Validity and Reliability Issues in the Large-Scale Assessment of English Language Proficiency

The 5th International Conference on ELT in China
Beijing, China
May 21, 2007

Richard J. Patz, Ph.D.
CTB/McGraw-Hill
• The measurement of language proficiency
• Assessment design: How to build a valid English-language assessment that reliably measures English-language proficiency
  • Building blocks
  • Constraints
    • Trade-offs
• Selected validity issues
Dimensions of Language Proficiency

- In measurement proficiency is defined as the unobserved ("latent") variable that explains individual differences in performance on an observable set of measures.

- How is language proficiency measured?
  - Listening
  - Speaking
  - Reading
  - Writing
  - Composite scores for Oral (Listening & Speaking), Comprehension (Listening & Reading), Production (Speaking & Writing) may be of interest
Tests are validated by systematically collecting evidence to support the appropriateness of the intended use of the assessment? Requires clear statements regarding the intended uses; validity is not assumed apriori.

E.g., Will the assessment be used to classify English language proficiency in
- academic settings?
- business/employment settings?
Assessment Building Blocks

- Content standards
- Test Blueprints
- Forms configuration
- Test items
- Measurement and Scaling models
- Equating and linking procedures
- Standard setting procedures
- Score reporting methods
Content Standards

• Organize expectations about what students should know and be able to do

• Example: Washington students in grades 3-5 at an intermediate level of achievement should be able to:
  • Respond to directions, questions, and some idiomatic expressions.
  • Use simple sentences to retell or state main point and details of conversations and stories.
  • Recognize inappropriate use of register.
Test Blueprints

- Specify number and types of items for each of the content standards
- Match is in the eye of beholder
- Example: LAS Links

<table>
<thead>
<tr>
<th>Grade Spans</th>
<th>Content</th>
<th>Item Type</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>Listen for Information</td>
<td>MC</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Listen in the Classroom</td>
<td>MC</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Listen and Comprehend</td>
<td>MC</td>
<td>6</td>
</tr>
<tr>
<td>Speaking</td>
<td>Speak in Words</td>
<td>DCR</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Make Conversation</td>
<td>CR</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Speak in Sentences</td>
<td>CR</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Tell A Story</td>
<td>CR</td>
<td>1</td>
</tr>
<tr>
<td>Reading</td>
<td>Analyze words</td>
<td>MC</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Head Words</td>
<td>MC</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Head for Understanding</td>
<td>MC</td>
<td>18</td>
</tr>
<tr>
<td>Writing</td>
<td>Use Conventions</td>
<td>MC</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Write About</td>
<td>SCR</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Write Why</td>
<td>SCR</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Write in Detail</td>
<td>CR</td>
<td>1</td>
</tr>
</tbody>
</table>

Key—MC: Multiple-choice—CR: Constructed-response—DCR: Dichotomous CR—SCR: Short CR
Forms Configuration

• Within One Administration
  • Single form
  • Parallel forms
  • Non-parallel forms
  • Hybrid

• Across Administrations
  • Constant form(s)
  • Overlapping forms
  • Non-overlapping forms
Degrees of Test Comparability

- **Equated forms**
  - Measure same construct
  - Forms are parallel
  - Each student indifferent to assignment of form
    - Same expected score, measurement error
  - “Strict comparability”

- **Linked forms**
  - Measure the same construct
  - Forms may differ in length, coverage, reliability
  - Statistically related (e.g., regression)

Required comparability depends on use
Accountability Requirements

• Accountability requires fairness
• Fairness demands equivalent measurement
• Equivalent measures may be obtained from
  • Use of same instrument
    • Same test form(s)
  • Use of equated instruments
    • Documented technical quality
Breadth vs. Uniformity

- **High Breadth, High Uniformity:**
  - Fairness of Student Comparisons
  - Accuracy of Group Progress Measure
  - Different form Each Student
  - NAEP, TIMMS
  - No student scores

- **Low Breadth, Low Uniformity:**
  - Same Form All Students
  - NCLB Statewide Assessments
  - Student, school, ..., state scores

---

**CTB McGraw-Hill**
Test Items

- Multiple Choice
  - Contrived, not “authentic” (perhaps)
  - Inexpensive to score
  - More items per unit time
  - Can measure complex thinking skills
- Constructed Response
  - More authentic, natural (not guaranteed)
  - Captures thinking process
  - Expensive to score (if by human)
  - Rater differences affect validity (within and across administrations)
  - Machine-scoreable constructed response looks promising
- Extended Response

Good and bad examples of all types exist
Measurement Models

- Specify relationship between student proficiency and success on items
- Unidimensional item response theory
  - 1-, 2-, 3-parameter logistic models
  - Partial credit models
- Multidimensional approaches
  - Multidimensional IRT
  - Bayesian inference networks
  - Cognitive diagnosis models
- Model fit critical
- Useful parameters need not make useful scores
Equating and Scaling

- Tied to forms configuration
- Complete before administration?
  - Requires equating study in advance
  - Enables immediate scoring
- Uses data from live administration?
  - Delays score reporting
  - Quicker development cycle
- Vertical scaling allows comparisons over age spans
Standard Setting

- Maps test scores to proficiency level
- Requires eliciting, synthesizing judgments
  - Categorizing students: “contrasting groups”
  - Categorizing (ordered) items: “bookmark”
- Good to involve variety of stakeholders
- Multiple methods, replication, support validity
- Descriptions of performance-levels follow
Reporting scores

- Statistically optimal estimates based on model/data
  - E.g., use information in full response pattern
- Simple and transparent rules
  - Use only number correct or total points
Large-Scale English Language Assessment

Large scale assessments bring challenges:
  • Volume of work to administer, score
  • Controlling exposure, timing of administrations

And opportunities:
  • See population trends, characteristics
  • Survey broad content efficiently
  • Collect rich background information
  • Research relationship of proficiency to background variables
Some Design Considerations

Designing for Validity
• Maintain broad definition of content domain
• Control exposure of items and item types
• Scoring algorithms (human, machine) robust

Designing for Reliability
• Optimize level of accuracy in scoring
• Sufficient test length: numbers of items/points

Designing for Efficiency
• Leverage technology: online assessment, speech processing, text analysis and AI scoring
Similar challenges in large scale science assessment in United States. One design:

• Detailed framework of content standards
• Large development effort
  • Want to measure whole domain, not sample
• Multiple layers for multiple purposes
  • Public domain
  • Secure for teacher &/or district use
  • Secure for large-scale assessment testing

Test blueprint

- 2/3 student test common to all
  - Common items or strictly parallel form
  - Reliable, comparable student scores
- 1/3 matrixed content
  - Matrixed (BIB) anchor test
  - Field test, link, new content
  - Background, OTL surveys
Matrixed Anchor Test

- Measures entire domain
- Arranged in balanced incomplete blocks
- Constant across administrations
- Provides accurate measures of progress in domain for groups (schools, districts, province, etc.)
- Supports research on growth in English-language proficiency
- Low exposure
Common Form

- Reliable, comparable scores for students, schools
- Released immediately with scoring rules (e.g., Raw-score-to-performance-level)
- Linked, not equated to domain, last year’s common form
- May purposefully wander through domain over years
  - In conjunction with multi-year professional development program
Configuring other building blocks

- Test items
  - MC+CR
  - Many instances of items from item templates possible
- Measurement models
  - Best fitting
- Scaling and equating procedures
  - Pre-link common form
  - Provides immediate reporting
- Standard setting
  - Bookmark better for BIB
  - Annual review of common-form cutpoints, interpretations
- Scoring and Reporting
  - Simple, interpretable rules for common form
  - Best available technical solution for anchor, trends
Breadth vs. Uniformity

- **High Breadth**
  - Fairness of Student Comparisons
  - Accuracy of Group Progress Measure
  - Same Form All Students: NCLB Statewide Assessments, Student, school, …, state scores

- **Low Breadth**
  - Different form Each Student: NAEP, TIMMS, No student scores

- **High Uniformity**
  - "Hybrid" Design
Limitations of Design

- Student comparisons on common form only
- Common form, matrixed form present different picture of achievement
  - Use both in accountability system
  - One immediate; one refined
- Common form does not support year-to-year comparisons
  - Use matrixed anchor for this
- Small schools can be problematic
- Domain interpretations have technical challenges
- Matrixed anchor could grow stale
- Large anchor, multiple forms, add expense
Building an Assessment: Required Expertise

- English-language development domain knowledge
- Psychometrics
- Item development
- Policy experience
- Communication
- Project Management
Discussion

• Design challenge is significant
  • Goals and constraints need refinements
• Building blocks are flexible, configurable
• Priorities of reliability, validity, efficiency, cost affect design
• Any design balances priorities
• Resources are available to help
Discussion (continued)

• English language proficiency tests need to reflect the complexity of the domain.
• Sampled items and item types need to be representative of the domain.
• Good test development procedures will lead to rich information, valid results and interpretations.
• Technology is a key enabler for large-scale assessment of English language proficiency