Aligning Achievement Level Descriptors To Mapped Item Demands To Enhance Valid Interpretations Of Scale Scores And Inform Item Development

M. Christina Schneider, CTB/McGraw-Hill
Kristen L. Huff, The College Board
Karla L. Egan, CTB/McGraw-Hill
Margie Tully, CTB/McGraw-Hill
Steve Ferrara, CTB/McGraw-Hill

June 20, 2010
Achievement-level descriptors (ALDs)

- ALDs describe what students should know and be able to do.
- Because ALDs define the intended interpretations of test scores as a field (e.g., Bejar, Braun, and Tannenbaum, 2007; Perie, 2008, Schneider, Egan, Siskind, Brailsford, and Jones, 2009) we are moving toward a vision of:
  - ALDs guiding test development
  - ALDs guiding standard setting
  - ALDs guiding score interpretation
Egan, Schneider, and Ferrara (in press) propose three stages/types of ALDs to correspond with their three intertwined uses:

- Target ALDs
- Range ALDs
- Reporting ALDs

States do not define who the ALD represents but this has important implications for each ALD use.
Specify a state’s expectations for students at the **threshold** of the achievement level

Define the state’s policy and content-based expectations

Guide the cut score recommendation workshop
Range ALDS

- Expand *target* ALDs.
- Define the knowledge and skills expected of examinees in the Proficient **range**
- Guide item development
Reporting ALDs

- Reconcile the target ALDs with the final cut scores
- Define actual student performance based upon the final approved cut scores
- Guide stakeholders in making appropriate, validated inferences about examinee knowledge and skills based upon the student’s test score.
Within ALDS we expect that:

- Achievement levels reflect the more complex knowledge, skills, and abilities (e.g., the difference between what is expected at proficient versus advanced) as they increase.
- Items measuring skills within these achievement levels should increase in complexity, and therefore, in difficulty.
- Claims (descriptions) about what students should be able to do within an achievement level should be reflected in the tasks students are asked to perform and should align in truth on the test scale from a cognitive and content perspective.
However

- We often do not have a specific coherent cognitive and contextual framework that is embedded into the ALD development process.
- We do not have empirical evidence that cognitive and contextual frameworks are related to item difficulty along the test scale.
Therefore

- We worked to
  - analyze the relationships among various cognitive and contextual coding frameworks and item difficulty when reviewing items in test administration order.
- and
  - infer the relationships between features of items and item difficulty when the item difficulty was known.
- Why?
  - We want to improve procedures for developing ALDs.
Item Coding Frameworks

- Depth of knowledge
- Reading load
- Difficulty drivers
- NAEP mathematical complexity
- Question type
Frameworks

DOK (Webb, 2005)
- Level 1
- Level 2
- Level 3
- Level 4

Difficulty Drivers (O’Callaghan, Morley & Schwartz, 2004)
- Applying Math Knowledge 1
- Applying Math Knowledge 2
- Reasoning
- Managing Complexity
- Creating Representations/Insight

Reading Load (Ferrara, Svetina, Skucha, & Murphy, 2009)
- Low
- Medium
- High

NAEP Mathematical Levels of Complexity (NAGB, 2007)
- Low
- Moderate
- High
Frameworks

Question Type (Ferrara et al., 2007)

- use and apply
- Analyze/categorize/hypothesize
Procedures (Part 1)

- Grade 4 mathematics items rated by one team and Grade 8 mathematics items rated by second team
- Assessment editors (AE’s) first recorded what each item measured
- In sequence for each coding framework AE’s were trained in the framework and then immediately rated items according to the framework. Training comprised
  - Review of coding framework
  - Review of sample items
  - Coding of training sets
- With the exception of differently drivers, ratings were not resolved
Items were mapped to the reporting scale and placed in an ordered item book (similar to standard setting)—one for Grade 4 and one for Grade 8.

We investigated item content within each performance level so characteristics could be identified and then potentially used to target future item writing to a particular achievement level.

Moved from target ALD to working with a range ALD.
Rater Agreement

<table>
<thead>
<tr>
<th>Grade</th>
<th>Framework</th>
<th>% Perfect</th>
<th>% Adjacent</th>
<th>% Perfect &amp; Adjacent</th>
<th>Kappa</th>
<th>WT Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>DOK</td>
<td>75</td>
<td>25</td>
<td>100</td>
<td>0.53</td>
<td>0.63</td>
</tr>
<tr>
<td>4</td>
<td>R_Load</td>
<td>69</td>
<td>31</td>
<td>100</td>
<td>0.43</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>DD</td>
<td>91</td>
<td>8</td>
<td>98</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>4</td>
<td>Complexity</td>
<td>94</td>
<td>6</td>
<td>100</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>4</td>
<td>Q_Type</td>
<td>89</td>
<td>n/a</td>
<td>89</td>
<td>0.49</td>
<td>0.55</td>
</tr>
<tr>
<td>8</td>
<td>DOK</td>
<td>76</td>
<td>24</td>
<td>100</td>
<td>0.53</td>
<td>0.59</td>
</tr>
<tr>
<td>8</td>
<td>R_Load</td>
<td>52</td>
<td>45</td>
<td>97</td>
<td>0.19</td>
<td>0.36</td>
</tr>
<tr>
<td>8</td>
<td>DD</td>
<td>71</td>
<td>24</td>
<td>95</td>
<td>0.56</td>
<td>0.6</td>
</tr>
<tr>
<td>8</td>
<td>Complexity</td>
<td>87</td>
<td>13</td>
<td>100</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>8</td>
<td>Q_Type</td>
<td>81</td>
<td>n/a</td>
<td>81</td>
<td>0.25</td>
<td>0.21</td>
</tr>
</tbody>
</table>

- NAEP Levels of complexity had highest perfect agreement rates

- Reading Load had lowest perfect agreement rates but raters ranked it as one of the easiest to use
Findings – Correlations and significance tests

Grade 4

<table>
<thead>
<tr>
<th>Item Difficulty</th>
<th>DOK</th>
<th>R Load</th>
<th>DD Resolved</th>
<th>Complexity</th>
<th>Q Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Difficulty</td>
<td>1</td>
<td>0.14435</td>
<td>-0.03093</td>
<td>0.1766</td>
<td>0.1548</td>
</tr>
<tr>
<td>DOK</td>
<td>&lt;.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>&lt;.0001</td>
<td>0.1216</td>
</tr>
<tr>
<td>R Load</td>
<td>1</td>
<td>0.50934</td>
<td>0.46478</td>
<td>0.4941</td>
<td>0.1955</td>
</tr>
<tr>
<td>DD Resolved</td>
<td></td>
<td></td>
<td></td>
<td>&lt;.0001</td>
<td>0.2467</td>
</tr>
<tr>
<td>Complexity</td>
<td></td>
<td></td>
<td></td>
<td>&lt;.0001</td>
<td>0.15058</td>
</tr>
<tr>
<td>Q Type</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>0.08642</td>
</tr>
</tbody>
</table>

Findings

• For both Grade 4 & 8, item difficulty was not related to cognitive complexity. Some frameworks were related, however.
Visual Inspection

Plot of DOK_R1 by RP50

Plot of R_Load_R1 by RP50

Plot of DD_Resolved by RP50

Large item difficulty ranges in lowest coding level – therefore no correlations found between coding frameworks and item difficulty
Visual Inspection

Math Grade 4
Plot of Com_R1 by RP50

Conditional Stats

<table>
<thead>
<tr>
<th></th>
<th>56</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>596.0515</td>
<td>808.4763</td>
</tr>
<tr>
<td>Mean</td>
<td>24.03072</td>
<td>15.92447</td>
</tr>
<tr>
<td>Std Dev</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Math Grade 4
Plot of Q_Type_R1 by RP50

Conditional Stats

<table>
<thead>
<tr>
<th></th>
<th>59</th>
<th>4</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>597.8703</td>
<td>592.84</td>
<td>576.15</td>
</tr>
<tr>
<td>Mean</td>
<td>24.08708</td>
<td>16.85081</td>
<td></td>
</tr>
<tr>
<td>Std Dev</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Com_R1

Q_Type_R1
AEs hypothesized that subtle nuances in content area skill were driving item difficulty.

- **retrieving information from a bar graph**
  - 1 data point maps to lowest ALD
  - 2 data points maps to higher ALD
- **Addition**
  - add one-digit numbers
  - add three digit numbers
  - add decimals
  - add whole numbers and fractions
Conclusions

- Unable to model item difficulty based on complexity and contextual variables
- Vertically articulating ALDs according to content is a standard practice, however, there may be more subtle content shifts than we oftentimes account for when writing ALDs.
- There is increased focus on writing items to varying levels of complexity so tests are more rigorous; we need more research in this area.
- There may be an interaction between cognitive complexity and content in regard to item difficulty that is not captured using current approaches.
- Our next goal is to have item writers successfully use our newly developed range ALDs and a subset of coding frameworks to predict the achievement level to which a new set of items will map.