FUTURE RESEARCH BEYOND INSTRUCTIONALLY SENSITIVE ITEMS

Stephen C. Court
Educational Research and Evaluation, LLC

A Presentation at the First International Conference on Instructional Sensitivity
Achievement and Assessment Institute
University of Kansas, Lawrence, Kansas,
November 13-15, 2013
Item-level

- **Effective Less**
  - A
  - C
- **Effective More**
  - B
  - D

**IS** = \( \frac{A + D}{N} \)

20 / 25 = .80

- **Effective Less**
  - 5
  - 2
- **Effective More**
  - 15

**IS** = \( \frac{A + D}{N} \)

- **Effective Less**
  - 20%
  - 8%
- **Effective More**
  - 12%
  - 60%

20% + 60% = .80
Beyond Items

Instructional sensitivity may reside not only in items and tasks.

While crunching the numbers for the results I shared yesterday, the data patterns raised several ideas regarding areas for future research.

Here are some of them.
Typically, developers select a wide range of item difficulties for the test overall and even within subscales.

This was desirable in NRT context because doing so helped to separate students.

But for purposes of school accountability and teacher evaluation, we need to revisit this notion.

If we use standards-based (CRT) selection criteria – all items of similar difficulty – do the indicator and total-test results become more sensitive to instruction?
Typically, items that are too easy (p-value > .85 or .90) are deleted from final test forms.

Yet, these are the items that tend to discriminate best between effective and less effective instruction... At the classroom level and above, the most poorly taught groups of students tend to be the ones who get the most items wrong. Conversely, the best taught groups tend to earn the highest scores.

RQ: Is there an “ideal” p-value for items used on evaluative tests?
Standard-setting and Optimum Cut Scores

In Kansas, a popular report is the instructional planning graph.
For each indicator, use ROC analysis to set optimum cut scores that maximize instructional sensitivity.

For example, given an indicator measured by four items, what cut score (number correct) is most sensitive to instruction?
Diagonal segments are produced by ties.

### Area Under the Curve

<table>
<thead>
<tr>
<th>Test Result Variable(s)</th>
<th>Area</th>
<th>Std. Error</th>
<th>Asymptotic Sig</th>
<th>Asymptotic 95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:0</td>
<td>.500</td>
<td>.033</td>
<td>1.000</td>
<td>.436 - .564</td>
</tr>
<tr>
<td>13:1</td>
<td>.519</td>
<td>.034</td>
<td>.570</td>
<td>.452 - .585</td>
</tr>
<tr>
<td>13:2</td>
<td>.686</td>
<td>.036</td>
<td>.009</td>
<td>.514 - .657</td>
</tr>
<tr>
<td>13:3</td>
<td>.668</td>
<td>.033</td>
<td>0.000</td>
<td>.604 - .733</td>
</tr>
<tr>
<td>13:4</td>
<td>.500</td>
<td>.033</td>
<td>1.000</td>
<td>.436 - .564</td>
</tr>
</tbody>
</table>
Differential Sensitivity

Examine how item, indicator, and total test sensitivity varies across...

✓ N-size
✓ AYP or CCR subgroups
✓ other disaggregations of interest

Also, check concordance between confirmatory (intended) and exploratory (actual) factor patterns:
✓ Same number of factors?
✓ Items load onto the intended factors?

Greater concordance \(\rightarrow\) greater sensitivity.
Who remembers the formula for standard error of measurement (SEM) or for the standard error of the mean?

Do you know where to find each formula?

Do you understand what they are, when each is applicable, and how to use them?

For an elementary or middle school student, such a problem may exist for the difference between the area and the circumference of a circle.
Why do so many geometry items call for such distinctions with, at most, the student being given “hints” such as:

\[ A = \pi r^2 \quad \text{and} \quad C = 2 \pi r \]

Might items and tasks be more instructionally sensitive if students could have access to resources such as textbooks, the Internet, and other resources...just as they would in real life?
In turn, if constructs were redefined to involve inquiry...and if tests were administered in an “enriched” environment ...

Would teachers widen the curriculum and teach procedural and inquiry skills rather than just drilling-and-killing declarative knowledge for the sake of the test?

In addition, if we exchange the sterile environment for an enriched environment, might such an **authentic** testing approach diminish incentives to “teach the test?”

- Further, would such an approach be more engaging for students in terms of (a) learning and (b) being tested?
- Would achievement gaps narrow with respect to test results and actual learning?
- Would such an approach spawn an upward spiral of increased success, self-efficacy, and motivation?
We need more research into...

Items, tasks, reading passages, and other stimuli...

But also into all other aspects of testing, instruction, and educational excellence.

Just as important: We need to communicate our research findings not only among ourselves (the measurement, assessment, and research community) but also to policy-makers, practitioners, and stakeholders at the local, state, and national levels.