
    b. Training
    c. Space
    d. Equipment
    e. Money

When selecting a test, it is important to understand:

1. Types of Tests Available:

   a. Norm-referenced tests = Compare given individual with so called "norm" or average of a group of similar individuals i.e. the purpose is to discriminate
   E.g. Developmental tests such as the Bayley Scales of Infant Development or screening tools such as the Harris Infant Neuromotor Test

   b. Criterion-referenced tests = Compare performance within same child in terms of absolute level of mastery i.e. the purpose is to evaluate
   E.g. Gross Motor Function Measure

   Please note: A few tests are both norm and criterion-referenced.
   E.g. Peabody Developmental Motor Scales

2. Test Scores and Characteristics:

   a. Raw score = Total number of items passed on a test
   b. Standard score = (Child's score – Test mean score)/Test standard deviation
   c. Percentile = Percentage of children of the same age expected to score lower than the child being assessed.
   d. Age equivalent score = Score compared with mean age when 50% of children would have achieved those skills. (*N.B.* Age equivalents should be used with caution as they can easily be misinterpreted; they are typically considered a less desirable way of communicating test results than percentile ranks.)
   e. Developmental quotient = Ratio between the child's actual score based on age equivalent score divided by the child's chronological age.

3. Test Administration Terminology:

   a. Chronological Age = Age from birth
   b. Corrected Age = Age from birth corrected for prematurity (Typically, test developers recommend correcting age for infants born before 37 weeks gestation but assessors should check individual test guidelines. Tests also vary as to what point in time they stop using a corrected age.)
   c. Basal score = Minimum number of passed items needed to complete/stop testing
   For e.g. The Bayley Scales of Infant Development require that the first 3 items in a subtest be passed; if this is not achieved, items must be attempted at a less advanced level.
   d. Ceiling score = Number of failed items needed to complete/stop testing
   For e.g. The Bayley Scales of Infant Development require that 5 consecutive items in a subscale be "failed"; if a child has not failed five consecutive items in a subscale, items must be attempted at a more advanced level. If a child has "failed" five consecutive items, further items are not to be attempted.

4. Measurement Properties: (From Finch et al., 2002, Chapter 4)

a. Scaling Issues

i. Levels of Measurement (Important to understand as these dictate what mathematical equations can be performed with the scale. *For e.g. To add responses a scale should be interval or ratio, and to multiply, divide, or create percentages a scale should be ratio. There are, however, some criteria that can be used to determine whether an ordinal scale can be treated as an interval scale.*)

1. Nominal Scales = Categorical scale with no hierarchy.
   For e.g. Gender or ethnic background
2. Ordinal Scales = Ordered scale with unequal spacing between levels.
   For e.g. Education stated as high school, undergraduate degree, graduate degree or first, second, third place in a running race where participants' times were not equally separated
3. Interval Scales = Ordered scale with equal spacing between levels but no meaningful zero.
   For e.g. Temperature in degrees Celsius where 0 does not signify no temperature even though the difference between 10 and 20 degrees Celsius is the same as the difference between 25 and 35 degrees Celsius
4. Ratio Scales = Ordered scale with equal spacing between levels and a meaningful zero.
   For e.g. Temperature in degrees Kelvin where 0 is absolute 0

ii. Floor and Ceiling Effects: A measure must be able to show both improvement and deterioration. If scores are grouped around the lower end of a scale, the measure won't be able to detect deterioration (i.e. the measure has a floor effect) where as if the scores are grouped around the upper end of the scale, the measure won't be able to detect improvement (i.e. the measure has a ceiling effect).

b. Reliability = To be reliable, a measure must: 1) Provide consistent values with small errors of measurement (Absolute Reliability); 2) Be capable of differentiating between clients with whom the measure is being used (Relative Reliability).

i. Internal Consistency = The extent to which items measure various aspects of the same characteristic and nothing else.

1. Split-half reliability = Method of calculating internal consistency by splitting test items into two comparable halves and correlating the results for each half.

2. Cronbach's Alpha = Method of calculating internal consistency by averaging the coefficients of all possible split-half reliabilities.

ii. Standard Error of Measurement = Index of the degree to which obtained scores differ from true scores.

iii. Confidence Intervals = Range of scores within which you can be highly confident that a true score actually lies.
For e.g. If calculating a 95% confidence interval, one can be 95% confident that the true score is within this interval.

iv. Rater Reliability

1. Intra-Rater Reliability = Examines how similar a measure's results are when one assessor completes the same assessment twice.
For e.g. Comparing scoring of initial assessment with score given when the assessor re-scores a videotaped version of the initial assessment at a different time.
2. Inter-Rater Reliability = Examines how similar a measure's results are when two or more assessors complete the same assessment at the same point in time.

v. Test-Retest Reliability = Examines how similar a measure's results are when one assessor completes the assessment at two different yet not too distant time points.

vi. Intra-Class Correlation Coefficient = Index used to calculate inter-rater and test-retest reliability.

c. Validity = A measure is valid to the extent that it measures what it is intended to measure. Validity implies that a measurement is relatively free from error i.e. a valid test is also reliable.

i. Face Validity = Extent to which a measure appears to be measuring what it is intended to measure.

ii. Content Validity = Extent to which a measure is composed of a comprehensive sample of items that completely assess the domain of interest.

iii. Criterion-Related Validity = Extent to which a measure's results compare to a gold standard measure's results.

1. Concurrent Validity = Extent to which a measure's results compare to a gold standard measure's results when completed at approximately the same point in time.
2. Predictive Validity = Extent to which a measure is able to predict a future criterion event.

iv. Construct Validity = Extent to which a measure provides results that are consistent with theories regarding the attribute of interest.

1. Convergent Validity = Extent to which a measure's results agree with the results of another measure that is believed to be assessing the same attribute. (If this measure is the gold standard, this represents criterion validity; if not, i.e. there is no gold standard, it represents construct validity.)
2. Divergent Validity = Extent to which a measure's results agree with the results of another measure that is believed to be assessing different attributes.
3. Known-Group Validity = Extent to which a measure differentiates between two or more distinct groups.

v. Characteristics of Diagnostic and Screening Tools:

1. Sensitivity = Test's ability to obtain a positive test when the target condition is really present, or a true positive (Recommended = 80%).
2. Specificity = Test's ability to obtain a negative test when the condition is really absent, or a true negative (Recommended = 90%).
3. Positive Predictive Value = Estimate that a person who tests "positive" actually has the condition in question (Recommended = 70%).
4. Negative Predictive Value = Estimate that a person who tests "negative" actually doesn't have the condition.

d. Describing Reliability and Validity: Correlation coefficients are used to describe many elements of a measure's reliability and validity.

i. Correlation = The relationship between two variables
Of note: Correlation does not equal causation.

ii. Correlation Coefficient (*r*) = Number that indicates the strength and direction of the relationship between two variables.
Of note: Negative correlations do not indicate low or poor correlations but rather an inverse correlation.
For e.g. A test such as the Bayley Scales of Infant Development, in which a higher score indicates more advanced development, may be inversely correlated with the Movement Assessment Battery for Children, in which a higher score indicates increased impairment or less advanced development.

iii. Defining Correlation Coefficients: Although there is no widely accepted criteria for describing correlation coefficients, Portney & Watkins suggest that correlations ranging from 0.00 to 0.25 indicate little or no relationship, those from 0.25 to 0.50 indicate a fair degree of relationship, those from 0.50 to 0.75 indicate a moderate to good degree of relationship, and those from 0.75 to 1.00 indicate a good to excellent relationship.

References

1. Bottomley J. 2000. Quick Reference Dictionary for Physical Therapy. Slack Incorporated: Thorofare, NJ.
2. Law M, Baum C, Dunn W. 2001. Measuring Occupational Performance. Slack Incorporated: Thorofare, NJ.
3. Finch E, Brooks D, Stratford PW, et al. 2002. Physical Rehabilitation Outcome Measures: A Guide to Enhanced Clinical Decision Making, 2nd Ed. Canadian Physiotherapy Association; BC Decker Inc: Hamilton, ON.
4. Portney LG, Watkins MP. 2000. Foundations of Clinical Research: Applications to Practice, 2nd Ed. Prentice Hall: Upper Saddle River, NJ.
5. Glascoe FP, Byrne KE, Ashford LG, et al. Accuracy of the Denver-II in developmental screening. *Pediatrics.* 1992;89:1221-1225.

This evidence summary is one part of a series on pediatric rehabilitation outcomes measures.

Other summaries in this series include:

- The Alberta Infant Motor Scale (AIMS)
- The Bayley Scales of Infant Development, 3rd Ed. (BSID-III)
- The Bruininks-Oserestky Test of Motor Performance, 2nd Ed. (BOT-2)
- The Developmental Test of Visual Perception. 2nd Ed. (DTVP-2)
- The Gross Motor Function Measure (GMFM)
- The Movement Assessment Battery for Children, 2nd Ed. (MABC-2)
- The Peabody Developmental Motor Scales, 2nd Ed. (PDMS-2)
- The Sensory Profile (SP)