What Are We Learning about Learning from Experiments? (It Depends.)

Justin Sandefur

Center for Global Development

July 17, 2013
Outline

Cataloging what works
- Lexicographic evidence rankings
- Highly clustered evaluations
- Meta-analysis
- Global policy prescriptions

What have we learned?
- Contract teachers
- Class size
- Private schools

Where are we going?
- Rigor versus relevance
- Evaluation & scale-up
- Conclusions
Outline

Cataloging what works
- Lexicographic evidence rankings
- Highly clustered evaluations
- Meta-analysis
- Global policy prescriptions

What have we learned?
- Contract teachers
- Class size
- Private schools

Where are we going?
- Rigor versus relevance
- Evaluation & scale-up
- Conclusions
Outline

Cataloging what works
   Lexicographic evidence rankings
   Highly clustered evaluations
   Meta-analysis
   Global policy prescriptions

What have we learned?
   Contract teachers
   Class size
   Private schools

Where are we going?
   Rigor versus relevance
   Evaluation & scale-up
   Conclusions
1 of 4: Lexicographic evidence rankings

- Randomization
  - Yes
    - Attrition
      - Low
        - Meets Evidence Standards
      - High
        - Equivalence
          - Yes
            - Meets Evidence Standards with Reservations
          - No
            - Does Not Meet Evidence Standards
  - No
Outline

Cataloging what works
- Lexicographic evidence rankings
- Highly clustered evaluations
- Meta-analysis
- Global policy prescriptions

What have we learned?
- Contract teachers
- Class size
- Private schools

Where are we going?
- Rigor versus relevance
- Evaluation & scale-up
- Conclusions
Caroline Fiennes, SSIR 2013, “Most Charities Shouldn’t Evaluate Their Work”

\[ \text{impact} = \text{idea} \times \text{implementation} \]

“the ideas used by charities don’t need to be evaluated again, because they’ve been amply evaluated already. . . . All the charity then needs to do is run the programs well.”
Outline

Cataloging what works
Lexicographic evidence rankings
Highly clustered evaluations
Meta-analysis
Global policy prescriptions

What have we learned?
Contract teachers
Class size
Private schools

Where are we going?
Rigor versus relevance
Evaluation & scale-up
Conclusions
3 of 4: Meta-analysis
Outline

Cataloging what works
- Lexicographic evidence rankings
- Highly clustered evaluations
- Meta-analysis
- Global policy prescriptions

What have we learned?
- Contract teachers
- Class size
- Private schools

Where are we going?
- Rigor versus relevance
- Evaluation & scale-up
- Conclusions
<table>
<thead>
<tr>
<th>Program costs</th>
<th>Source and method of calculation</th>
<th>Randomized?</th>
<th>Cost (US$2000 millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School inputs (uniforms and textbooks)</td>
<td>Source: Kremer, Moulin, and Namanya 2013. Calculation: Unit InputCost from Kremer et al. (p. 44). Transportation costs ignored. Country cost = InputCost × SchoolAge × PPP Correction</td>
<td>Yes</td>
<td>2,268</td>
</tr>
<tr>
<td>Schooling vouchers</td>
<td>Source: Angrist et al. 2002. Calculation: Assumes that everyone is sufficiently motivated to achieve satisfactory performance, hence qualifying for the voucher, and ignoring general equilibrium effects due to the resultant increase in private school fees. Unit VoucherCost used in our calculations is the increase in public educational expenditure per lottery winner, given in Angrist et al. (153). This is multiplied by four because at any time there are four cohorts in high school. Country cost = (4/15) × VoucherCost × ChildPop + GDP Correction</td>
<td>Yes</td>
<td>1,478</td>
</tr>
<tr>
<td>Monetary rewards to parents for sending children to school</td>
<td>Source: Behrman, Segura, and Todd 2001. Calculation: Assumes that if the subsidy is large enough, everyone will wants to send their children to school, and therefore everyone will get the subsidy. Unit SubsidyCost is calculated from data given in Behrman et al. (2001). This is multiplied by seven because at any time there are four cohorts getting the subsidy. Country cost = (7/15) × SubsidyCost + ChildPop + GDP Correction</td>
<td>Yes</td>
<td>2,3142</td>
</tr>
</tbody>
</table>
Outline

Cataloging what works
- Lexicographic evidence rankings
- Highly clustered evaluations
- Meta-analysis
- Global policy prescriptions

What have we learned?
- Contract teachers
- Class size
- Private schools

Where are we going?
- Rigor versus relevance
- Evaluation & scale-up
- Conclusions
Outline

Cataloging what works
- Lexicographic evidence rankings
- Highly clustered evaluations
- Meta-analysis
- Global policy prescriptions

What have we learned?
- Contract teachers
- Class size
- Private schools

Where are we going?
- Rigor versus relevance
- Evaluation & scale-up
- Conclusions
Outline

Cataloging what works
  Lexicographic evidence rankings
  Highly clustered evaluations
  Meta-analysis
  Global policy prescriptions

What have we learned?
  Contract teachers
  Class size
  Private schools

Where are we going?
  Rigor versus relevance
  Evaluation & scale-up
  Conclusions
Contract teachers

- Muralidharan & Sundararaman (2008)
  Andhra Pradesh
  Contract teachers ⇒ +0.15 std. dev.

- Duflo, Dupas, & Kremer (2009)
  Western Kenya
  Contract teachers ⇒ +0.21 std. dev.
  Class size reduction ⇒ no effect on scores
Treatment Effect of Contract Teachers on Test Scores

ITT Effect

Overall: 0.075
NGO: 0.225
MOE: -0.025
Outline

Cataloging what works
   Lexicographic evidence rankings
   Highly clustered evaluations
   Meta-analysis
   Global policy prescriptions

What have we learned?
   Contract teachers
   Class size
   Private schools

Where are we going?
   Rigor versus relevance
   Evaluation & scale-up
   Conclusions
## Class-size effects

<table>
<thead>
<tr>
<th>Country</th>
<th>No Controls</th>
<th>Controlling for observables</th>
<th>Controlling for unobservables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RDD</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angrist &amp; Lavy (1999)</td>
<td>Israel</td>
<td>0.322</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.039)</td>
<td>(.044)</td>
</tr>
<tr>
<td>Urquiola (2006)</td>
<td>Bolivia</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.03)**</td>
<td>-0.03</td>
</tr>
<tr>
<td>Asadulla (2005)</td>
<td>Bangladesh</td>
<td>0.25</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.115)**</td>
<td>(1.03)**</td>
</tr>
<tr>
<td><strong>RCT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krueger (1999)</td>
<td>USA</td>
<td>-0.271</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.072)**</td>
<td>(.072)**</td>
</tr>
<tr>
<td>Banerjee et al (2007)</td>
<td>India</td>
<td>0.027</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0125)**</td>
<td>(.118)</td>
</tr>
<tr>
<td>Duflo et al (2012)</td>
<td>Kenya</td>
<td></td>
<td>-0.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.024)**</td>
</tr>
</tbody>
</table>
Class-size effects: Woessman & West (2006) TIMSS data
Class-size effects: Lessons

1. Most striking feature is high variance, not high mean.

2. Heterogeneity is real (statistically significant) and affects both treatment and selection parameters.

3. No clear encompassing theory to explain widely variant OLS and (quasi-) experimental results – as required for any external validity claims.
Outline

Cataloging what works
  Lexicographic evidence rankings
  Highly clustered evaluations
  Meta-analysis
  Global policy prescriptions

What have we learned?
  Contract teachers
  Class size
  Private schools

Where are we going?
  Rigor versus relevance
  Evaluation & scale-up
  Conclusions
## Return to Private Schooling

<table>
<thead>
<tr>
<th>Country</th>
<th>Point estimate &amp; std. error</th>
<th>Controls</th>
<th>Observables</th>
<th>Unobservables</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cox &amp; Jimenez (1991)</td>
<td>Colombia</td>
<td>0.22</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Cox &amp; Jimenez (1991)</td>
<td>Tanzania</td>
<td>-0.14</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>Aggregation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsieh &amp; Urquiola (2006)</td>
<td>Chile</td>
<td>-0.714</td>
<td>-0.51</td>
<td>(1.390)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.188)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tabarrok (2013)</td>
<td>India</td>
<td>0.224</td>
<td></td>
<td>(0.036)***</td>
</tr>
<tr>
<td>Bold et al (2012)</td>
<td>Kenya</td>
<td>0.79</td>
<td>0.98</td>
<td>(0.41)**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.046)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angrist et al (2002)</td>
<td>Colombia</td>
<td>0.379</td>
<td>0.291</td>
<td>(0.153)**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.111)***</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

Cataloging what works
- Lexicographic evidence rankings
- Highly clustered evaluations
- Meta-analysis
- Global policy prescriptions

What have we learned?
- Contract teachers
- Class size
- Private schools

Where are we going?
- Rigor versus relevance
- Evaluation & scale-up
- Conclusions
Outline

Cataloging what works
  Lexicographic evidence rankings
  Highly clustered evaluations
  Meta-analysis
  Global policy prescriptions

What have we learned?
  Contract teachers
  Class size
  Private schools

Where are we going?
  Rigor versus relevance
  Evaluation & scale-up
  Conclusions
Outline

Cataloging what works
  Lexicographic evidence rankings
  Highly clustered evaluations
  Meta-analysis
  Global policy prescriptions

What have we learned?
  Contract teachers
  Class size
  Private schools

Where are we going?
  Rigor versus relevance
  Evaluation & scale-up
  Conclusions
OLS ‘here’ vs an RCT ‘there’

For non-experimental estimates:

$$\text{MSE}(\tilde{\beta}_k) = \text{Var}(\tilde{\beta}_k) + (\tilde{\beta}_k - \beta_k)^2$$

- **Sampling error**
- **Omitted var. bias**

For experimental estimates:

$$\text{MSE}(\hat{\beta}_j) = \text{Var}(\hat{\beta}_j) + \text{Var}(\beta)$$

- **Sampling error in context j**
- **Variance of true effect across contexts**
Experimental estimate from context \( j \) is a better estimate of the true causal effect in context \( k \)

Non-experimental estimate from context \( k \) is a better estimate of the true causal effect in context \( k \)
Outline

Cataloging what works
  Lexicographic evidence rankings
  Highly clustered evaluations
  Meta-analysis
  Global policy prescriptions

What have we learned?
  Contract teachers
  Class size
  Private schools

Where are we going?
  Rigor versus relevance
  Evaluation & scale-up
  Conclusions
Alternative evaluation models

Straw man model?
▶ Define an intervention defined at school level; evaluate in a controlled context; extrapolate.

Alternative models:
▶ Evaluate scale-up within large programs (Duflo & Kiessel 2013, Ghana TCAI).
▶ Mechanism experiments (Ludwig, Kling, and Mullainathan) applied within organizations to production of outputs.
▶ Move up the bureaucratic supply chain (Rasul & Roger 2013, Nigerian Civil Service).
Teacher Community Assistants Initiative

A project of the Ghana Education Service in collaboration with:

Ghana National Association of Teachers (GNAT)
National Youth Employment Program (NYEP)
Innovations for Poverty Action (IPA)
Outline

Cataloging what works
Lexicographic evidence rankings
Highly clustered evaluations
Meta-analysis
Global policy prescriptions

What have we learned?
Contract teachers
Class size
Private schools

Where are we going?
Rigor versus relevance
Evaluation & scale-up
Conclusions
Conclusions

For evaluators:

▶ Specify context: Intervention may be both bigger and narrower than assumed.
▶ Encompass rather than trump: OLS estimates are facts demanding explanation.

For evaluation users:

▶ Avoid lexicographic preference for internal over external validity.
▶ Commission more, faster, cheaper RCTs with fewer grand ambitions.