The Validity of Standardized Tests for Evaluating Curricular Interventions in Mathematics and Science

Joshua Sussman
Postdoctoral Scholar
Berkeley Evaluation and Assessment Research (BEAR) Center
University of California, Berkeley
Talk overview

• Three studies that examine the use of standardized academic tests for evaluating the impact of curricular interventions

• Analyze the validity (AERA, APA, & NCME, 2014) of the test for evaluating the intervention

• The studies lead to political and methodological solutions to an enduring problem in applied educational measurement.
Three studies: Research questions

1. How often do investigators use standardized tests to evaluate the impact of educational interventions; are the tests valid for their intended purpose?

2. How much alignment at the item level is necessary for valid evaluation?

3. What research designs can investigators use to mitigate validity problems with standardized tests as outcome measures?
About me

• The goal of my work is to advance applied measurement in schools.
• My research experience includes curriculum development projects funded by the Institute of Education Sciences (IES) and National Science Foundation (NSF). Dissertation research funded by an IES pre-doctoral fellowship in the Research in Cognition and Mathematics Education Program
• Experience in test construction and validation (Black racial identity, sustained attention, early childhood development, non-cognitive predictors of academic success, mathematics and science).
Reasons to evaluate educational interventions using standardized tests as outcome measures

• They are reliable measures of grade-level academic proficiency, in a major subject area, for groups of students.

• They provide a “fair” measure of the impact of an academic intervention.
  • Curriculum-independent and not subject to researcher biases or “training effects.”

• Schools are accountable for improving test scores
Problems with the use of standardized tests as outcome measures
Problems with the use of standardized tests as outcome measures: content mismatch

• What if the domain of the educational intervention is narrower than “mathematics?”
  • E.g., fractions

• The broad test design can be problematic.

• A longstanding consensus is that we should evaluate interventions by determining the degree to which the goals of the program are being realized in students (Baker, Chung, & Cai, 2016; Tyler, 1942).
Problems with the use of standardized tests as outcome measures: cognitive mismatch

• Standardized tests do not measure everything that is important in academic competence (Darling-Hammond et al., 2013; NRC, 2001).
  • Specific issues: NRC (2004) found serious problems with the validity of standardized tests in 86 evaluations of 25 different math curricula.

• New standardized tests in mathematics do a better job of measuring modern learning goals but serious shortcomings continue to exist (Doorey & Polikoff, 2016).

• In science, existing tests are not designed to measure the modern learning goals in the Next Generation Science Standards (DeBarger, Penuel, & Harris, 2013; Wertheim et al., 2016).
Study 1: A focus on prevalence and validity of standardized tests as outcome measures

1. How often do investigators use standardized tests as key outcome measures?

2. Are the tests valid?
   - Do the goals of the intervention appear to align with the measurement target of the standardized test?
   - Do investigators establish validity evidence for the specific use of the test per recommendations in the literature (AERA, APA, & NCME, 2014)?
   - Is the validity evidence adequate?
A focus on the alignment aspects of test validity

• Evaluate the validity evidence with an emphasis on the alignment between the tests and the interventions (Bhola, Impara, & Buckendahl, 2003; Roach, Niebling, & Kurz, 2009; Porter, 2002)

• A principled way to study the match between a test and an intervention
  • Content alignment
  • Cognitive process alignment

• Well developed investigations into the alignment between standardized tests and interventions are a relatively new area of the literature (e.g., May, Johnson, Haimson, Sattar, & Gleason, 2009)
Method

• A secondary analysis of 85 projects funded by the IES mathematics and science education program (2003 – 2015).

• Data sources
  a) IES database entries (study goals, description of intervention, key measures etc...)
  b) Reports to IES received from project PI’s
  c) Peer-reviewed articles associated with projects
  d) Test information on the internet
The prevalence of standardized tests as outcome measures

Analysis: Calculate the proportion of the projects that evaluated a curricular intervention using data from a standardized test

Results:

• Most projects developed and evaluated a curricular intervention (82%)

• Most intervention projects used, or planned to use, a standardized test for impact evaluation (72%)

• Thus, evaluation of new curricular interventions using standardized tests is widespread practice
The validity of standardized tests as outcome measures

Analysis: Three raters, using a validity rubric to score each project, reached consensus on the projects with misalignment between the intervention and the standardized test used as an outcome measure.

Results: The raters flagged 54% of the projects for a mismatch between the intervention and the test.

• Tests measured too much academic content
• Learning goals were difficult to measure with a typical standardized test
  • E.g., Conducting scientific investigations; participating in a learning community.
The validity of standardized tests as outcome measures

Analysis: For each project flagged for validity issues, the same three raters closely examined the corpus of data for validity evidence and to judge the adequacy of the validity evidence.

Data: Reports from PIs
• Emailed 68 unique PI’s for reports and 48 responded (70.6%)
• 33 PI’s provided reports
25 projects flagged

11 projects

33 reports provided

Reports from PIs
The validity of standardized tests as outcome measures

- Analyzed reports and published articles
Results: Validity discussions

• Five out of the 11 did not even mention validity issues.
• Six out of the 11 contained validity discussions.
Results: Adequacy of validity evidence

- Only one established adequate validity evidence
Measurement issues uncovered during the analysis

• The standardized test did not have enough test items that tapped the content taught by the intervention.

• I learned a lesson to “be more specific about the learning outcomes I want to measure and select an assessment that will be more sensitive to measuring those outcomes.”

• One investigator could not evaluate the intervention because the standardized test did not measure the appropriate construct.

• In follow up research, one investigator selected a subset of items from the test (i.e., the useful ones).
Summary

• Majority of projects engaged in applied research and evaluation using a standardized test
• About half of these projects were flagged as potentially problematic
• Only 6 of 11 projects established *any* validity evidence for the specific use of the test
• Only 1 of 11 established adequate validity evidence
Recommendations

• Cautiously interpret evaluations of new curricula that position data from standardized tests as the primary outcome measure—they may not provide accurate and useful information for data-based decision making.

• Careful item selection

• Proposals that include impact evaluation should require investigators to discuss measurement in detail
Study 2: How much alignment is enough for valid evaluation?

- In many cases, only a few items on the standardized test align with the intervention (Sussman, 2016).
- This data simulation study develops a psychometric model of the relationship between alignment and the *treatment sensitivity* of an evaluation defined as the ability of an evaluation to detect the effect of an educational intervention (Lipsley, 1990; May et al., 2009).
- The practical goal is to develop a method, akin to power analysis, that helps researchers account for misalignment when they design evaluations.
Alignment between a math test and an intervention

<table>
<thead>
<tr>
<th>Cognitive complexity</th>
<th>Academic content</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Addition</td>
</tr>
<tr>
<td>Single digit</td>
<td></td>
</tr>
<tr>
<td>Double digit</td>
<td></td>
</tr>
<tr>
<td>Double digit with carrying or borrowing</td>
<td></td>
</tr>
</tbody>
</table>

干预教这些领域

干预教这些领域

Intervention teaches this area
Method

• Data simulation of hypothetical evaluations with an outcome measure that is more or less aligned with an intervention.
• The primary outcome is the average statistical power, calculated as a function of test alignment and intervention effect size.
  • Power to detect a true difference between experimental and control.
• Psychometric models for data generation and for data analysis from the Rasch family of item response models (Rasch, 1960/1980; Adams, Wilson, & Wang, 1997).
Key assumptions of the simulation

• Effective treatments increase the probability that a student succeeds on an a test item that is aligned

• The treatment has no impact on an item that is considered not aligned

• The control group is unaffected
Simulation variables

Hold sample size consistent but vary alignment over a range of effect sizes for the intervention.

• Fixed sample size ($N = 600$; 300 each in experimental and control groups)
• Fixed test length ($N = 50$ items)
• Vary alignment (1 – 50 items)
• Vary effect size of the intervention (0.1 – 2.0 $SD$)
Results

A graph showing the relationship between Simulated Alignment (%) on the x-axis and Statistical Power on the y-axis. The graph includes curves for different Effect Sizes (0.1 to 2.0).
Conclusions

• Alignment should be no less than 60%, for adequate statistical power to detect treatment effects with an effect size of 0.2 $SD$.

• Researchers must balance alignment against sensitivity to detect small effect sizes.

• Use of multiple measures with different levels of alignment represent an ideal scenario for developing a compelling evaluation argument (Cronbach, 1963; House, 1977; Penuel, 2016).
Study 3: One solution to the alignment problem

- Research methods that coordinate data and theory can present stronger arguments for the efficacy of an intervention.

- Empirical study that documents the effectiveness of the *Learning Mathematics through Representations* (LMR) lesson sequence for teaching English Learners (ELs) mathematics.

- The evaluation coordinates data from a researcher developed test and a standardized test with theory about how the features of LMR meet the needs of ELs in the mathematics classroom.
Learning Mathematics through Representations (LMR)

• LMR is a 19-lesson number line-based curriculum unit that supports upper elementary students’ understandings of integers and fractions (Saxe, de Kirby, Le, Sitabkhan, & Kang, 2015)

• The unit supports mathematical learning through (a) the use of the number line as a central representational context, and (b) the building of mathematical definitions in classroom communities that become resources to support student argumentation and problem solving.
Method

- 571 students in 21 classrooms (4th and 5th grade) containing both ELs and English Proficient (EP) students participated in a quasi-experimental study.
  - There were 95 ELs in the sample: 44 ELs in 11 LMR classrooms and 51 ELs in 10 comparison classrooms.

- Students completed a set of four (pre, interim, post-test, and follow up) researcher developed assessments of integers and fractions

- Students also completed the state test in mathematics in the prior year and the end of the intervention year
The empirical results support the efficacy of LMR for students classified as ELs

• Multilevel analysis revealed that the ELs in LMR classrooms gained more in mathematics than the ELs in the matched comparison group on both an assessment of integers and fractions ($p = 0.011; ES = 0.68$) and a standardized assessment in mathematics ($p = 0.010, ES = 0.49$)

• LMR eliminated or narrowed the achievement gap between ELs and EPs

• In addition, theory supports LMR’s potential as a mathematics intervention benefitting ELs’ achievement; narrowly (integers and fractions) and broadly (grade-level achievement)
Two resources for meeting the needs of ELs in the mathematics classroom

1. Participation in mathematical communication & argumentation (Darling-Hammond, 2007; Moskovich, 2012; NCTM, 2000; Schoenfeld, 2002).

2. Multimodal opportunities for learning using visual and embodied representations (Bustamante & Travis, 1999; Hakuta & Santos, 2012; Moschkovich, 1999, 2002; Schleppegrell, 2007.)
Providing ELs access to participating in mathematics lessons

1. Participation in mathematical communication & argumentation

2. Multimodal learning (visual and embodied representations)
Providing ELs access to mathematical discussions

1. Participation in mathematical communication & argumentation

2. Multimodal learning (visual and embodied representations)
Visual resources for mathematical learning

1. Participate in mathematical communication & argumentation

2. Multimodal opportunities for learning (visual and embodied representations)
Embodied resources for mathematical learning

1. Participation in mathematical communication & argumentation

2. Multimodal opportunities for learning (visual and embodied representations)
Meeting the needs of ELs in the mathematics classroom

1. Participation in mathematical communication & argumentation

2. Multimodal opportunities for learning (visual and embodied representations)
Conclusions

• Standardized tests have a place and purpose, but they need to be well aligned to serve as outcome measures.

• Alignment should be no less than 60% to detect reasonable effect sizes (0.2 $SD$).

• High quality evaluations of educational interventions coordinate data and theory.
Plans for future research

• Measurement in special education
• Measuring progress towards Individualized Education Plan goals
• Support data-based decision making for future student eligibility, goals and services (interventions).
References


Sussman, J. Standardized tests as outcome measures in applied research: A psychometric simulation of the relationship between alignment and treatment sensitivity. To be submitted to Applied Measurement in Education.


jsussman@berkeley.edu