easyCBM Test Item Development: Merging Researcher and Practitioner Expertise for Student Improvement

P. Shawn Irvin
Behavioral Research & Teaching
College of Education – UO
Road Map

• Foundations of Item Development
• Item Development Process
  – Item Writing
  – Editing and Review
  – Graphics/Audio
  – Standards Alignment/Quality
  – Piloting and Scaling
• Test Form Creation/Equating
• Ongoing Research
Foundations

- Accountability
- Standards-based Instruction
- Research
  - English Language Arts and *The Big 5* (NICHD, 2000)
    - phonemic awareness, alphabetic principles, fluency, vocabulary, and comprehension
  - Mathematics
    - numeracy, operations, reasoning skillsets, etc.
Foundations cont.

• Developing technically adequate interim-formative assessment measures to:
  – Screen for risk, gauge status, and monitor change (McConnell, McEvoy, & Priest, 2002)
  – Establish valid/parsimonious factor structures (Justice, Invernizzi, Geller, Sullivan, & Welsch, 2005)

• easyCBM
  – Reading (early/emergent) and Math
  – RTI framework to improve student learning outcomes through school-wide improvement
Item Development Process

1. Item Writing (P, R)
2. Editing and Review (P, R)
3. Graphics/Audio (P, R)
4. Standards Alignment/Quality (P, R)
5. Piloting and Scaling (P, S, R)

Key stakeholders: Practitioners (P); Students (S); Researchers (R)
1. Item Writing

**Recruitment** of item writers/reviewers

- Representative sample of practitioner experts
- Experience/expertise (i.e., content, years of experience, position held, education level)
- General/Special educators
- e.g., K-5 CCSS Math: 18 individuals, 16 with Masters, ave of 14 yrs experience (r = 3-32), GenEd/SPED
1. Item Writing cont.

Training of item writers (and reviewers)

• Half-day, webinar/in-person sessions

• High-quality items, according to principles of:
  – Universal Design for Assessment (UDA; precise construct targets, accessible to diverse popns, lack of bias) (Thompson, Johnstone, & Thurlow, 2002)
  – Research-based construction (e.g., Haladyna, 2002; 2004)
  – Logistics (e.g., written >> operational, alignment, style, formatting, templates)
  – Examples/non-examples of quality items
  – Targeted practice
2. Editing and Review

- Multi-stage and iterative
  - Concurrent with item writing
  - Subsequent to item writing, concurrent with graphics/audio

- Employing both in- and out-of-house content and test development experts
3. Graphics and Audio Development

- Professional graphic artists hired to create graphics according to UDA
- In-house audio for most items
  - Students with diverse learning/assessment needs
  - English and Spanish audio created for items/measures (e.g., NCTM/CCSS)
4. Item Alignment/Quality

Alignment/quality addressed two-fold:

• Before and during writing/review
• Formal alignment research studies using the Distributed Item Review (DIR)
  – Content/instructional experts judge test items as student would see them in the operational measure
  – Address issues of bias, sensitivity, accessibility
  – Feedback for further improvement (i.e., items revised or discarded)
Distributed Item Review (DIR; BRT, 2013)
• Distribute test items to expert users across appropriate geography (e.g., national, state)
• Examine dimensions of item quality (e.g., alignment/linkage, bias, sensitivity, accessibility)
• Essential features: diverse item types, pertinent support resources, organized assignment to participants, review contexts (e.g., development, review/improvement).
4. Item Alignment/Quality cont.
4. Item Alignment/Quality cont.

- 4,245 assessment items
- ELA, Math, Science – easyCBM/OR alternate assessment
- 121 SPEDucators
- 110 GenEducators
- 38 states
- Multi-purpose studies (alignment, b-s-a)
- More on the horizon!!! 😊

<table>
<thead>
<tr>
<th>Year</th>
<th>Grade</th>
<th>Subject</th>
<th>Items</th>
<th>SPED</th>
<th>Gen-Ed</th>
<th>Total</th>
<th>Bias</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>K-8</td>
<td>Rdg/Math</td>
<td>61</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

*Note. Abbreviations are as follows: ELA = English Language Arts; RC = Reading Comprehension; EL = Early Literacy; Rdg = Reading; SPED = Special Education; Gen-Ed = General Education; B-S = Bias-Sensitivity; SA = Standards Alignment.*
5. Item Piloting and Scaling

Students of varying ability take multiple test items in carefully designed pilot forms to analyze the quality of item functioning and to calibrate items (from a given measure) to a common scale. This makes it so that item difficulty is directly comparable within (and sometimes across) grades.

### 6th and 8th Grade Piloting Plan

| Form | 5A₁ | 30U₁ | 10VS₁ | 5A₂ | 30U₂ | 10VS₂ | 5A₃ | 30U₃ | 10VS₃ | 5A₄ | 30U₄ | 10VS₄ | 5A₅ | 30U₅ | 10VS₅ | 5A₆ | 30U₆ | 10VS₆ | 5A₇ | 30U₇ | 10VS₇ | 5A₈ | 30U₈ | 10VS₈ | 5A₉ | 30U₉ | 10VS₉ | 5A₁₀ | 30U₁₀ | 10VS₁₀ | 5A₁₁ | 30U₁₁ | 10VS₁₁ | 5A₁₂ | 30U₁₂ | 10VS₁₂ | 5A₁₃ | 30U₁₃ | 10VS₁₃ | 5A₁₄ | 30U₁₄ | 10VS₁₄ | 5A₁₅ | 30U₁₅ | 10VS₁₅ | 5A₁₆ | 30U₁₆ | 10VS₁₆ | 5A₁₇ | 30U₁₇ | 10VS₁₇ | 5A₁₈ | 30U₁₈ | 10VS₁₈ | 5A₁₉ | 30U₁₉ | 10VS₁₉ | 5A₂₀ | 30U₂₀ | 10VS₂₀ | 5A₂₁ | 30U₂₁ | 10VS₂₁ | 5A₂₂ | 30U₂₂ | 10VS₂₂ | 5A₂₃ | 30U₂₃ | 10VS₂₃ | 5A₂₄ | 30U₂₄ | 10VS₂₄ | 5A₂₅ | 30U₂₅ | 10VS₂₅ |
|------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|-------|
|      | 25  | 24   | 23    | 22  | 21   | 20    | 19  | 18   | 17    | 16  | 15   | 14    | 13  | 12   | 11    | 10  | 9    | 8     | 7   | 6    | 5     | 4   | 3    | 2     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     | 1   | 1    | 1     |

**Note.** A – horizontal anchor items; VS – anchor items for vertical scaling; U – unique items to the form.

...and pilot forms always have unique items.

Vertical anchor items link test forms across grades allowing calibration to a common scale.
5. Item Piloting and Scaling cont.

- Items analyzed using *item response theory* (IRT)
- Item-level stats, pre-defined criteria (e.g., Wright and Linacre, 1994)
  - *Mean square outfit* – indicator of item performance given item difficulty and student ability
  - *Discrimination* – indicator of relation b/t item and test success, i.e., Does the item yield unique info? Does the item distinguish b/t students with higher-lower performance?
- Poorly functioning items edited/discarded
Test Form Construction/Equating

• Standard (domain) representation
• Range of difficulty – sensitivity at “lower” end of the performance spectrum
• Alternate forms of appx equivalent difficulty (status and growth, teacher/school DM)
• Nuances to reduce construct-irrelevant variance (e.g., domain clustering, ramping difficulty)
Ongoing Research and Collaboration

- Reliability
- Validity
- Cross-validation and Diagnostic Efficiency
- National and Regional Norms
- Test Use and Associated Teacher Decision-making
Thank you! Questions?

http://www.brtprojects.org
http://easyCBM.com