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

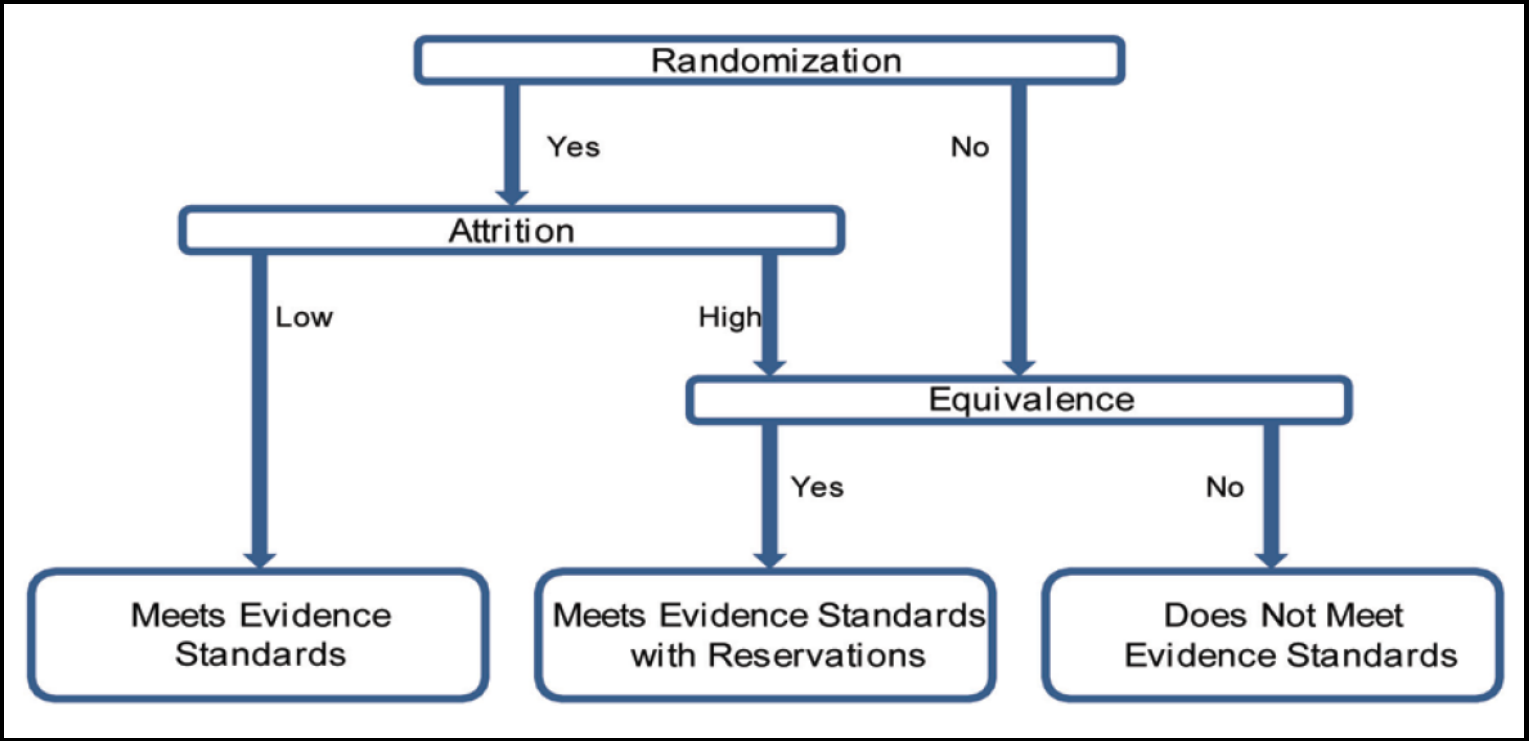
Outline

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



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

2 of 4: Highly clustered evaluations

Caroline Fiennes, *SSIR* 2013, "Most Charities Shouldn't Evaluate Their Work"

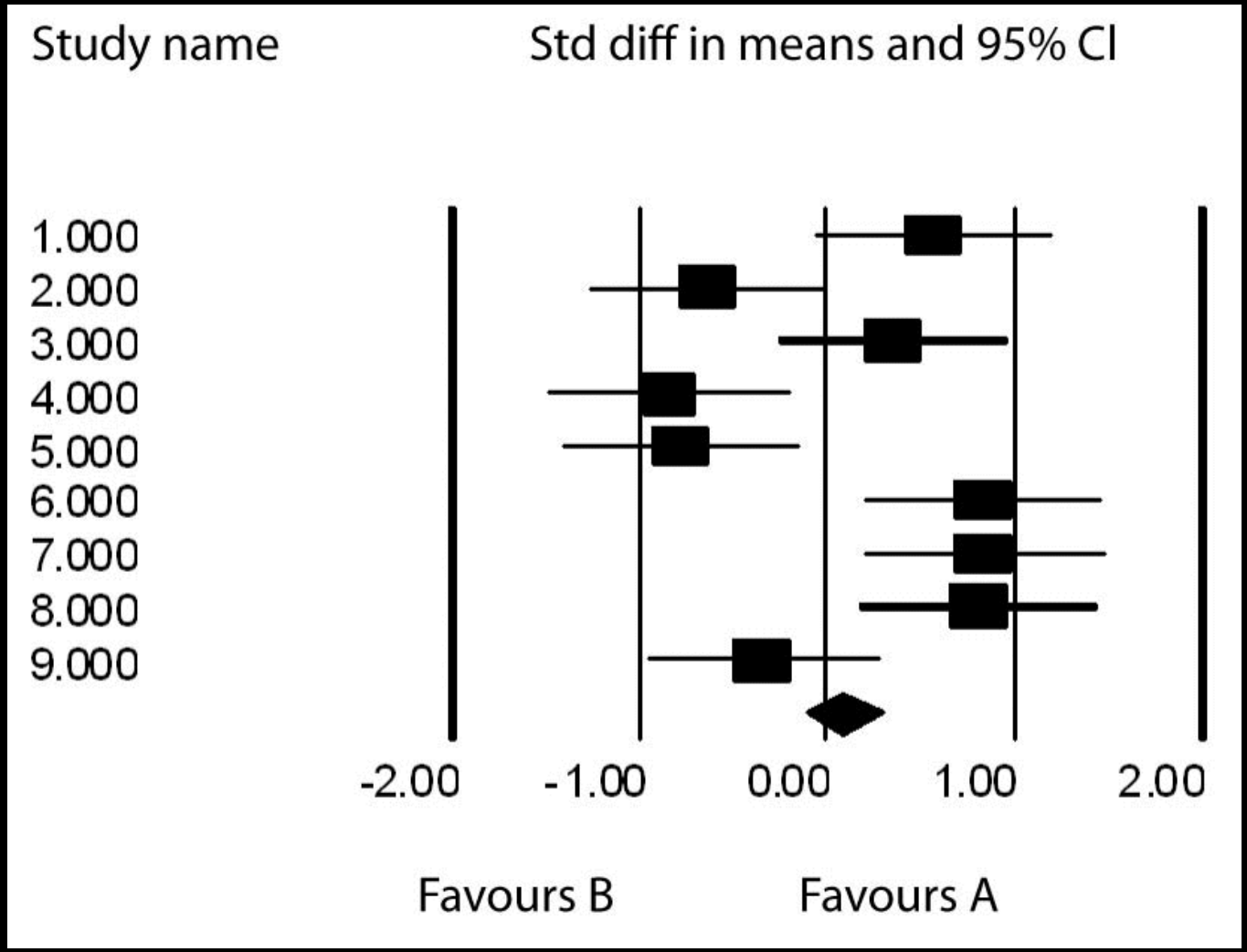
$$impact = idea \times implementation$$

"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

4 of 4: Global policy prescriptions

Table 2.4
Program costs

Program (recurring annual expenditure)	Source and method of calculation	Randomized?	Cost (US$2000 millions)
Education			
Remedial teaching based on the Balsakhi model developed by Pratham	Source: Banerjee et al. 2003. Calculation: Unit (cost per child-year), *BalkhashiCost*, from the Pratham 2001–2002 Reports (cited on June 28, 2003). Available at www.pratham.org/reports. We use the mean of costs reported for the Dehli, Mumbai, and Pune regions. Country cost = *BalkashiPop* * *BalkhashiCost* * *GDPCorrection*	Yes	644
Universal education based on a 40:1 pupil-teacher ratio	Source: Angrist and Lavy 1999[1] Calculation: Country cost = *NotInSchool* * *SchoolCost*	No	1,544
School inputs (uniforms and textbooks)	Source: Kremer, Moulin, and Namunyu 2003. Calculation: Unit *InputsCost* from Kremer et al. (p. 44). Transportation costs ignored. Country cost = *InputsCost* * *SchoolAge* * *PPPCorrection*	Yes	2,268
Schooling vouchers	Source: Angrist et al. 2002. Calculation: Assumes that everyone is sufficiently motivated to achieve satisfactory performance, hence qualifying for the vouchers, 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. (1535). This is multiplied by four because at any time, there are four cohorts in high school. Country cost = (4/15) * *VoucherCost* * *ChildPop* * *GDPCorrection*	Yes	1,478
Monetary rewards to parents for sending children to school	Source: Behrman, Segupta, 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, 1). This is multiplied by seven because at any time there are four cohorts getting the subsidy. Country cost = (7/15) * *SubsidyCost* * *ChildPop* * *GDPCorrection*	Yes	2,3142
Nutrition supplementation			
Iron	Source: Bobonis, Miguel, and Sharma 2004. Calculation: Unit *IronCost* data from Miguel and Bobonis, private communication. The program covers five cohorts aged between ages two and six. Country cost = (5/15) * *IronCost* * *ChildPop* * *PPPCorrection*	Yes	346

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

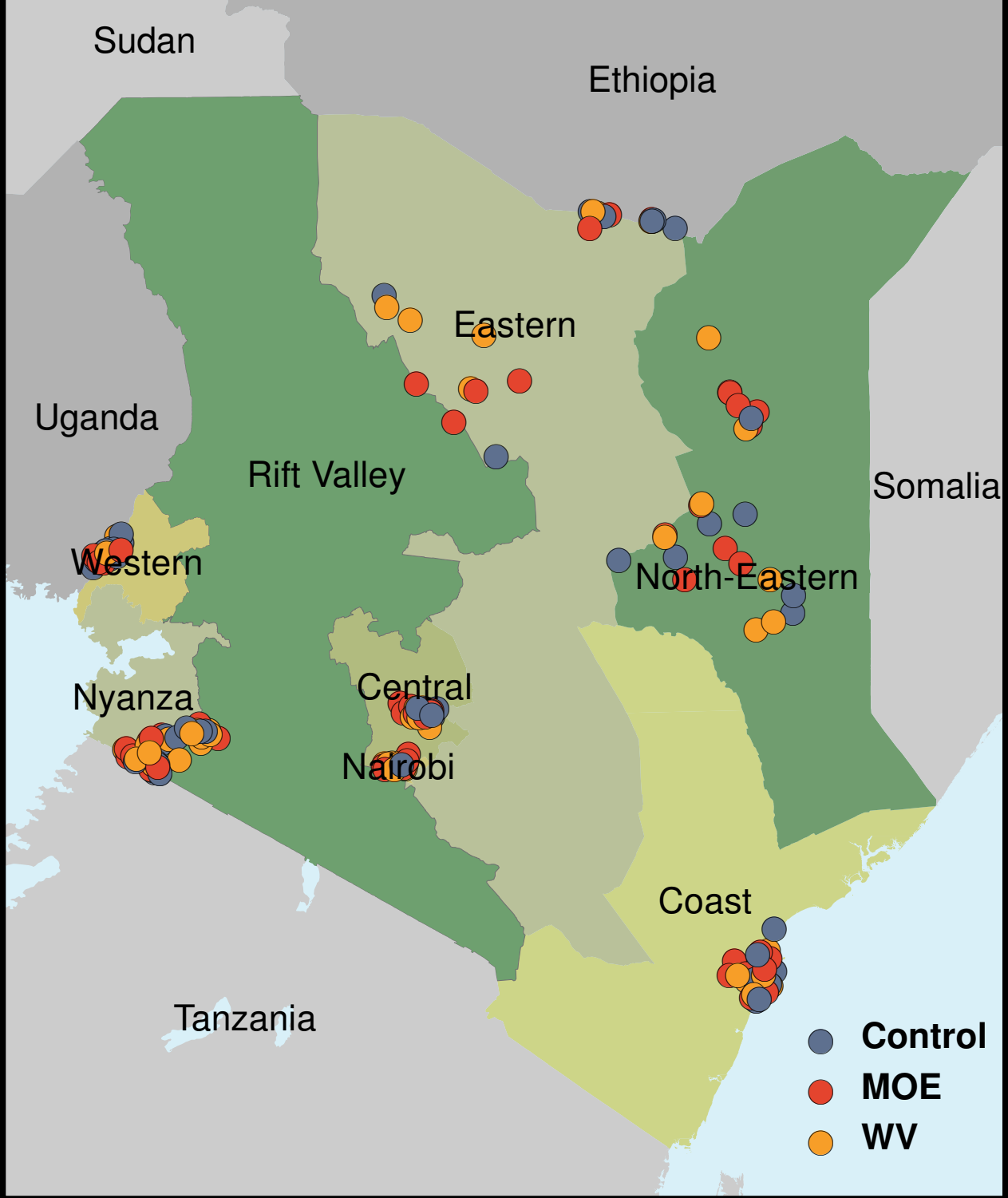
What have we learned?
- Contract teachers
- Class size
- Private schools

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

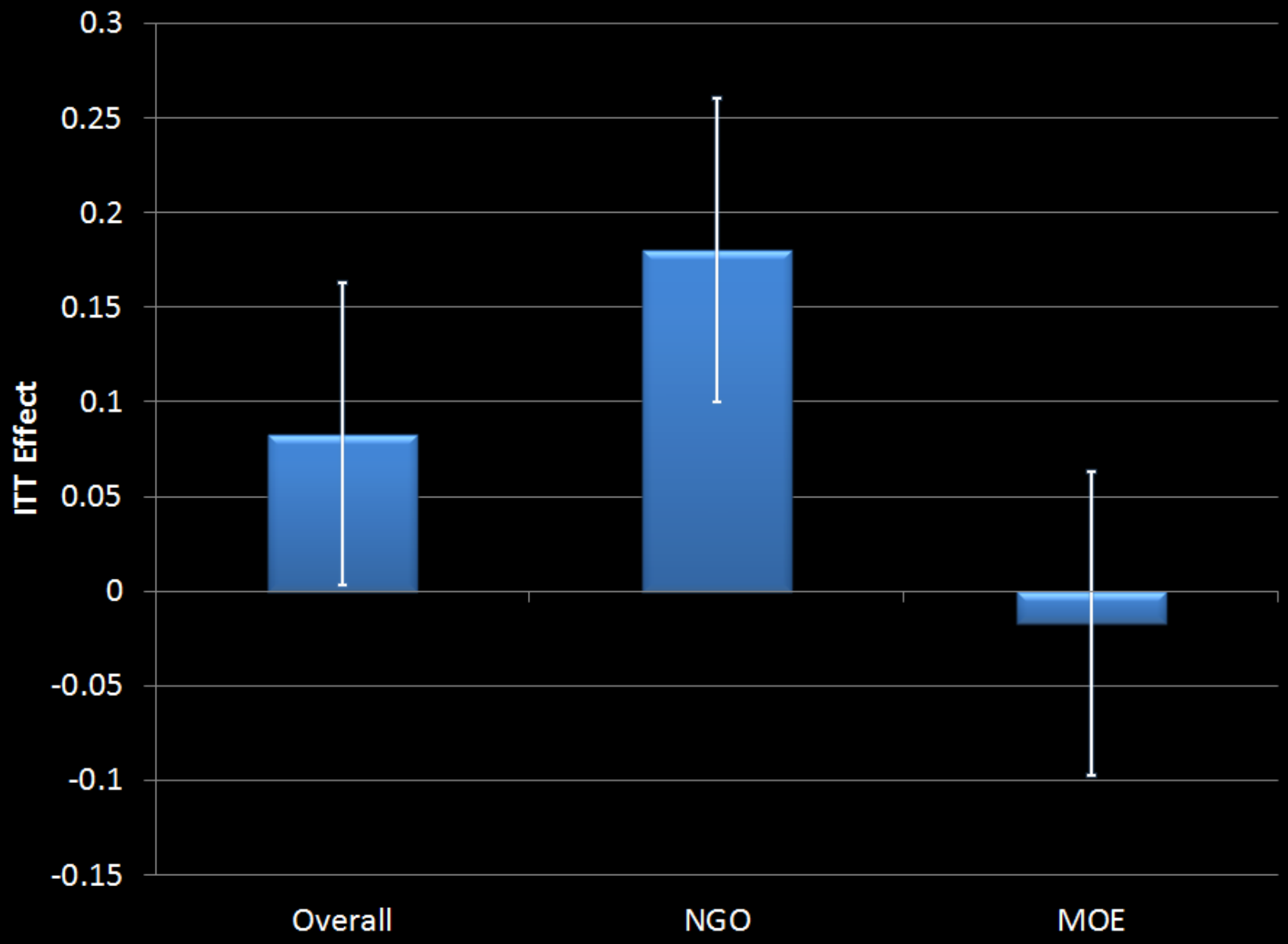
Contract teachers

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

- Duflo, Dupas, & Kremer (2009)
  Western Kenya
  Contract teachers $\Rightarrow$ $+0.21$ std. dev.
  Class size reduction $\Rightarrow$ no effect on scores


Sudan
Ethiopia
Eastern
Uganda
Rift Valley
Somalia
Western
North-Eastern
Nyanza
Central
Nairobi
Coast
Tanzania
Control
MOE
WV

Treatment Effect of Contract Teachers on Test Scores



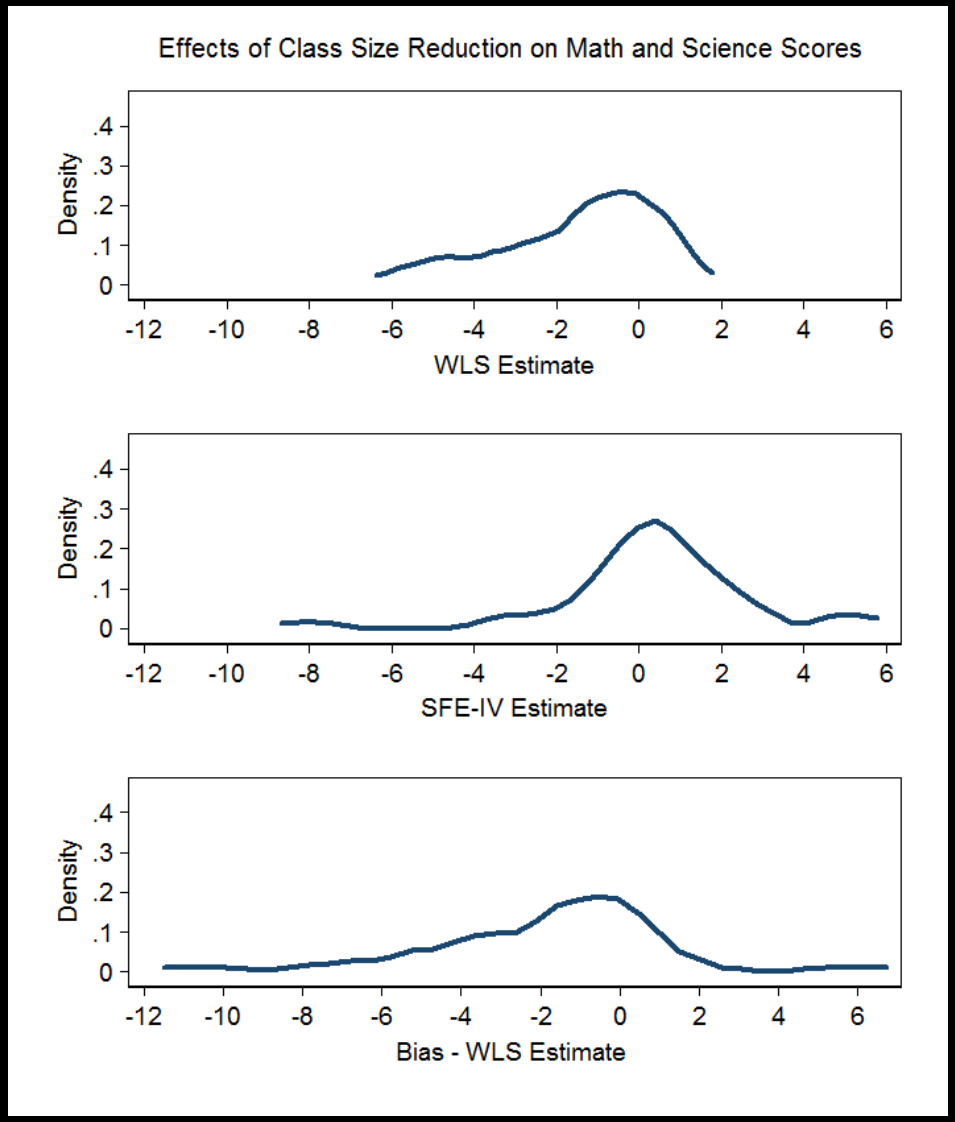
Outline

- 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

		Point estimate & std. error		
	Country	No Controls	Controlling for observables	Controlling for unobservables
RDD				
Angrist & Lavy (1999)	Israel	0.322 (.039)	0.019 (.044)	-0.261 (.113)***
Urquiola (2006)	Bolivia	0.07 (0.03)**	0.01 -0.03	-0.21 (0.07)**
Asadulla (2005)	Bangladesh		0.25 (0.115)***	3.5 (1.03)***
RCT				
Krueger (1999)	USA			-0.271 (.072)***
Banerjee et al (2007)	India	0.027 (.0125)**		0.064 (.118)
Duflo et al (2012)	Kenya			-0.064 (.024)**

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

		Point estimate & std. error		
	Country	No Controls	Controlling for observables	Controlling for unobservables
None				
Cox & Jimenez (1991)	Colombia	0.22	0.55	
Cox & Jimenez (1991)	Tanzania	-0.14	0.97	
Aggregation				
Hsieh & Urquiola (2006)	Chile	-0.714 (1.188)		-0.51 (1.390)
Tabarrok (2013)	India			0.224 (0.036)***
Bold et al (2012)	Kenya	0.79 (0.046)***		0.98 (0.41)**
RCT				
Angrist et al (2002)	Colombia		0.379 (0.111)***	0.291 (0.153)**

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

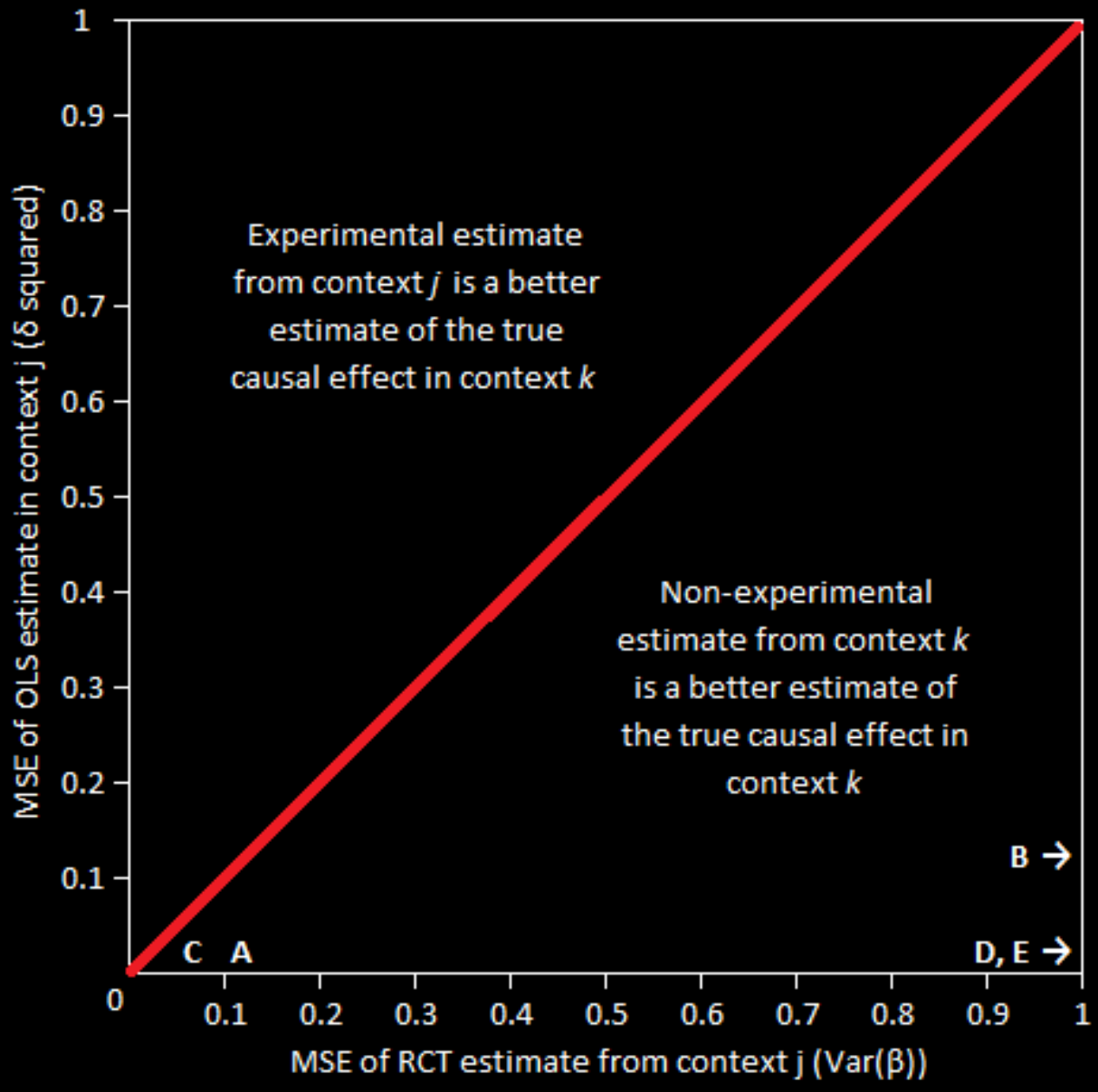
OLS ‘here’ vs an RCT ‘there’

For non-experimental estimates:

$$\text{MSE}(\tilde{\beta}_k) = \underbrace{\text{Var}(\tilde{\beta}_k)}_{\substack{\text{Sampling} \\ \text{error}}} + \underbrace{(\tilde{\beta}_k - \beta_k)^2}_{\substack{\text{Omitted} \\ \text{var. bias}}}$$

For experimental estimates:

$$\text{MSE}(\hat{\beta}_j) = \underbrace{\text{Var}(\hat{\beta}_j)}_{\substack{\text{Sampling error} \\ \text{in context } j}} + \underbrace{\text{Var}(\beta)}_{\substack{\text{Variance of true effect} \\ \text{across contexts}}}$$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
MSE of OLS estimate in context j (δ squared)
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
B →
C
A
D, E →
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
MSE of RCT estimate from context j (Var(β))

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

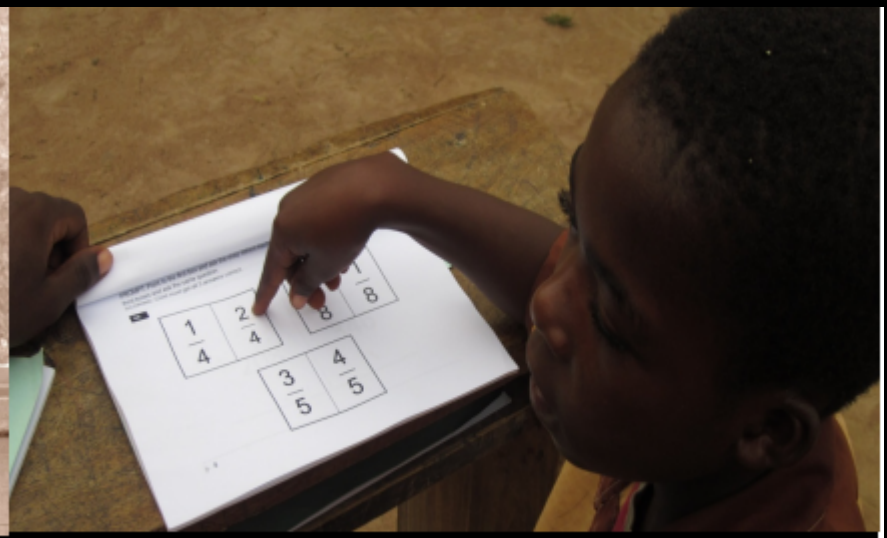
Alternative evaluation models

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 (Ludgwig, Kling, and Mullainathan) applied within organizations to production of outputs.
- Move up the bureaucratic supply chain (Rasul & Roger 2013, Nigerian Civil Service).

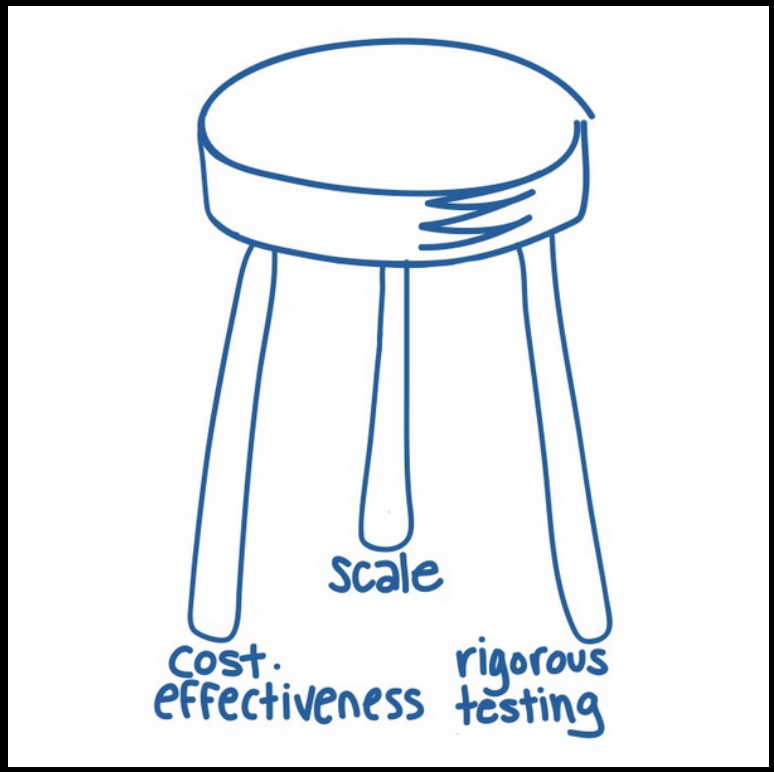


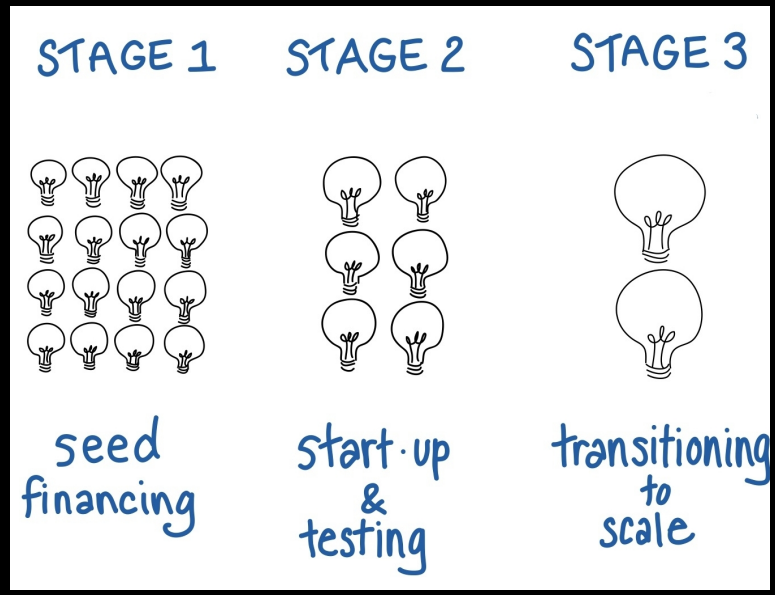
Teacher Community Assistants Initiative

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)

Development Innovation Ventures (DIV) @ USAID





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.