Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures

Rebecca Thornton
University of Illinois at Urbana-Champaign

Jason Kerwin
University of Minnesota
How to improve learning?

- Hundreds of program effectiveness studies
- Meta-analyses, systematic reviews, meta-meta-analyses
“What works”

Figure 2: Effects of treatments on language or composite test scores

Mean = 0.084
p-value = 0.014

McEwan (2015)
How to (use research to) improve learning?

- Never able to “replicate” implementation of a program
  - Context
  - Budget constraints
  - Logistical constraints
  - Are enough details even provided?
What happens as we modify a program?

- Large variation in effectiveness across programs
  - Across setting & intervention type
  - Within setting & intervention type (Evans and Popova 2016b, Vivalt 2017)

- Most evidence across studies, not within (McEwan 2015)

- This paper: examines variation within a single study, holding context and intervention type constant
Outcomes from a single study

Figure 2: Effects of treatments on language or composite test scores

Mean = 0.084
p-value = 0.014
Mango Tree Literacy Program

• Community engagement

• Pedagogy
  • Mother tongue, slower pace, phonics, scripted lessons

• Materials

• Teacher training and support
  • Each term: 1 residential + 3 non-res workshops, 3 class visits
Reduced-cost version

• Modified to resemble implementation at scale
  1. Cascade model of training and support (non-res)
  2. Fewer classroom visits (5 vs. 2)
  3. No slates or wall clocks
Differences?

- **Teacher training indicators** (Arancibia, Popova, and Evans 2016)
  - Codes 26 teacher training programs, including NULP
  - Out of 51 indicators, three (5.9 percent) differ

- 325 pairwise combinations, compute % indicators different
Research design

- 38 primary schools
- 50 grade one students/school (N=1,900)

Public randomization
- Control (Government status quo)
- Full-cost program (Mango Tree)
- Reduced-cost program (Cascade model)
Research design

• Exams
  • Baseline
  • Endline (78% of baseline, N=1481)
  • Outside examiners blinded to study arm

• Learning
  • Reading Leblango (EGRA)
  • Writing Leblango (EGWA)

• Results
  • Each module + PCA index
  • Normalize
  • Randomization inference p-values
Program effects on writing

[Graph showing the effects of program on writing, with categories such as Index, Last name, First name, Ideas, Organization, Voice, Word choice, Sentences, Conventions, Presentation, Full-cost vs Reduced-cost]
Mechanisms

• What led to the success of the full-cost program?
• What led to the failure of the reduced-cost program?

• Use classroom observations data to explore mechanisms
  1. Time on task
  2. Productivity
  3. Complementarities
Classroom observations

- 3 visits: two 30-min literacy lessons/classroom
- Factors

<table>
<thead>
<tr>
<th>Specific Lesson Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
</tr>
<tr>
<td><strong>FIRST</strong></td>
</tr>
<tr>
<td>10 minutes:</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Other:</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Productivity?

Returns to time on task (SDs)

<table>
<thead>
<tr>
<th></th>
<th>Full-cost program</th>
<th>Reduced-cost program</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hour reading</td>
<td>0.011</td>
<td>0.004</td>
<td>0.011</td>
</tr>
<tr>
<td>Hour writing</td>
<td>0.024</td>
<td>0.008</td>
<td>0.002</td>
</tr>
</tbody>
</table>

- Reading: more time on sounds (not sig between full & reduced)
- Writing: more focus on names
- Teachers more engaged in both reading and writing
Complementarities?

• Did not randomize each input

• Across inputs (materials, human capital)
  • Reading: more use of materials (& in full-cost)
  • Writing: large differences

• Across skills (reading and writing, advance and basic)
Complementarities?

- Mediation analysis with linear regression (Sequential $g$-estimator Acharya, Blackwell, and Sen 2016)

  \[
  \%	ext{ explained by mediators}
  \begin{array}{cc}
  \text{EGRA} & \text{Writing} \\
  0.020 & 0.037 \\
  \end{array}
  \]

- Machine-learning allowing for complementary inputs and non-linearities

  \[
  \text{R-squared}
  \begin{array}{cc}
  \text{EGRA} & \text{Writing} \\
  0.80 & 0.99 \\
  \end{array}
  \]
Summing up

• Massive gains in learning possible
  Even in most resource-deprived schools and using existing teachers

• Small changes in inputs may dramatically change program effectiveness

• Cutting costs may leave some students worse off

• What we measure is crucial for seeing the entire picture
Implications

• Researchers
  • Focus on isolating individual inputs
  • Systematically underestimate possible effects
  • Attention to types of learning metrics

• Policy makers
  • Limited resources – infeasible to provide most-effective programs
  • Almost always need to modify or eliminate some inputs

• Take advantage of complementarities, rather than focus on individual inputs

• What is the most effective research for knowing what/how to implement what works?