Assessment Basics: Understanding and Evaluating Assessments for the NRS

State Directors of Adult Education

National Conference

Presenters:
Larry Condelli and Dwayne Norris
American Institutes for Research

February 3, 2010
Agenda

- Introductions, goals for the seminar, pretest
- Presentation: Assessment issues and the NRS
- General discussion
- Break
- Small group activity and large group discussion of activity
Goals of the Seminar

By the end of this training participants will be able to:

- Describe different types of assessments;
- Discuss the characteristics and principles of good assessment;
- Understand the assessment requirements for measuring educational gain for the NRS; and
- Apply these principles to evaluate and make informed judgments about assessments to use in their states.
Presentation Topics

- What is assessment?: Model of the assessment process
- Types of assessments
- Basic assessment terms: validity, reliability, standardization
- Assessment principles for measuring educational gain
- Evaluating assessments: NRS requirements
Student Assessment Process

Adapted from D.M. Kenyon and C. Van Duzer, *Valid, Reliable, and Appropriate Assessments for Adult English Language Learners*, November 2003
http://www.cal.org/caela/accountability
Types of Assessments and Their Use

- **Selection/admission**—skills needed to enter a program – e.g., SAT, MCAT, LSAT, TOEFL
- **Certification/credential** – skills to guarantee student has certain skills required by a curriculum of study – e.g., GED tests, bar exam
- **Diagnosis** – to identify areas of weakness or in need of remediation – e.g., Woodcock-Johnson, Nelson-Denny
- **Achievement** – to measure whether student is learning skills – e.g., TABE
Selecting an Assessment

“States have the discretion to establish the standardized assessment method used within the State ... and must develop a written statewide assessment policy describing assessments and procedures for approval from DAEL. All assessments must conform to standard psychometric criteria for reliability and validity as defined by DAEL.”

(NRS Implementation Guidelines, p. 8)
Validity

“...... refers to the degree to which evidence and theory support the interpretations of test scores... Validity is, therefore, the most fundamental consideration in developing and evaluating tests ...... It is the interpretations of test scores... that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated”

(Standards for Educational and Psychological Testing, p.9)
In other words...

- It is the **use** of a test that is subject to evaluation, not the test itself.
- A test may have more than one proposed use; however, **each** use must be demonstrated to be valid.
- Just because a test is valid for one use, **does not** mean that it is valid for another.
Reliability

- Refers to the consistency of measurement across varying conditions
- Same results achieved regardless of who administers, when where and what form (more on this later)
- Requires *standardization of administration*
Importance of Standardization (1)

- “Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer. ... Specifications regarding instructions to test takers, time limits, the form of item presentations or response, and test materials or equipment should be strictly observed.”

(Standards for Educational and Psychological Testing, p.63)
Importance of Standardization (2)

- Data on validity and reliability of an assessment is created by standardized administration:
  - Mode of administration, conditions
  - Time
  - Scoring

- If procedures are not followed during administration, validity and reliability data are *no longer valid*. 
Questions??

- Types of assessments
- Validity
- Reliability
- Standardization
Meaning for the NRS: Educational Gain

- The primary purpose of assessments used for the NRS is to assess the educational gain of adult learners with respect to the NRS Educational Functioning Levels.

- The ability of an assessment to serve this purpose that should be the focus in determining whether to use it in your state for the NRS.
Assessing Educational Gain

True EFL (Time 1) → True EFL (Time 2)

Observed EFL (Time 1) → Observed EFL (Time 2)
Assessing Educational Gains (1)

- Implications for assessment
  - Content of the assessment must be linked to the EFL descriptors
  - Same assessment must be used for both pre- and post-testing, but with predominantly different sets of items across administrations
  - There must be sufficient time between pre- and post-test to show effect of instruction
Assessing Educational Gains (2)

- Implications for assessment (cont’d)

  - For traditional “fixed-form” assessments:
    - Multiple forms of the assessment must be available, and
    - The form administered during pre-testing to a given examinee must be different than the form administered during post-testing
Assessing Educational Gains (3)

- Implications for assessment (cont’d)
  - For assessments without fixed forms (e.g., computer adaptive tests), other controls must be established to minimize the overlap in items across administrations:
    - Development of very large item banks, with many items at each level being assessed
    - Development and use of separate (non-overlapping) item banks for pre- and post-testing
Evaluating Assessments for the NRS

- Required by Federal regulation
- Annual review process
- Review panels of independent experts in psychometrics and test development
- Evaluated according to criteria described in *Standards for Educational and Psychological Testing*
Information for Evaluating Suitability of Assessments for NRS

- Intended purpose of the assessment
- Development/maintenance procedures
- Match between test content and NRS descriptors (“content validity”)
- Match between test scores and NRS levels
- Reliability/classification consistency
- Construct validity
Intended Purpose of the Assessment

- What the assessment was intended to measure and for what use
- The greater the similarity between the intended use of a given assessment and the purpose of the NRS, the greater the likelihood of suitability, all else being equal
Evaluation of Intended Purpose of the Assessment

- Includes:
  - Test content (skill domains; levels)
  - Inferences/decisions to be made based on examinee performance
  - Targeted population(s)
Adequacy of Procedures used to Develop/Maintain the Assessment

Development

- Steps taken to assure item quality
  - Trained/qualified item writers, editors, and reviewers
  - Thorough item review process
  - Use of pilot/field test to “weed out” bad items
- Appropriateness of pilot/field test sample(s)
  - Should match population to which the assessment will be administered operationally
Adequacy of Procedures used to Develop/Maintain the Assessment (1)

- Maintenance
  - Rate of form/item replenishment
    - Test forms/items that have been in use for a long time have a greater chance of being exposed
  - Steps taken to ensure form comparability
    - Comparisons of pre-test and post-test results assume that the two sets of scores have the same meaning
Adequacy of Procedures used to Develop/Maintain the Assessment (2)

- Maintenance (cont’d)
  - Steps taken to maintain test security
    - Scores associated with items/forms that have been exposed tend to be inflated
    - Procedures need to be in place to limit and control the physical and electronic access to test material
      - Locked storage areas
      - Password protection
      - Trained test administrators
Match between Test Content and NRS Descriptors ("Content Validity")

- The extent to which items/tasks on an assessment cover the domain of interest (and do not cover unrelated skills)
- Usually determined by expert judgement
- Refers to both depth and breadth of content coverage
  - Just because an assessment does a good job of covering some EFLs does not mean that it does an adequate job of covering all EFLs
# Hypothetical Example of Matching Items with NRS Educational Functioning Levels

<table>
<thead>
<tr>
<th>Educational Functioning Level</th>
<th>Number of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>
Match between Test Scores and NRS Educational Functioning Levels

- Adequacy of the procedures used to specify range of test scores corresponding to a given EFL

- Rules for translating test scores to educational functioning level are needed for:
  - classifying examinees into EFLs; and
  - assessing educational gain (re: NRS criterion)
**Sample Table Matching Test Scores to Educational Functioning Levels**

<table>
<thead>
<tr>
<th>Scores on Hypothetical Assessment</th>
<th>Educational Functioning Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 12</td>
<td>1</td>
</tr>
<tr>
<td>13-16</td>
<td>2</td>
</tr>
<tr>
<td>17-23</td>
<td>3</td>
</tr>
<tr>
<td>24-32</td>
<td>4</td>
</tr>
<tr>
<td>33-40</td>
<td>5</td>
</tr>
<tr>
<td>41-45</td>
<td>6</td>
</tr>
</tbody>
</table>
Reliability/Classification Consistency

- Refers to the **consistency** of measurement across varying conditions
- For the NRS, it is necessary to determine the consistency of measurement across alternate forms (or sets of items)
- Of particular importance is consistency of EFL classifications, rather than raw or scale scores
Reliability/Classification Consistency

- Classification consistency is typically examined using “alternate forms reliability” studies
  - Data should be collected from sample(s) similar to the population of interest
  - Design is similar to pre-test/post-test, but with little or no time separating the 2 administrations
### Hypothetical Results of Alternate Forms Classification Consistency Study (1)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>.90</td>
<td>.08</td>
<td>.02</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>.09</td>
<td>.82</td>
<td>.07</td>
<td>.02</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>.01</td>
<td>.07</td>
<td>.84</td>
<td>.06</td>
<td>.02</td>
<td>--</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>--</td>
<td>.02</td>
<td>.06</td>
<td>.82</td>
<td>.07</td>
<td>.01</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>--</td>
<td>--</td>
<td>.01</td>
<td>.07</td>
<td>.83</td>
<td>.09</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>.03</td>
<td>.09</td>
<td>.88</td>
</tr>
</tbody>
</table>
### Hypothetical Results of Alternate Forms Classification Consistency Study (2)

<table>
<thead>
<tr>
<th>Pre-test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.66</td>
<td>.29</td>
<td>.05</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2</td>
<td>.25</td>
<td>.68</td>
<td>.05</td>
<td>.02</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>.01</td>
<td>.07</td>
<td>.84</td>
<td>.06</td>
<td>.02</td>
<td>--</td>
</tr>
<tr>
<td>4</td>
<td>--</td>
<td>.02</td>
<td>.06</td>
<td>.82</td>
<td>.07</td>
<td>.01</td>
</tr>
<tr>
<td>5</td>
<td>--</td>
<td>--</td>
<td>.01</td>
<td>.07</td>
<td>.83</td>
<td>.09</td>
</tr>
<tr>
<td>6</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>.03</td>
<td>.09</td>
<td>.88</td>
</tr>
</tbody>
</table>
Construct Validity

- The extent to which an assessment measures what it is intended to measure
- Supported by different types of evidence, including:
  - Convergent validity
  - Discriminant validity
  - Criterion-related validity
  - Other
Evidence for Construct Validity

- **Convergent validity**
  - Relatively high (positive) correlations with instruments intended to measure the same or similar constructs

- **Discriminant validity**
  - Relatively low or zero correlations with instruments intended to measure unrelated constructs
Evidence for Construct Validity

- **Criterion-related validity**
  - Correlations with important outcome measures
    - Attainment/retention of employment
    - Acquisition of academic credentials (e.g., GED)

- **Other types of evidence**
  - Freedom from sources of variance not related to the construct of interest (e.g., practice effects or culturally-based knowledge)
Questions?

- NRS Review Criteria?
- Assessment Principles?
- Your state’s assessments?
- Other?
We Appreciate Your Participation!

- Thank you!
- Enjoy the meeting
- Visit NRSWeb.org
- Safe travel home!