Constructing aligned assessments using automated test construction

Morgan S. Polikoff
Andrew C. Porter
Katherine M. Barghaus
Rui Yang

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C080009 and Grant R305C050041-05. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.
Standards-based reform emphasizes the importance of coherence – alignment among tests, standards, and curriculum.

- NCLB requires state achievement tests to be aligned with content standards.
- But research shows that NCLB-era assessments were, in fact, modestly aligned with their corresponding content standards (Polikoff et al., 2011)

Alignment (not mentioned by name) an important part of AERA/NCME/APA standards.
CONTEXT – WHERE WE BEGAN

- Two parallel (FOSS, Holt) three-arm cluster randomized trial of two interventions (cognitive science, content knowledge) and their effects on instruction and student outcomes relative to a control.
  - Teacher PD on cognitive science principles, content knowledge
  - Student-level outcomes: state science test
    - Concerns this test would not be sensitive to our intervention given distal nature, questionable alignment
PURPOSE

- To demonstrate the development and application of a test construction algorithm for creating customized aligned assessments of student achievement and teacher knowledge.
GENERAL APPROACH

- Content analyze the “target” of the test
  - Teacher content PD
  - Science unit (unmodified)
- Assemble a pool of existing, high-quality items and content analyze them
  - Sources such as NAEP, state tests, validated research instruments, PRAXIS
- Use an automated process to select the items from the pool that create the most highly aligned test possibly to the target
**CONTENT ANALYSIS PROCEDURE**

- Surveys of Enacted Curriculum (Porter, 2002)
- Content taxonomy in science: 211 topics, 5 cognitive demand levels (intersection is a cell)

<table>
<thead>
<tr>
<th>Time on Topic</th>
<th>Grades K-12 Science Topics</th>
<th>Expectations for Students in Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2200</td>
<td>Meteorology</td>
<td></td>
</tr>
<tr>
<td>2201</td>
<td>Earth's atmosphere</td>
<td>0 1 2 3</td>
</tr>
<tr>
<td>2202</td>
<td>Air pressure and winds</td>
<td>0 1 2 3</td>
</tr>
<tr>
<td>2203</td>
<td>Evaporation, condensation, and precipitation</td>
<td>0 1 2 3</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td>Elements and The Periodic System</td>
<td></td>
</tr>
<tr>
<td>2300</td>
<td>Early classification system(s)</td>
<td>0 1 2 3</td>
</tr>
<tr>
<td>2301</td>
<td>Modern periodic table</td>
<td>0 1 2 3</td>
</tr>
<tr>
<td>2302</td>
<td></td>
<td>0 1 2 3</td>
</tr>
</tbody>
</table>
ALIGNMENT

Alignment of any two documents can be estimated by comparing the proportions in the content matrix describing document 1 to those in the matrix describing document 2.

\[
\text{Alignment} = 1 - \frac{\Sigma |X - Y|}{2}
\]

where \(X\) denotes cell proportions in one matrix and \(Y\) denotes cell proportions in another matrix (Porter, 2002).

Index ranges from 0 to 1, with 1 representing perfect alignment.
Adjusted alignment for any document pair is the unadjusted alignment divided by the maximum alignment possible for a given target and test length.

Maximum alignment not actually 1 – estimated using simulations

Assumptions:
1. Infinite item pool
2. Four content analysts who could each place an item in up to three cells
3. Complete rater agreement or disagreement
**Test Construction Algorithm**

1. Content analyze the target domain.
2. Decide the test length in terms of number of score points.
3. Content analyze each item in the population of items.
4. Calculate the alignment of each item to the target.
5. Select the most aligned item (if a tie, randomly select from among tied items) and adjust the target to account for content of the item.
6. Recalculate the alignment for each item.
7. Repeat Steps 5 and 6 until enough items have been selected to generate a test of the desired length.
## Test-Target Alignment

**Table 1**

End of Unit Student and Teacher Achievement Test Alignments With Target Domains

<table>
<thead>
<tr>
<th></th>
<th>Holt</th>
<th>FOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cell</td>
<td>ITM</td>
</tr>
<tr>
<td>Student Achievement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alignment</td>
<td>0.53</td>
<td>0.58</td>
</tr>
<tr>
<td>Agree max</td>
<td>0.74</td>
<td>0.85</td>
</tr>
<tr>
<td>Disagree max</td>
<td>0.89</td>
<td>0.95</td>
</tr>
<tr>
<td>Agree adjusted</td>
<td>0.70</td>
<td>0.68</td>
</tr>
<tr>
<td>Disagree adjusted</td>
<td>0.58</td>
<td>0.61</td>
</tr>
<tr>
<td>Teacher Knowledge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alignment</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>Agree max</td>
<td>0.62</td>
<td>0.71</td>
</tr>
<tr>
<td>Disagree max</td>
<td>0.91</td>
<td>0.93</td>
</tr>
<tr>
<td>Agree adjusted</td>
<td>0.52</td>
<td>0.51</td>
</tr>
<tr>
<td>Disagree adjusted</td>
<td>0.35</td>
<td>0.39</td>
</tr>
</tbody>
</table>

*Note.* Alignment is the unadjusted alignment index of the assessment. Agree max is the maximum possible alignment assuming complete agreement among content analysts. Disagree max is the maximum possible alignment assuming complete disagreement among content analysts. Agree adjusted is the raw alignment index adjusted for the maximum alignment given complete agreement. Disagree adjusted is the raw alignment index adjusted for the maximum alignment given complete disagreement. ITM = Introduction to Matter; IRE = Inside the Restless Earth; EH = Earth History; WW = Weather and Water; DOL = Diversity of Life.
FIGURE 1. Topographical maps comparing the content of the target domain and the constructed teacher content knowledge test for the cells unit.
ANOTHER VARIANT – SUBTESTS

- If it is not possible, or not desired, to administer a supplementary test, an algorithm can also be used to create subtests of an existing test (e.g., state test), assuming item-level performance data exist.
  - Procedure is similar – content analyze items, select items to maximize alignment.
  - Can allow number of items to vary – start at 1 item, calculate alignment, then 2, etc. Graphing maximum alignment for each number of item will allow you to choose maximally-aligned subtest.
Reliability

- Generalizability theory was used to determine the reliability of the content analyses of the custom tests.

- Generalizability coefficient were generally satisfactory:
  - Holt Cells student test – .79
  - FOSS Earth History student test – .95
  - Holt Cells teacher test – .63
  - FOSS Earth History teacher test – .74
LIMITATIONS

- Based on the SEC conception of content
- Depends on the quality of the item pool (though this applies to all tests)
- Exclusive focus on alignment – of course things like item bias, reliability are also important
- Doesn’t necessarily build maximally aligned test
APPLICATIONS

- Guide item writing to improve test alignment to a target (e.g., Common Core standards)
- Constructing parallel forms – select enough items for two tests and randomly assign
- Useful in variety of contexts for intervention studies, e.g., evaluating whether magnitude of effect size depends on test content
- May help make distal assessments less distal – improving their sensitivity?