Engineering PLDs and Test Content to Engineer Cut Scores

OR Aligning Test Development with Intended Score Interpretations

Steve Ferrara

A presentation in *Engineered Cut Scores: Aligning Standard Setting Methodology with Contemporary Assessment Design Principles*, a session in the National Conference on Student Assessment June 22, 2016
Overview

- Define and illustrate test content-PLD alignment
- Principled design and validity argumentation call for this alignment
- Engineering PLDs
- Engineering test forms
- Engineering items
- Concept not new; engineering PLDs, test forms, and items is WIP
Premise

• Need a broader conception of alignment to engineer cut scores

• Proposed: Alignment is the degree to which an item’s response demands are consistent with the knowledge and skill requirements described in the corresponding PLD

• Aside: Need a broader type of alignment to provide evidence to support intended score interpretations and uses, as well
What is test content-PLD alignment?

<table>
<thead>
<tr>
<th>Level 4: Approaches Expectations for assessed content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very complex text: General accuracy and understanding</td>
</tr>
<tr>
<td>Moderately complex text: General accuracy and understanding</td>
</tr>
<tr>
<td>Readily accessible text: Mostly accurate analyses, showing understanding</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3: Approaches Expectations for assessed content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very complex text: Minimal accuracy and understanding</td>
</tr>
<tr>
<td>Moderately complex text: General accuracy, basic understanding</td>
</tr>
<tr>
<td>Readily accessible text: Mostly accurate analyses, showing understanding</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2: Partially Meets Expectations for assessed standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very complex text: Inaccurate analysis, limited understanding</td>
</tr>
<tr>
<td>Moderately complex text: Minimal accuracy and understanding</td>
</tr>
<tr>
<td>Readily accessible text: Partial accuracy and understanding</td>
</tr>
</tbody>
</table>

## Test content-PLD alignment and articulation not automatic

### Table 7. Summary of Item Response Demands Within and Across Grade Levels and Contributions to Coherent and Disarticulated Inferences About Achievement Growth

<table>
<thead>
<tr>
<th>Response Demand Category</th>
<th>Within Grade</th>
<th>Across Grades</th>
<th>Coherent Articulation?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cognitive Demands</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading load</td>
<td>Mostly low reading load at all grades</td>
<td>Some moderate reading load at grade 5 (only)</td>
<td>Reasonable</td>
</tr>
<tr>
<td>Depth of knowledge</td>
<td>Mostly recall (level 1) at all three grades</td>
<td>Skill/concept (level 2) higher in grade 3</td>
<td>Disarticulation at grade 3</td>
</tr>
<tr>
<td>Mathematical complexity</td>
<td>Generally low</td>
<td>No increase across grades</td>
<td>Reasonable to expect an increase across grades</td>
</tr>
<tr>
<td>Question type</td>
<td>Majority are use/apply items</td>
<td>Fewer analyze/categorize/hypothesize items at grade 5</td>
<td>Disarticulation at grade 5</td>
</tr>
<tr>
<td><strong>Vocabulary Demands</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguous words</td>
<td>75% of all items contain 1–5 ambiguous words</td>
<td>Slight increase at grade 5</td>
<td>Reasonable</td>
</tr>
<tr>
<td>Difficult mathematics</td>
<td>~67% of all items contain 1–2 mathematics terms</td>
<td>Number of items with no terms is lower at grade 5; only grade 5 items contain 3+ terms</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>terms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Syntactical Complexity Demands</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex verbs</td>
<td>Few items contain complex verbs, at all grades</td>
<td>No increase</td>
<td>Reasonable</td>
</tr>
<tr>
<td>Pronouns</td>
<td>~67% of all items contain 1–2 pronouns</td>
<td>Number of pronouns decreases across grades</td>
<td>Disarticulation</td>
</tr>
<tr>
<td>Prepositions</td>
<td>Almost all items contain 1–2 prepositions, some as many as 5 or more</td>
<td>Percentage with 3 or more prepositions is slightly higher at grade 3</td>
<td>Reasonable</td>
</tr>
</tbody>
</table>

Ferrara, Svetina, Skucha, & Davidson, 2011

Engineered Cut Scores NCSA Session
Linking reading comprehension items to the Common European Frame of Reference (CEFR)

Figure 1: Confidence bands for the assigned CEFR levels

Figure 2: Item difficulty (difficulty/mean rating)

Figueros, N., Kaftandjieva, F., & Takala, S. (2013)
PRINCIPLED APPROACHES
Principled approaches to assessment design, development, and implementation

- Several names, several conceptualizations
- Common elements, varying details (Ferrara, Lai, Reilly, & Nichols, 2016)
  - Evidence Centered Design (ECD)
  - Assessment Engineering (AE)
  - Cognitive Design Systems (CDS)
  - BEAR Assessment System
  - Principled Design for Efficacy (PDE)
Principled approaches: Design (etc.) for alignment

Table 1

<table>
<thead>
<tr>
<th>Framework Elements</th>
<th>Assessment Triangle Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizing Element</td>
<td></td>
</tr>
<tr>
<td>Ongoing accumulation of evidence to support validity arguments</td>
<td>Overall evidentiary reasoning goal</td>
</tr>
<tr>
<td>Foundational Elements</td>
<td></td>
</tr>
<tr>
<td>Clearly defined assessment targets</td>
<td>Cognition</td>
</tr>
<tr>
<td>Statement of intended score interpretations and uses</td>
<td>Cognition</td>
</tr>
<tr>
<td>Model of cognition, learning, or performance</td>
<td>Cognition</td>
</tr>
<tr>
<td>Aligned measurement models and reporting scales</td>
<td>Interpretation</td>
</tr>
<tr>
<td>Manipulation of assessment activities to align with assessment targets and intended score interpretations and uses</td>
<td>Observation</td>
</tr>
</tbody>
</table>


Principled approaches are Assessment Engineering, BEAR Assessment System, Cognitive Design System, Evidence Centered Design, and Principled Design for Efficacy

Engineered Cut Scores NCSA Session
ENGINEERING PLDS
Logic of inferences

Ferrara, Lai, Reilly, & Nichols, 2016, Fig. 1

Engineered Cut Scores NCSA Session
Framework for developing PLDs

<table>
<thead>
<tr>
<th>PLD Type</th>
<th>Use</th>
<th>Purpose</th>
<th>Author</th>
<th>Timing</th>
<th>Intended Audience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy</td>
<td>Test development/conceptualization</td>
<td>Set tone for the rigor of performance standards expected by sponsoring agency</td>
<td>Policymakers, senior staff at sponsoring agency</td>
<td>Created at the genesis of the testing program, perhaps as part of legislation</td>
<td>Sponsoring agency staff, practitioners, item writers, policymakers, the public</td>
</tr>
<tr>
<td>Range</td>
<td>Item writing guidance</td>
<td>Define content range and limits for item writers</td>
<td>Content experts, cognitive scientists, sponsoring agency staff</td>
<td>Created prior to or in conjunction with content standards. Adjusted based upon final cut score approval</td>
<td>Item writers, test developers</td>
</tr>
<tr>
<td>Target</td>
<td>Cut-score recommendation workshop guidance</td>
<td>Define the threshold student for item writers and standard setters</td>
<td>Content experts, sponsoring agency staff, standard setters</td>
<td>Created prior to item development and cut-score recommendation workshop</td>
<td>Standard setting panelists, item writers, test developers</td>
</tr>
<tr>
<td>Reporting</td>
<td>Test score interpretation</td>
<td>Define KSPs that test taker demonstrates</td>
<td>Content experts, sponsoring agency staff</td>
<td>Created after approval of final cut scores by sponsoring agency.</td>
<td>The public (e.g., parents, examinees, teachers)</td>
</tr>
</tbody>
</table>

From Egan, Schneider, & Ferrara, 2012, Table 2
Engineered Cut Scores NCSA Session
Guidance on engineering PLDs

- Develop PLDs to guide test development, articulate standards across grades, link to learning trajectories (Bejar, Braun, & Tannenbaum, 2006, 2007)

- Policy definitions
  - Generic, by policy makers (Loomis & Bourque, 2001; Perie, 2008)
  - Number of levels; labels and meaning (Beck, 2003; Burt & Stapleton, 2010; Cizek & Bunch, 2007; Egan, Schneider, & Ferrara, 2012; Perie, 2008; Zieky, Perie, & Livingston, 2008)

- Range and other PLDs
  - Explicit about content knowledge (Egan et al., 2012; Mills & Jaeger, 1998; Perie, 2008; US Department of Education 2004)
  - Explicit about cognitive processes (Egan et al., 2012; Perie, 2008)
  - Nouns and verbs, defining phrases (Egan et al., 2012)

- Would like to find guidance on working from policy PLDs to content standards to range PLDs
PURSUING PLD-ITEM ALIGNMENT
How can we pursue alignment?

- Code items for response demands
- Determine which items are aligned with the KSA requirements in the corresponding PLD

- Build aligned test forms
  - Select only aligned items
- Item development
  - Develop item specifications and item writer training to improve alignment
  - Re-field test items that are not aligned to PLDs
ENGINEERING TEST FORMS FOR ALIGNMENT: ILLUSTRATION
Selected item response demand codes

- **Question Type**
  - The cognitive task an item poses (e.g., explain, analyze)

- **Depth of Knowledge**
  - Recall, Skill/concept, Strategic thinking

- **Relational Complexity**
  - No. of facts, concepts, and skills to be processed

- **Linguistic Complexity**
  - No. of prepositional phrases, as a proxy

- **Command of Textual Evidence**
  - Single, multiple pieces of text

- **Response Mode**
  - Select response, multiple responses, construct responses
Excerpt from Of Fat, Feathers...; item 3874

Part A

According to the article, why do chickadees eat a lot of fatty foods every day?

A. Chickadees consider fatty foods to be special treats.
B. People who have chickadee bird feeders fill them with fatty foods.
C. Other types of birds fly south and leave mostly fatty foods for chickadees.
D. Fatty foods provide chickadees with just enough fuel to stay warm overnight.

To a chickadee, body fat is a fuel. Since they can store food to eat later, they must eat long, 18-hour night. By feasting on fatty foods body weight by about 8 to 10 percent—which gaining 8 to 10 pounds in one day. The amount depends on what kind of food is available and Chickadees living near well-stocked bird feeders than those living far from humans. Increasing like stoking the fire before going to bed. Throughout the night chickadees burn their recently accumulated fat, and by the next morning, none is left.
Item 3874

**PLD Alignment Target**

Moderate Complexity text: Minimal accuracy and understanding

Do correct responses to this item support this intended interpretation (or claim)?

**Item Response Demands**

<table>
<thead>
<tr>
<th>Question Type</th>
<th>USE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOK level</td>
<td>2. Skill/concept</td>
</tr>
<tr>
<td>Relational Complexity</td>
<td>5</td>
</tr>
<tr>
<td>Number of Prepositions</td>
<td>5</td>
</tr>
<tr>
<td>Command of Textual Evidence</td>
<td>Low</td>
</tr>
<tr>
<td>Response Mode</td>
<td>Low</td>
</tr>
</tbody>
</table>

Engineered Cut Scores NCSA Session
Excerpt from Turn, Turn My Wheel and item 3182

20 And while he plied his magic art—
For it was magical to me—
I stood in silence and apart,
And wondered more and more to see
That shapeless, lifeless mass of clay

Prior knowledge not going to help

Part A
In line 22 of “Turn, Turn, My Wheel,” the speaker stands in silence and apart as the potter creates his work. Why is the speaker silent?

A. The speaker is in awe of the potter’s power and skill.
B. The speaker is frightened of the potter’s magical powers.
C. The speaker dislikes the final product.
D. The speaker wants to ask a question but is nervous.
Item 3182

PLD Alignment Target
Readily Accessible text: Partial accuracy and understanding
Do correct responses to this item support this intended interpretation (or claim)?

Item Response Demands
- Question Type: USE
- DOK level: 2. Skill/concept
- Relational Complexity: 5
- Number of Prepositions: 5
- Command of Textual Evidence: Low
- Response Mode: Low

Item Response Demands
- Question Type: USE and INF
- DOK level: 3. Strat. Thinking
- Relational Complexity: 6 or more
- Number of Prepositions: 7
- Command of Textual Evidence: Low-Moderate
- Response Mode: Low
“Arid” seems pretty clear; “cold,” “long,” and “rotten” seem unlikely. So what makes this relatively difficult?

Prior knowledge of “arid” might help, but “arid bones” is figurative language.
Item 3175

Level 3: Approaches Expectations

Level 2: Partially Meets Expectations

Level 2: Partially Meets Expectations

Very complex text: Inaccurate analysis, limited understanding

Do correct responses to this item support this intended interpretation (or claim)?

Item Response Demands

Question Type: USE and INF
DOK level: 3. Strat. Thinking
Relational Complexity: 6 or more
Number of Prepositions: 7
Command of Textual Evidence: Low-Moderate
Response Mode: Low

PLD Alignment Target

Very complex text: Inaccurate analysis, limited understanding

Do correct responses to this item support this intended interpretation (or claim)?

Item Response Demands

Question Type: USE or INF
DOK level: 2. Skill/concept
Relational Complexity: 5 or more
Number of Prepositions: 4
Command of Textual Evidence: Low
Response Mode: Low
ENGINEERING ITEMS FOR ALIGNMENT
Framework for engineering items

- Content, cognitive, and linguistic response demands frameworks
  - Select item response demand codes most appropriate for the test and PLDs working with

- Empirical support
  - Reliable coding (some of it semi-automated)
  - Important predictors of difficulty and discrimination
  - $R^2 = \{.15, .30\}$: Need more work (e.g., additional or better frameworks, learning science lit reviews, OTL question)

- Still working out PLD-item alignment criteria
  - For now, normative criteria
Engineering items to target PLDs

1. Breakfast that morning was a feast—a bit of the rice boiled to a gruel in a castoff earthenware pot, served up in a bowl carved from a gourd. And Crane-man produced yet another surprise to add to the meal: two chicken leg-bones. No flesh remained on the arid bones, but the two friends cracked them open and worried away every scrap of marrow from inside.

<table>
<thead>
<tr>
<th>Item Response Demands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Type</td>
</tr>
<tr>
<td>DOK level</td>
</tr>
<tr>
<td>Relational Complexity</td>
</tr>
<tr>
<td>Number of Prepositions</td>
</tr>
<tr>
<td>Command of Textual Evidence</td>
</tr>
<tr>
<td>Response Mode</td>
</tr>
</tbody>
</table>

Part A
What is the meaning of arid as it is used in the first paragraph of the passage from A Single Shard?

- A. cold
- B. long
- C. dry
- D. rotten
Engineering items (cont.)

- Theoretically, once you know some things about the relationship between item response demands and item difficulty:
  - Create item templates to guide development of more items with similar difficulties
  - Engineer existing items to hit difficulty targets
- We’re just getting started
We-are-not-Alone (WANA) Alignment Model

Infrastructure <--- Policy <--- PLDs <--- Test Content

**Enactment**

**Infrastructure**: Professional development, curriculum materials, instructional approaches

**Policy**: CCSS, NGSS

**PLDs**: Range PLDs

**Test Content**: Standards targeted, item types, item response demands

A = alignment; Coburn, Hill, & Spillane (2016)

**Enactment and implementation**: McDonnell & Weatherford (2016)
Thanks!

Steve Ferrara
sferrara1951@gmail.com
References


References (cont.)


