Lessons Learned About the Implementation of Large-Scale Evaluations:

- Outcomes -

Douglas H. Clements
University at Buffalo, SUNY
Especially given their expense, large-scale evaluations should answer a comprehensive array of questions...

<table>
<thead>
<tr>
<th>Effect</th>
<th>Practice</th>
<th>Policy</th>
<th>Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective in</td>
<td>Effective in achieving</td>
<td>Effect size?</td>
<td>Why effective?</td>
</tr>
<tr>
<td>Learning goals?</td>
<td>curriculum goals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credible relative to</td>
<td>important?</td>
<td></td>
<td>Why credible relative to</td>
</tr>
<tr>
<td>alternatives?</td>
<td></td>
<td></td>
<td>alternatives?</td>
</tr>
<tr>
<td>Conditions</td>
<td>When and where?</td>
<td>Support requirements for</td>
<td>Why do conditions in (de)</td>
</tr>
<tr>
<td></td>
<td>Under what conditions?</td>
<td>various contexts?</td>
<td>crease effects?</td>
</tr>
<tr>
<td></td>
<td>Generalize?</td>
<td></td>
<td>How &amp; why do strategies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>produce previously unattained</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>results?</td>
</tr>
</tbody>
</table>
Requirements for Outcome Measures

- **Instructional**: Beyond fidelity:
  - extensions of “compliance” fidelity and
  - outcome measures of quantity and quality of environment and teaching for all treatment groups

- **Student**: Valid measures of child outcomes
Measures of Instruction

• Address “deep change” that “goes beyond surface structures or procedures… to alter teachers’ beliefs, norms of social interaction, and pedagogical principles” (Coburn, 2003, p. 4).

• We reviewed research concerning elements of successful, engaging instruction

• Two related measures:
  • Proximal: Fidelity
  • Distal: COEMET (Classroom Observation of Early Mathematics Environment and Teaching)
COEMET

• Quality of the mathematics environment, interactions, and activities
• Half-day observation
• Not connected to any curriculum
  • Thus, allows for experimental-comparison contrasts no matter what the source of the enacted curriculum
  • Distal (by analogy to transfer, “far fidelity”)
• Wide variety of items, empirically reduced to most predictive set

• Most Likert (SD D N A SA) or %

• Two sections, Specific Math Activity and Classroom Culture

• Specific done multiple times as needed; culture once
COEMET

- Culture
  - Environment and interaction (e.g., “interacted with and responsive to children” “used computers”)
  - Personal attributes (“knowledge, confident about math”)

- Activities
  - Mathematical focus (“appropriate content”)
  - Organization, teaching, interactions
  - Eliciting, supporting, extending understanding, strategies
  - Assessment and instructional adjustment
• Rasch model (all instruments)
• True interval scale (especially important for analyses of gains)
• In large-scale data collection, must be robust under conditions of missing items
• Especially for child outcomes, single scale to evaluate change over the duration of the intervention...and beyond
Psychometrics of the COEMET

- **Reliability**
  - *Inter-rater*: 88% exact; 100% of the disagreements had same polarity
  - *Rasch person reliability*: .96

- **Validity**
  - *Content*: Research on effective EC math
  - *Concurrent*: Correlation with teacher rankings
  - *Predictive*: Accounted for sig. amount of variance in children’s gain scores (correlation: .50)
Fidelity

• Based on same body of research

• *Proximal*, but beyond simple “compliance”
  
  • to evaluate *quality* and *quantity* of interaction, environment, and activities

  • many approaches require creative, adaptive enactment, not following script

  • so, philosophical foundations influence interpretation of fidelity on a *continuum from compliance to consonance* of enactment to a particular educational vision

• Ours emphasizes the interactions of teachers and children around activities
Child Outcome Measures

- General criteria for large-scale research
- Standardized instruments chosen or developed as valid measures of the curricular goals
- May be at least two assessment components:
  - *distal*: shared goals of all curricula + far transfer
  - *proximal*: unique goals + near transfer
    - because comparison curricula ideally selected on a principled basis, and use only of “traditional” curriculum less useful, goals of > 2 possible
- Instruments should be sufficiently valid, reliable, and differentiated to measure nuanced differences in various content and process areas.
PCER Measures for Math

• Commercial measures
  • E.g., Woodcock-Johnson III: Applied Problems
    • Too few items at youngest items, 8
    • Narrow—7 problems deal with sets of 3 in slightly varying contexts; 1 concrete 3 + 1. No other areas.
    • Gap developmentally—jump to quite advanced, formal knowledge
    • Not all subtests given to youngest children and reliability not reported for math for young ages.
  • Need comprehensive, sensitive measure
Additional PCER Measures

- Improvements
- More sensitive
- Broader coverage
- Some limitations remained
- Not all items well connected to later mathematics
- Need instrument for longitudinal evaluation, a single conjoint scale
Based on theory and consensus regarding content, beyond a single intervention
And measure the same latent trait throughout the early years
Learning Trajectories

• Three components
  • Goal, developmental progression, tasks
    • descriptions of children’s thinking and learning in a specific mathematical domain, and a related, conjectured route through a set of instructional tasks designed to engender those mental processes hypothesized to move children through a developmental progression of levels of thinking, created with the intent of supporting children’s achievement of specific goals in that mathematical domain (special issue, MTL)

• For assessment, developmental progression
  • explication of the mental constructions (actions-on-objects) constitute children’s thinking at each level
  • how they are incorporated in each subsequent level
Building Blocks Assessment

- Individual interview format, with explicit protocol and scoring procedures.
- Abilities assessed according to theoretically- and empirically-based developmental progressions.
- Accurate answers and solution strategies.
- Entry point determined via initial screening.
- Floor (4 consecutive correct) and ceiling (4 consecutive errors).
- Appropriate for children with disabilities:
  - individually administered.
  - no strict time limits on responses.
  - provides materials and illustrations.
  - allows for varied response formats.
Building Blocks Assessment

- Number
  - verbal counting (forward, back, before/after; identifying mistakes)
  - object counting
  - subitizing
  - comparison
  - sequencing
  - numerals
  - de/composition
  - add/subtracting
  - place value

- Geometry
  - shape identification
  - de/composition
  - congruence
  - construction
  - turns

- Measurement

- Patterns
Computer Assessment and Reporting

- Along the learning trajectories
- Automatic data collection (item level for drills, learning trajectory levels for other)
Building Blocks
Research Results
Small-Scale Summative Evaluation

Summative Research: Effect Sizes—Number

Score

Pretest
Posttest

ES .85
ES 1.70

9.744
20.566
32.342
Summative Research: Effect Sizes—Geometry

(Bloom’s famous “2-sigma effect”)
Measures

- Assessed reliability, validity of all instruments
- Quantitative analyses to eliminate non-predictive and redundant items, yielding a parsimonious instrument
- Qualitative observations to triangulate, evaluate validity
Limitations

• Small scale
  • low $n$; 4 classrooms
• child was unit of analysis

• Ideal conditions
  (Cronbach’s “Hyper-realization”)

• Therefore….
Building Blocks

Large-Scale Research

(see Web sites at end for papers)
Research Questions

• Can *Building Blocks* be implemented with high fidelity?

• Does *Building Blocks* have substantial positive effects on the quality of the mathematics environment and teaching?

• On children’s mathematics achievement?

• If so, does 2 mediate 3?
Results—Fidelity

- BB averaged 3.0 on Likert Scale (Comparison was 2.8)

- Rasch T Score showed no sig. difference by treatment, time, or interactions.

- Correlated positively, but not sig., with child gain (but volunteers...wait for TRIAD II; e.g., our phantom teacher)
Classroom Observation

• No sig. change over time or interactions—differences from beginning were maintained

• Building Blocks sig. more math, higher quality math, than control (p < .001, ES = 1.25), marginally than Comparison (p = .06)

• Predictive of child gain: r = .50

• Highest: % teacher active, built on/elaborated children’s ideas/strategies, and facilitated children’s responding
Results: Child Assessment

- $F(1, 32) = 40.52, \ p = .000+$
- T Scores:
  - 50 Mean
  - 10 SD
Mediational Hypothesis

• Does Classroom Observation mediate the effects of treatment on child gain scores?
  • Baron and Kenney…

• Not strong evidence, but reduced variance, so evidence for partial mediation
Research Q&A

• *Building Blocks* can be implemented with high fidelity in diverse settings; no changes

• *Building Blocks* has a positive effect on the quality of the mathematics environment and teaching

• *Building Blocks* has a substantial positive effect on mathematics achievement

• Some evidence that 2 mediates 3
Conclusions

- Large-scale evaluations should answer a comprehensive array of research questions
- To do so, need both *instructional* and *student* levels
- Measures should be
  - sensitive
  - comprehensive, connected to proximal and distal goals
  - theoretically grounded (e.g., research base for observation of instruction; developmental sequences for student outcomes)
Conclusions: Observations

- Need both *proximal* and *distal* measures
- Both address “deep change” and
- *Related, both based on common research foundation*
Conclusions: Observations

• Full individual observations (not “pieced together”) more important than frequency

• 2-3 observations adequate for evaluation

• although more frequent mentoring/coaching visits are promote professional development (especially analysis) and fidelity
Conclusions

• Measure coherent, unidimensional latent traits, producing interval measures that are linear and additive.

• True interval scale important for analyses of gains

• Robust under conditions of missing items

• Single scale to evaluate change over the duration of the intervention
Conclusions

• Connections among measures; otherwise, inadequate basis for contributing to theories of learning and teaching in complex settings
Randomized Trials… and Tribulations

• Teacher sharing
  (We needed the COEMET)
• IRB — Who is the subject?
  (district vs. teacher)
• Principal—My teachers are fine, so…
• Seniority/union rules—teachers moving in the middle; limited observations