INSTRUCTIONAL SENSITIVITY: THE CASE FOR INSTRUCTIONAL COVERAGE

Min Li\textsuperscript{1}, Maria Araceli Ruiz-Primo\textsuperscript{2}, & Kellie Wills\textsuperscript{1}

\textsuperscript{1}Uninveristy of Washington at Seattle
\textsuperscript{2}University of Colorado Denver

International Instructional Sensitivity Conference
November 14, 2013
The DEISA Team

University of Colorado Denver
- Maria Araceli Ruiz-Primo
- Kendra Occhipinti
- Michael Orgeron
- Mary Blair
- Renee Johnson
- Kristine Mathieseen

University of Washington at Seattle
- Min Li
- Kellie Wills
- Ming-Chih Lan
- Ting Wang
- Derek Zhao

Oregon State University-Cascades
- Michael Giamellaro
Overview

- The DEISA Project
  - Developing and Evaluating Instructionally Sensitive Assessments

- The Study

- Results
  - The gist of them

- Conclusions
DEISA Project in a nutshell

Through a four-tryout iterative process, we are:

- Building an approach to developing *instructionally sensitive assessment*
  - What item features can we systematically manipulate to write items more or less sensitive?

- Collecting empirical evidence to test the *validity* claims of the developed assessments in relation to the instruction students received
  - Did the items behave as what we expected?
Dimensions of instructional sensitivity

Instructional sensitivity of a test (or a test item) refers to the degree of:

- sensitivity to the **curriculum content** that was taught to the examinees/students
- sensitivity to the **quality of the instruction**
- **formative value** of the assessment information to instruction
Items at different proximities

- **Distal**: State/national standards-based large-scale assessment

- **Proximal**: Assessments items tap same concepts but in a different way than what was taught in the class

- **Close**: Assessment items are similar to activities in the units taught

- **Immediate**: Classroom artifacts (notebooks, worksheets) and embedded assessments

Source: Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002
Developing items at different proximities: the “Bundle” concept

Transfer of Learning

Far

Distal

Far proximal

Proximal

Near proximal

Close

Near proximal

Close
Main hypotheses on item sensitivity

- A greater gain of student learning is reflected in the close items and is less when the items become less sensitive.

- The patterns of student performance should reflect what, how much, and how students were taught.
  - If we are interested in comparing teachers with high coverage vs. teachers with low coverage, close items are more indicative of the instructional coverage when compared to items less sensitive.
Research questions

- Are the DEISA items sensitive to the instructional coverage?

- Do the three methods provide comparably similar results about item statistics regarding the instructional sensitivity?
Methods: participants and curricula

- 17 teachers from the Landforms module; 14 teachers from the Mixtures and Solutions module
- Those 5th grade teachers of the two coverage groups differed in their self-reports on whether the content/skills for the test items was taught. But no difference was found in teaching experience, class size, or qualification.

<table>
<thead>
<tr>
<th>Module</th>
<th>Low Coverage Group</th>
<th>High Coverage Group</th>
<th>Statistical Test by Coverage Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n²</td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>LF</td>
<td>9</td>
<td>0.59</td>
<td>0.07</td>
</tr>
<tr>
<td>MS</td>
<td>8</td>
<td>0.41</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Items were developed in bundles (close, near proximal, and far proximal).
Close item example

Close item

C  Sue’s stream table erodes more earth material than Tom’s.

What is probably true?

A. Sue’s stream channel is less deep than Tom’s.
B. Sue’s stream channel is narrower than Tom’s.
C. Sue’s delta is smaller than Tom’s.
D. Sue’s delta is larger than Tom’s.
Near proximal item example

Near Proximal Item

P1 The average speed of water in Cherry Creek was faster in 2008 than in 2010. Because of this, the creek eroded more material in 2008.

What else was probably true in 2008?

A. The creek deposited less material downstream.
B. The creek deposited more material downstream.
C. The creek channel was less steep than in 2010.
D. The creek level was lower than in 2010.
Far Proximal Item

P2 One form of erosion takes place when rocks break off from a cliff. What is the deposition of these rocks?

A. They land at the base of the cliff.
B. They get picked up by a river.
C. They break off other rocks when they hit them.
D. They are not deposited.
Distal Item

27. Which statement explains why the Sun appears to rise and set each day?

A. Earth rotates.
B. The Sun rotates.
C. The Sun revolves around Earth.
D. Earth revolves around the Sun.
Methods: item scores

- The assessments included items varying in proximity, organized into two booklets.
  - We only used the common items for each module, including roughly 3 close items, 3 near proximal items, 3 far proximal items, and 7 distal items.

- The assessment booklets were administered to the students before and after the module was taught.
## Methods: the three statistical methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Interpreting Item Parameters</th>
</tr>
</thead>
</table>
| Coverage difference index  | **P value difference between the two coverage group**  
Instructionally sensitive items should show a larger difference between the high and low coverage groups. |
| Variance component estimation | **Percent of variance attributable to the quality facet**  
Instructionally sensitive items should have a higher % of variance that is accounted for by the coverage facet. |
| Mixture IRT                | **Difficulty parameter for each coverage group**  
Instructionally sensitive items should have the estimated parameters reflective of the group membership. |
Methods: the three statistical methods (ii)

<table>
<thead>
<tr>
<th>Method</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage difference Index</td>
<td>$p \ (hc) - p \ (lc)$</td>
</tr>
<tr>
<td>VC s:t:c nested model</td>
<td>$\sigma_{s:t:c}^2 = \sigma_{q_2}^2 + \sigma_{t:c}^2 + \sigma_{s:t:c, e}^2$</td>
</tr>
<tr>
<td>IRT 2 parameter factor mixture IRT model with known class membership</td>
<td>$pU=1, g, \theta_{\downarrow g} = e^{{a_{\downarrow i g} (\theta_{\downarrow g} - b_{\downarrow g})}} / 1 + e^{{a_{\downarrow i g} (\theta_{\downarrow g} - b_{\downarrow g})}}$</td>
</tr>
</tbody>
</table>
Findings (i):

item statistics

- Landforms module
  - $M(d \text{ index}) = 0.40$
  - $M(p \text{ posttest}) = 0.67$

- Mixtures and Solutions module
  - $M(d \text{ index}) = 0.35$
  - $M(p \text{ posttest}) = 0.61$
Findings (i):
item ppdi estimates by module

- Close: LF = 0.3, MS = 0.3
- Near proximal: LF = 0.25, MS = 0.2
- Far proximal: LF = 0.15, MS = 0.2
- Distal: LF = 0.1, MS = 0.2
Findings (ii): item sensitivity regarding coverage

- So what were the good, bad, and ugly items?

<table>
<thead>
<tr>
<th>Coverage index</th>
<th>LF (n=16)</th>
<th>MS (n=15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G theory</td>
<td>IRT</td>
</tr>
<tr>
<td>Sensitive</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Insensitive</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Opposite</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Sensitive</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Insensitive</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Opposite</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Findings (iii): 
item sensitivity regarding coverage

- The items identified by the three methods were different.
Conclusions: what have we learned?

- Items judged to be sensitive for other aspects of instruction may not be sensitive to the instructional coverage.

- It may be challenging to develop items sensitive to the instructional coverage for a particulate target module.
  - Lack of variation between teachers
  - Influence of module activities
  - Confounding effect of instructional quality
Conclusions: what do we need to learn?

- Larger sample size is needed to run the analysis; more teachers with enough variation, more students, and more items
- Ways to capture accurate data for the instructional coverage
- Incorporating other variables, such as instructional quality, student pretest scores
The work reported herein was supported by National Science Foundation (DRL-0816123).

Thanks go to all the graduate and undergraduate RAs in this project.