Instructional Sensitivity as a Critical Component and Significant Challenge in Causal Effects Education Research

Chris Wilson, Molly Stuhlsatz & Kathy Roth, BSCS
Measurement in an Era of Accountability

NCLB and the associated accountability movement have influenced how states assess teacher and school effectiveness, but they have also resulted in a shift in the expectations for evidence in education research.

Many Federal policies now advocate for “evidence-based reform”, and decisions about effective educational interventions are based on research studies that demonstrate impacts on student learning (WWC).

This not only includes decisions about interventions at the student level (e.g. instructional approaches or curriculum materials), but also interventions at the teacher level.
“Education reform usually arrives with fanfare, great expectations, and overconfidence. Truth be known, typical education-reform effects tend to be small. Evaluations, if done at all, burst the reform balloon, having difficulty finding effects.”

Ruiz-Primo et al., 2002

It is one thing for a theory of change to describe how a change in instruction will impact student learning, it is quite another for a change in instruction to result in a difference in performance on student assessments.
# Measurement in an Era of Accountability

Effect sizes for educational interventions are SMALL!

<table>
<thead>
<tr>
<th>Types of Achievement Measure</th>
<th>Number of Effect Size Estimates</th>
<th>Mean Effect Size</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary School</td>
<td>389</td>
<td>0.33</td>
<td>0.48</td>
</tr>
<tr>
<td>Standardized test (<em>broad</em>)</td>
<td>21</td>
<td>0.07</td>
<td>0.32</td>
</tr>
<tr>
<td>Standardized test (<em>narrow</em>)</td>
<td>181</td>
<td>0.23</td>
<td>0.35</td>
</tr>
<tr>
<td>Specialized test</td>
<td>180</td>
<td>0.44</td>
<td>0.49</td>
</tr>
<tr>
<td>Middle School</td>
<td>36</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td>High School</td>
<td>43</td>
<td>0.27</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Hill, Bloom, Black, & Lipsey, 2008, Bloom et al., 2007
The Measurement Challenge:
1. An Instrument used to Assign Grades
The Measurement Challenge:
2. An Instrument to Measure Change over Time
The Measurement Challenge:
3. An Instrument to Differentiate between the Effects of Two Treatments over Time

*Treatment*

*Treatment*

*Comparison*

*Control for Pretest*

*Fidelity of Implementation and Achieved Relative Strength*
The Measurement Challenge:
4. An Instrument to Differentiate between the Effects of Two Treatments over Time of a Treatment at the Teacher Level on Learning at the Student Level.
We clearly need very precise measurement instruments...
But due to the scale of many efficacy studies, we are often limited to more crude measures.

(e.g. 200 teachers, pre and post, 4 topics, 30 student per class = ~48,000 student tests)
A Questionable Solution
Instructional Sensitivity
Lines of Research
Data Analysis: CRT

Comparing posttests...

adjusted for pretests

STeLLA
22 Schools
37 Teachers
925 Students

Comp
23 Schools
47 Teachers
1175 Students
Data Analysis

Level 1 (student or teacher)

\[ POST_{ij} = \pi_{0j} + \pi_{1jk} (PRE)_{ij} + e_{ij} \]

Level 2 (school)

\[ \pi_{0j} = \beta_{00} + \beta_{01} (TREAT)_{j} + r_{0j} \]
Item 9
You have a friend who lost 15 pounds of fat on a diet. Where did the mass go?
A) The mass was released as CO₂ and H₂O.
B) The mass was converted to energy which was used up.
C) The mass was converted to urine and feces and eliminated from the body.
D) The mass was converted into other substances inside the body that weighed less.

ALL OF THE ABOVE!
Monitoring the Difference between the Intended and Enacted Curriculum

- Enacted learning goals can differ significantly from the intended, both during the professional development program and during classroom teaching.

- Instruments may no longer be well aligned.

- May need to account for fidelity of implementation and/or achieved relative strength (ARS)
FINDINGS:
Increase in Instructional Sensitivity
## Effect Sizes for Each Outcome Measure

<table>
<thead>
<tr>
<th>Content/Participant</th>
<th>Mean (# correct)</th>
<th>N</th>
<th>SD</th>
<th>Effect Size (Hedges g)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lesson Analysis</td>
<td>Content Deepening</td>
<td>Lesson Analysis</td>
<td>Content Deepening</td>
</tr>
<tr>
<td>Food Webs Student</td>
<td>21.1</td>
<td>18.6</td>
<td>305</td>
<td>334</td>
</tr>
<tr>
<td>Food Webs Teacher</td>
<td>24.0</td>
<td>21.2</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Water Cycle Student</td>
<td>22.8</td>
<td>19.7</td>
<td>460</td>
<td>572</td>
</tr>
<tr>
<td>Water Cycle Teacher</td>
<td>28.9</td>
<td>24.7</td>
<td>17</td>
<td>25</td>
</tr>
</tbody>
</table>
Discussion

By focusing on test use and instructional sensitivity (validity) throughout instrument development, piloting, analysis and revision, one has a chance of overcoming seemingly impossible measurement challenges.

Evaluating Instructional Sensitivity for a research instrument must include piloting the instrument with instruction.

WWC and others must start taking measurement seriously.

“All of the Above” is often useful in developing instructionally sensitive instruments in science (as long as it is a distractor!)

Issues with measuring higher order cognitive skills and the NGSS
Instructional Sensitivity and the Next Generation
Science Standards: PERFORMANCE EXPECTATIONS

FROM: Students should know that:
• Individual organisms with certain traits are more likely than others to survive and have offspring.
• Natural selection leads to organisms that are well-suited for survival in particular environments.

TO: Students who demonstrate understanding can:
• Communicate scientific information that common ancestry and biological evolution are supported by multiple lines of empirical evidence
• Construct an explanation based on evidence for how natural selection leads to adaptation of populations.
• Create or revise a model to test a solution to mitigate adverse impacts of human activity on biodiversity.
Chris Wilson
cwilson@bscs.org