Comparing Two Study Designs to Evaluate Item Sensitivity

Kellie Wills and Min Li
University of Washington

Maria Araceli Ruiz-Primo
University of Colorado Denver
A challenge to instructional sensitivity claims

• The DEISA basic design is pretest-posttest.
  – 5th grade classrooms are tested before instruction in a science module, and tested again after instruction, using the same test booklet.

• How do we distinguish instructional sensitivity from alternative explanations for improved performance?
Two control strategies

Both strategies address *maturation threats* to validity (Shadish, Cook, & Campbell, 2002).

- Naturally occurring changes over time that may be confused with instructional sensitivity

1. Compare posttest performance of instructed 5<sup>th</sup> graders with their own repeated pretest scores.

2. Compare performance of 5<sup>th</sup> graders with performance of 4<sup>th</sup> graders.
Control strategy 1: Repeated pretesting of 5th graders

- Improvement without instruction may result from repeated exposure to an item (Rogers & Yang, 1996).
- Post-instruction improvement should exceed baseline (uninstructed) improvement.
Control strategy 2: 4\textsuperscript{th} graders

A. Differences between (uninstructed) 4\textsuperscript{th} graders and instructed 5\textsuperscript{th} graders (posttest) are evidence of instructional sensitivity (Wiliam, 2007).

B. Performance should be similar for 4\textsuperscript{th} graders and the pretest of 5\textsuperscript{th} graders.
   – Differences are evidence that other factors, such as reading ability, may account for performance improvements.
Participants

• Denver metro area urban school district
  – 29% non-white, 19% free/reduced lunch, 8% English language learners

• 4\textsuperscript{th} and 5\textsuperscript{th} graders
  – 125 5\textsuperscript{th} graders (six classrooms) studying FOSS Landforms participated in pre/post-testing.
  – Of these, 64 5\textsuperscript{th} graders (three classrooms) participated in two pretests prior to instruction on the module.
  – 99 4\textsuperscript{th} graders were tested once.
DEISA Instrument

• Three levels of instructional sensitivity:
  – **Close (C):** item scenarios same as classroom activities.
  – **Near-proximal (P1):** item scenarios are closely related to classroom activities.
  – **Far-proximal (P2):** items focus on module topics and concepts, but are not as close to classroom activities.

Cronbach’s alpha (posttest): **0.81**

Test-retest reliability (pretest 1 - pretest 2) **0.74**, (pretest 2 – posttest) **0.70**
# Hypothesis tests

| Control strategy | No control. 5\(^{th}\) graders: Was post-instruction proportion correct significantly larger than pre-instruction proportion correct? | \(H_0: p_{post} = p_{pre2}\)  
\(H_1: p_{post} > p_{pre2}\) | \(H_1\) represents evidence for instructional sensitivity |
|------------------|-------------------------------------------------------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| Control strategy 1. 5\(^{th}\) graders: Was gain after instruction significantly larger than pre-instruction gain? | \(H_0: p_{post} - p_{pre2} = p_{pre2} - p_{pre1}\)  
\(H_1: p_{post} - p_{pre2} > p_{pre2} - p_{pre1}\) | \(H_1\) represents evidence for instructional sensitivity | |
| Control strategy 2A. Was posttest proportion correct for 5\(^{th}\) graders significantly larger than proportion correct for 4\(^{th}\) graders? | \(H_0: p_{postG5} = p_{G4}\)  
\(H_1: p_{postG5} > p_{G4}\) | \(H_1\) represents evidence for instructional sensitivity | |
| Control strategy 2B. Was pretest proportion correct for 5\(^{th}\) graders significantly larger than proportion correct for 4\(^{th}\) graders? | \(H_0: p_{preG5} = p_{G4}\)  
\(H_1: p_{preG5} > p_{G4}\) | \(H_1\) represents evidence that development influences performance | |
Conditions which may indicate insensitivity to instruction

- **No control.** 5\textsuperscript{th} grade post-instruction proportion correct ($p_{\text{post}}$) not significantly larger than pre-instruction proportion correct ($p_{\text{pre}}$).

- **Control strategy 1.** 5\textsuperscript{th} grade gain after instruction ($p_{\text{post}} - p_{\text{pre2}}$) not significantly larger than pre-instruction gain ($p_{\text{pre2}} - p_{\text{pre1}}$).

- **Control strategy 2A.** Posttest proportion correct for 5\textsuperscript{th} graders ($p_{\text{postG5}}$) not significantly larger than the proportion correct for 4\textsuperscript{th} graders ($p_{\text{G4}}$).
<table>
<thead>
<tr>
<th>Item</th>
<th>No control</th>
<th>Control strategy 1</th>
<th>Control strategy 2A</th>
</tr>
</thead>
<tbody>
<tr>
<td>08C</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>26C</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>27C</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>28C</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>30C</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>31C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32C</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>09P1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27P1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28P1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30P1</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>32P1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32P1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>08P2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26P2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27P2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28P2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30P2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31P2</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>32P2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45P2</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Results (Control strategy 2B)

• Only one item showed a significant difference between 4th graders and 5th grade pretest.
  – One significant result out of 20 items is what we would expect from 5% Type 1 error.
  – This suggests that fourth grade to fifth grade development had a minimal effect for these items.
Discussion

• Most close and near-proximal items showed significant pretest to posttest improvement.
• Only one far-proximal item showed significant pretest to posttest improvement.
• This evidence supports the DEISA approach to manipulating item proximity.
Discussion

• The repeated-pretest design (control strategy 1) flagged three items that were not identified as problematic by other criteria.
  – Performance improved from the first pretest to the second, suggesting that repeated exposure may have influenced performance.

• A practical disadvantage of the repeated-pretest design is that it interrupts the classroom schedule three times for testing.
Discussion

• Six items did not show significant differences between 4\textsuperscript{th} graders and 5\textsuperscript{th} grade posttest (control strategy 2A).

• All were identified as potentially insensitive by other criteria as well.
  – Thus this comparison contributed no additional information regarding item sensitivity.

• However, the 4\textsuperscript{th} grade data can also be used to investigate potential development effects (control strategy 2B).
Conclusion

• The demonstrated control strategies address specific threats to the internal validity of an instructional sensitivity interpretation.

• The statistical hypothesis tests are straightforward to perform.

• The potential for additional evidence from these control strategies should be weighed against the costs.
  – Instructional time
  – Test administration and scoring
Acknowledgements

• The Development and Evaluation of Instructionally Sensitive Assessment (DEISA) project is funded by the National Science Foundation (DRL-0816123).

• DEISA team members participating in item development and data collection:
  – Michael Giamellaro
  – Hillary Mason
  – Jen Feehan
Item 30C

Below is a topographic map of Mount Shasta, showing two trails from the base to the peak of the mountain.

What can you tell about the two trails?

A. Trail 1 reaches a higher elevation than Trail 2.
B. Trail 2 reaches a higher elevation than Trail 1.
C. Trail 1 is steeper than Trail 2.
D. Trail 2 is steeper than Trail 1.