Returns on Scholarship (versus Organizational Learning) in Development

Using (mostly) Education as an Example

Lant Pritchett (based on work with Joan Gass)

April 24, 2017

University of Washington

Big Picture Motivation: How did the grand and glorious field of development reach this absurdist state of affairs in which a serious scholar with tenure at a great university could suggest with a straight face (and asserting that many of the field's luminaries agreed) that an **RCT of cash versus chickens was the highest ROI research**?

.

It would be straightforward to run a study with a few thousand people in six countries, and eight or 12 variations, to understand which combination [of giving people chickens versus cash] works best, where, and with whom. To me that answer is the best investment we could make to fight world poverty. The scholars at ***Innovations for Poverty Action*** *who ran the livestock trial in Science agree with me. In fact, we've been trying, together, to get just such a comparative study started.*

Chris Blattman, Professor Harris School, University of Chicago, VOX

Agenda

- Some motivation about learning in developing countries
- The ROI Criteria
- Tentative Implications for Practitioners & Researchers (if we get to it)

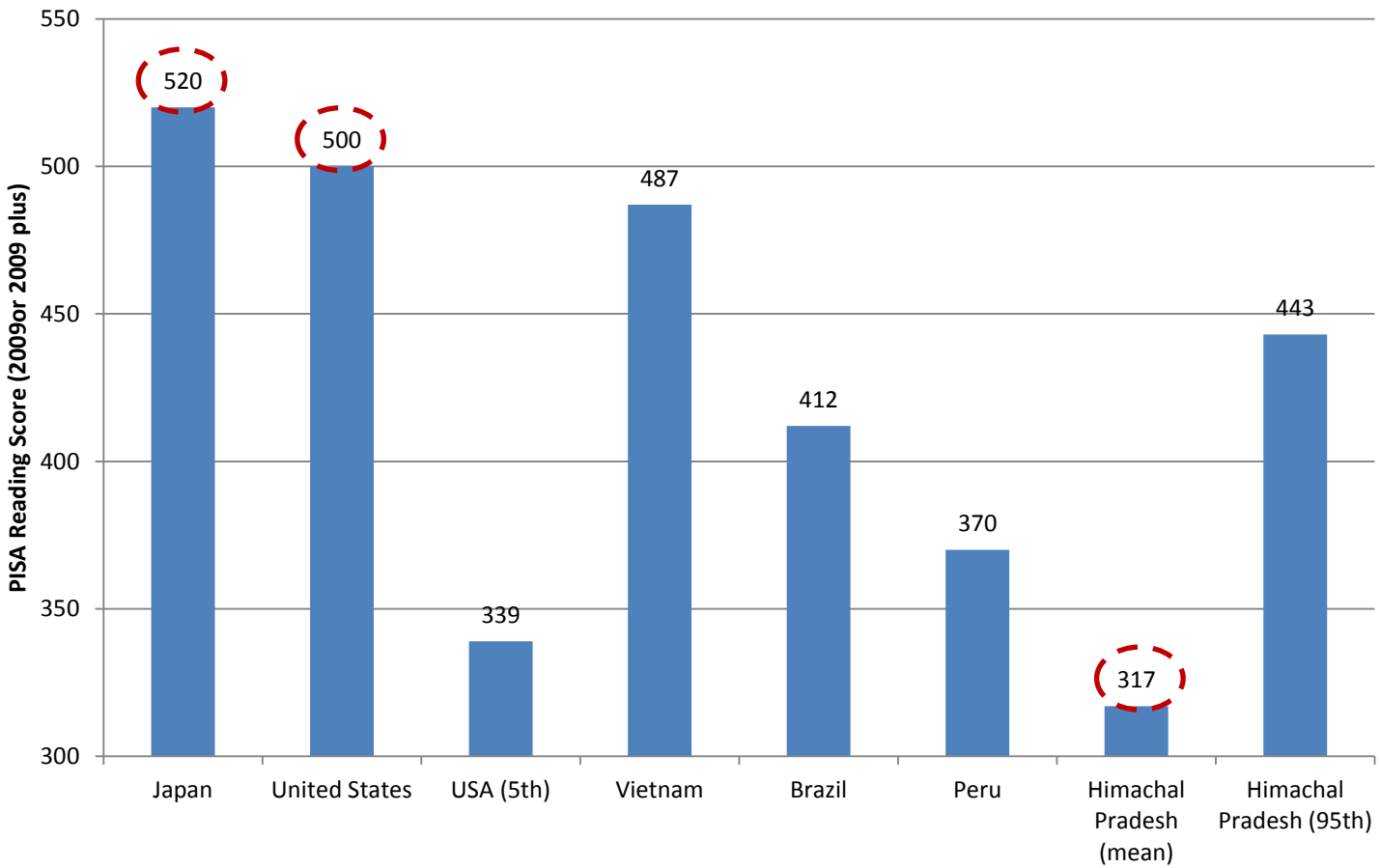
Motivation about education in developing countries

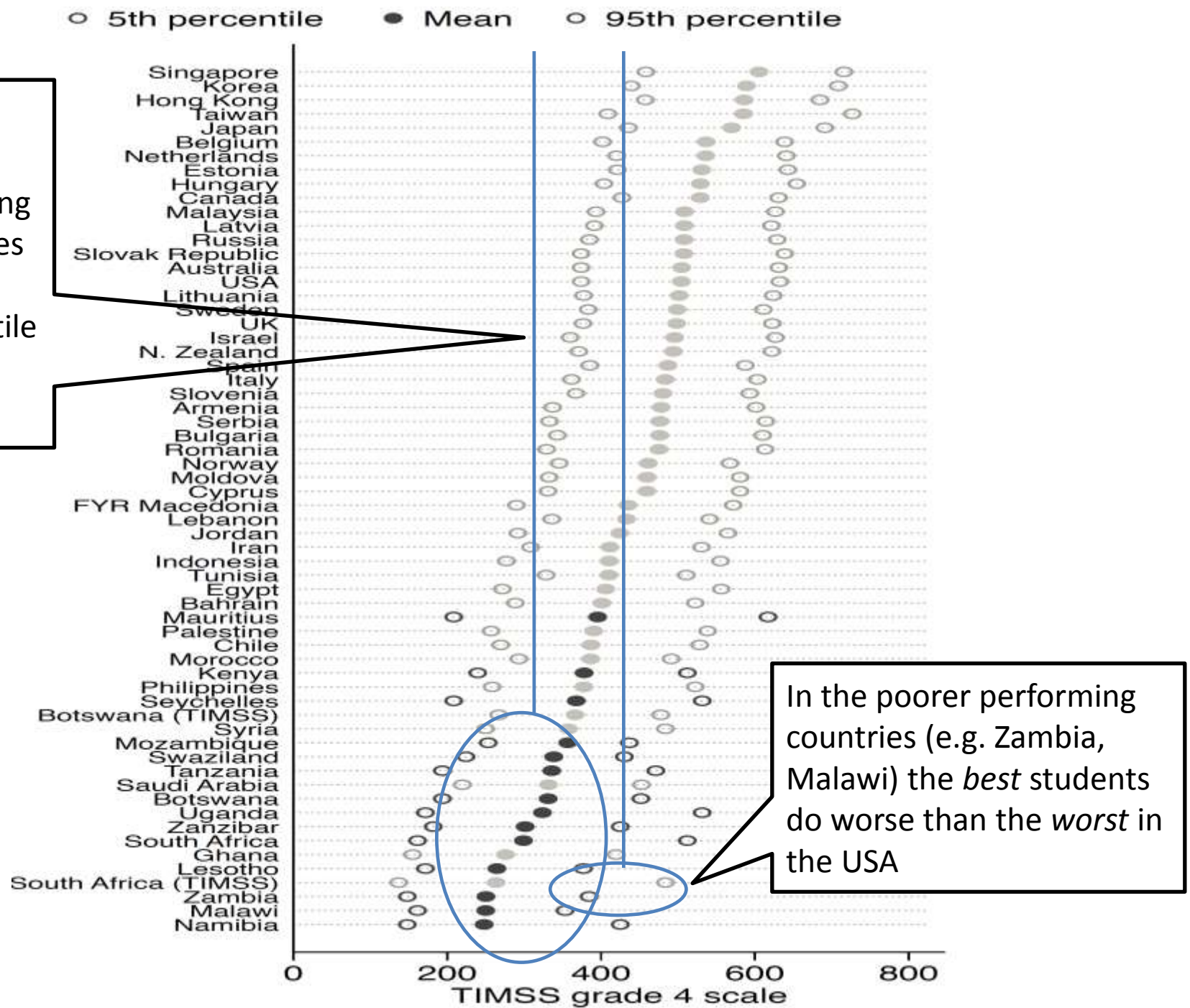
- In most low/lower-middle income countries students are one to two student standard deviations below their OECD counter-parts on standardized assessments (e.g. PISA, TIMSS)

- *College graduates* in Jakarta score lower the *high school drop-outs* in Denmark

- Roughly half of developing country young adult women who completed grade 6 *cannot read a single sentence* in their preferred language

India's state of Himachal Pradesh (a better performing state) is 9 times further behind USA than USA than Japan—the mean is lower than 5th percentile



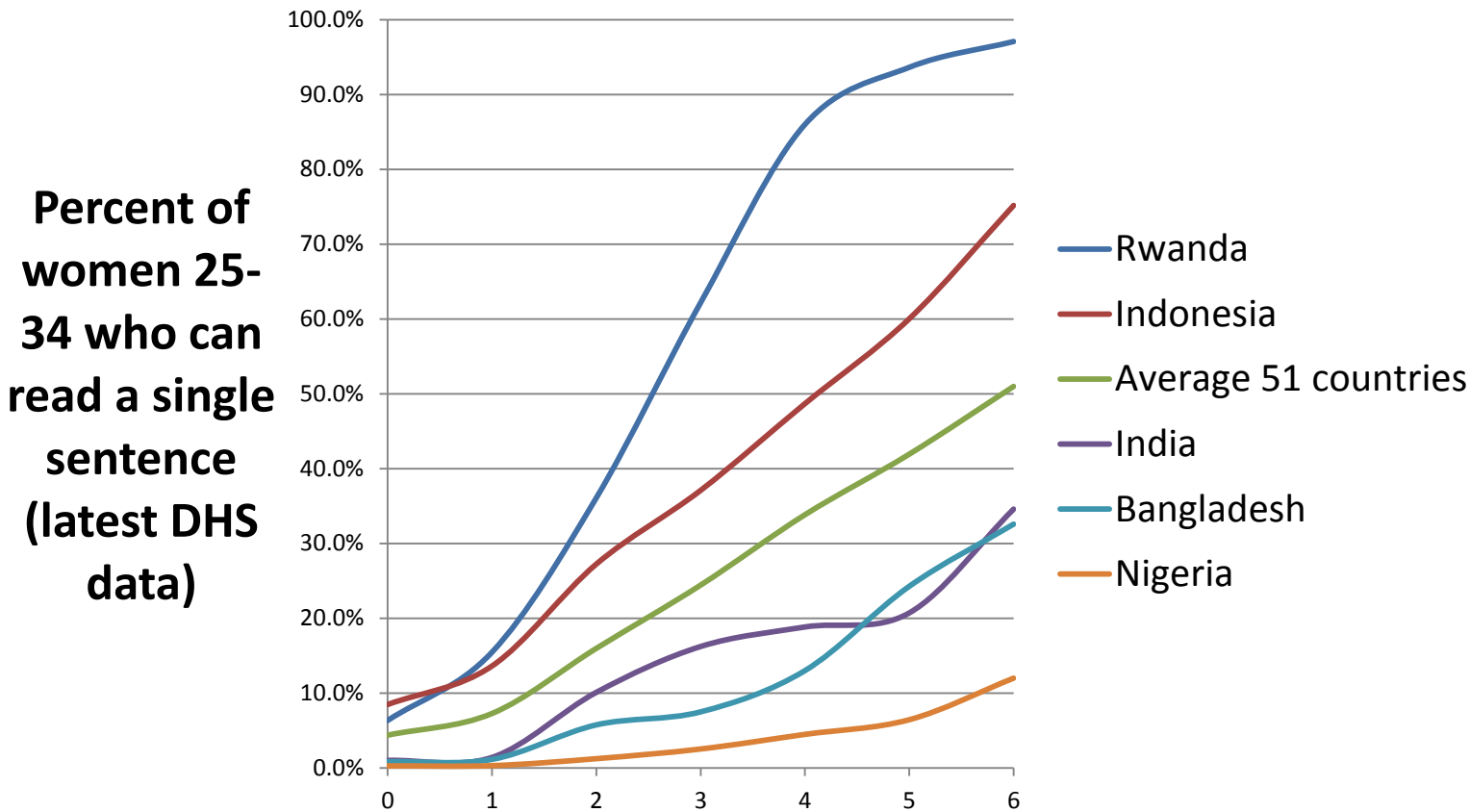


The *average* student in the lower performing African countries does less than the 5th percentile in *any* OECD country.

In the poorer performing countries (e.g. Zambia, Malawi) the *best* students do worse than the *worst* in the USA

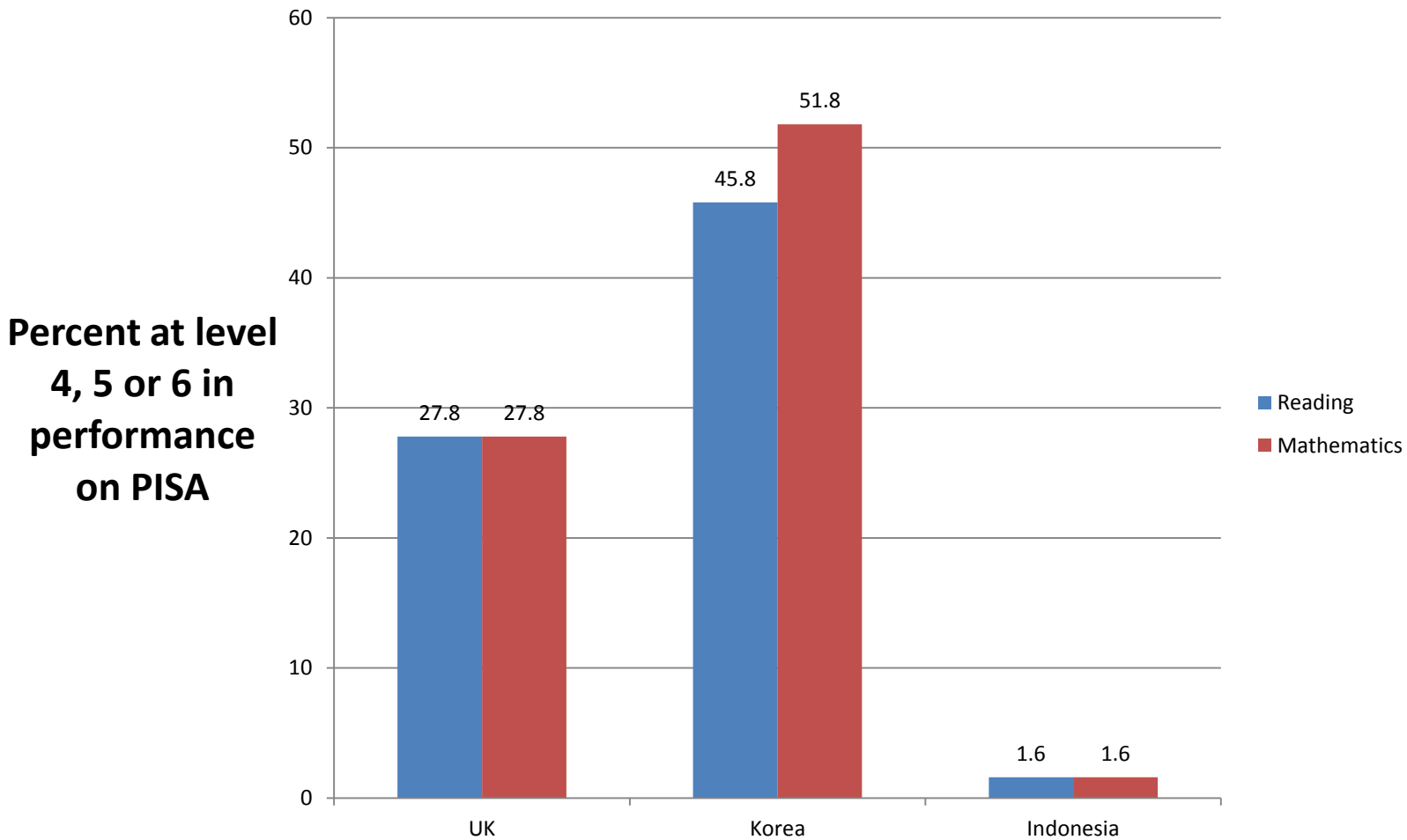
Source: Sandefur 2016, Crouch overlay

Half of adult women who completed grade 6 (but no higher) could not read a single sentence—12 percent in Nigeria

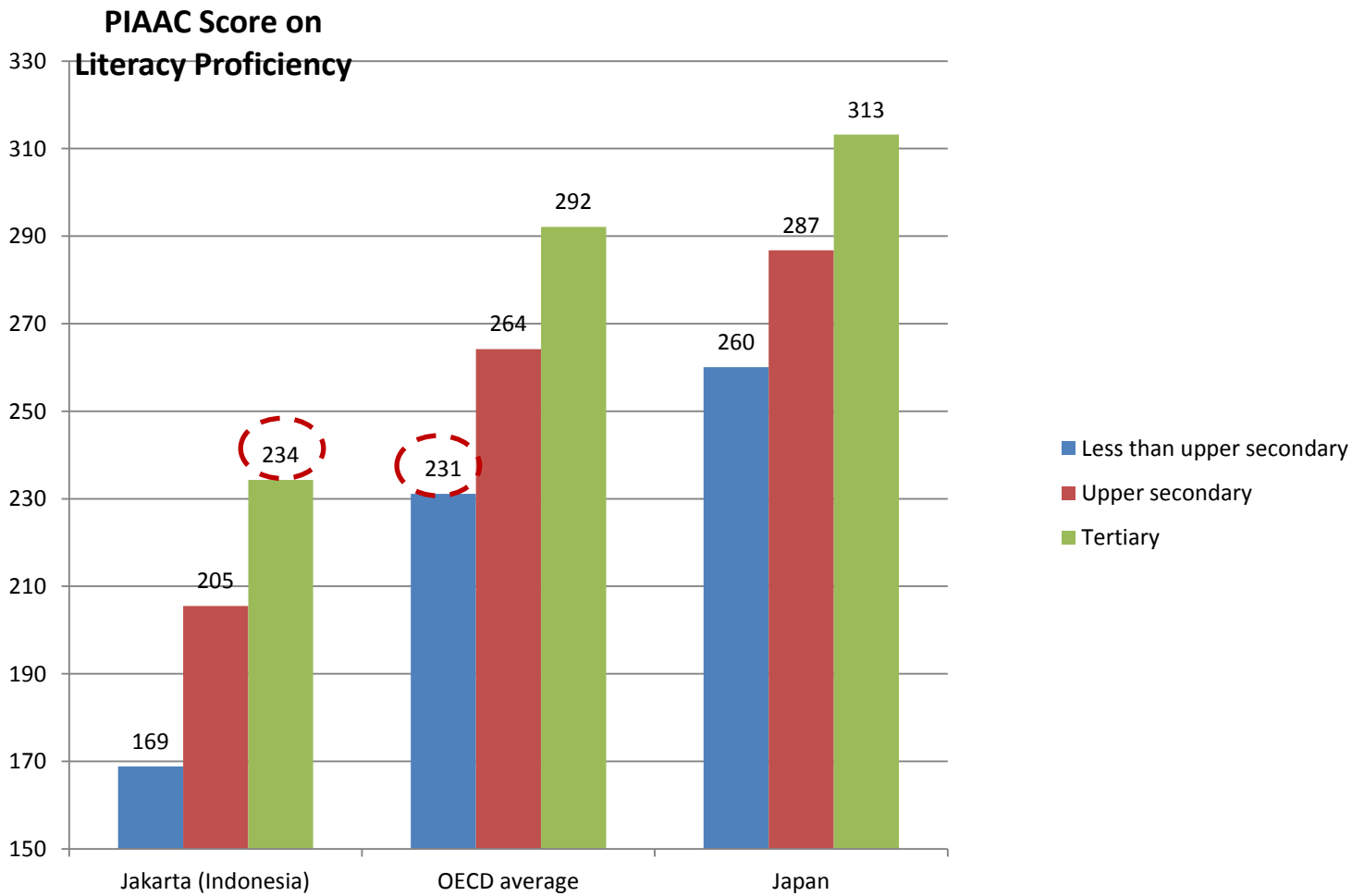


Source: Pritchett and Sandefur 2017

This isn't just that "the poor" are getting a crappy education, "the elite" are too…

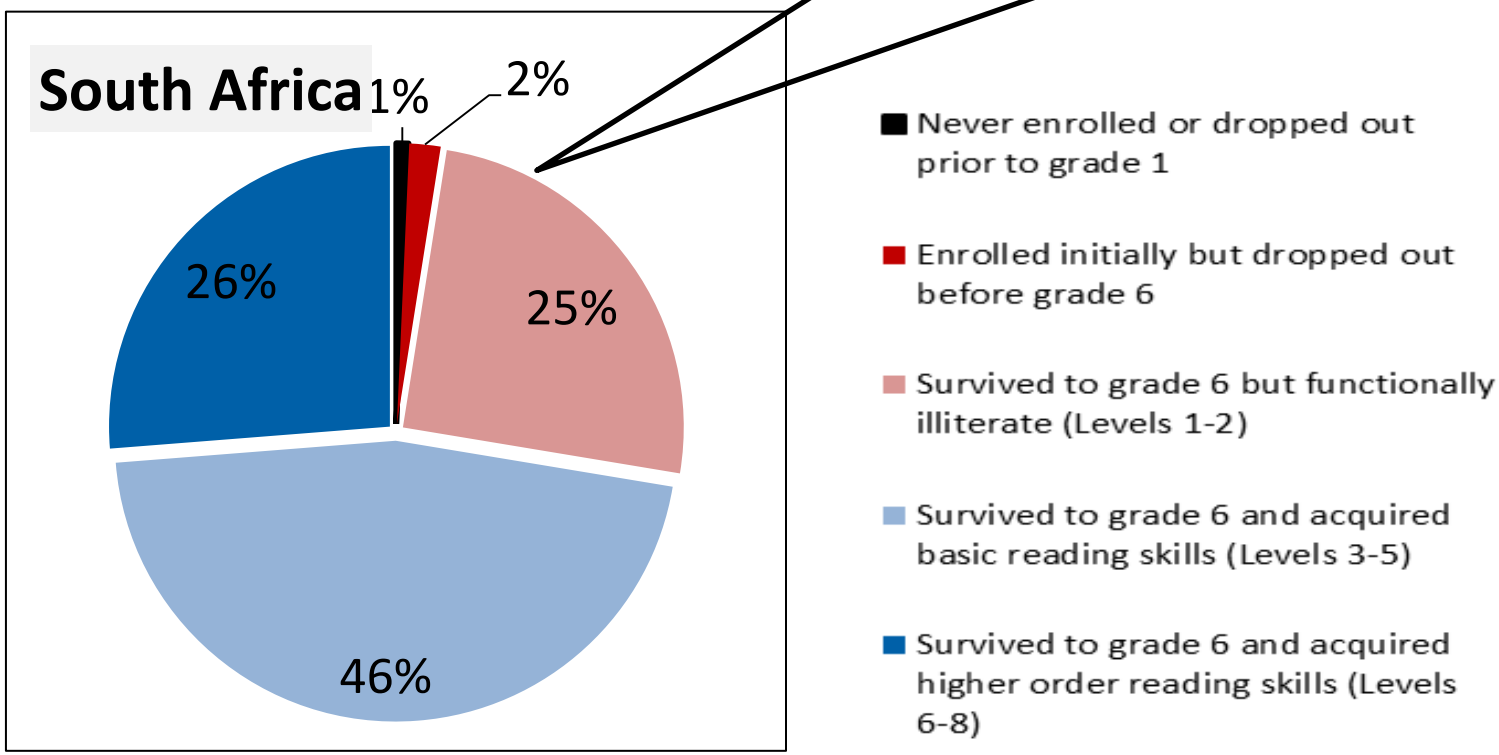


Jakartans with *tertiary* complete scored about the same as those without high school in OECD



Now the place to find an uneducated child is in school

In South Africa, of the 28 percent of children are Illiterate at age 12, 25 percentage points (90 percent) reached sixth grade so only 10 percent (3/28) of illiteracy are the unschooled



Source: Spaull and Taylor

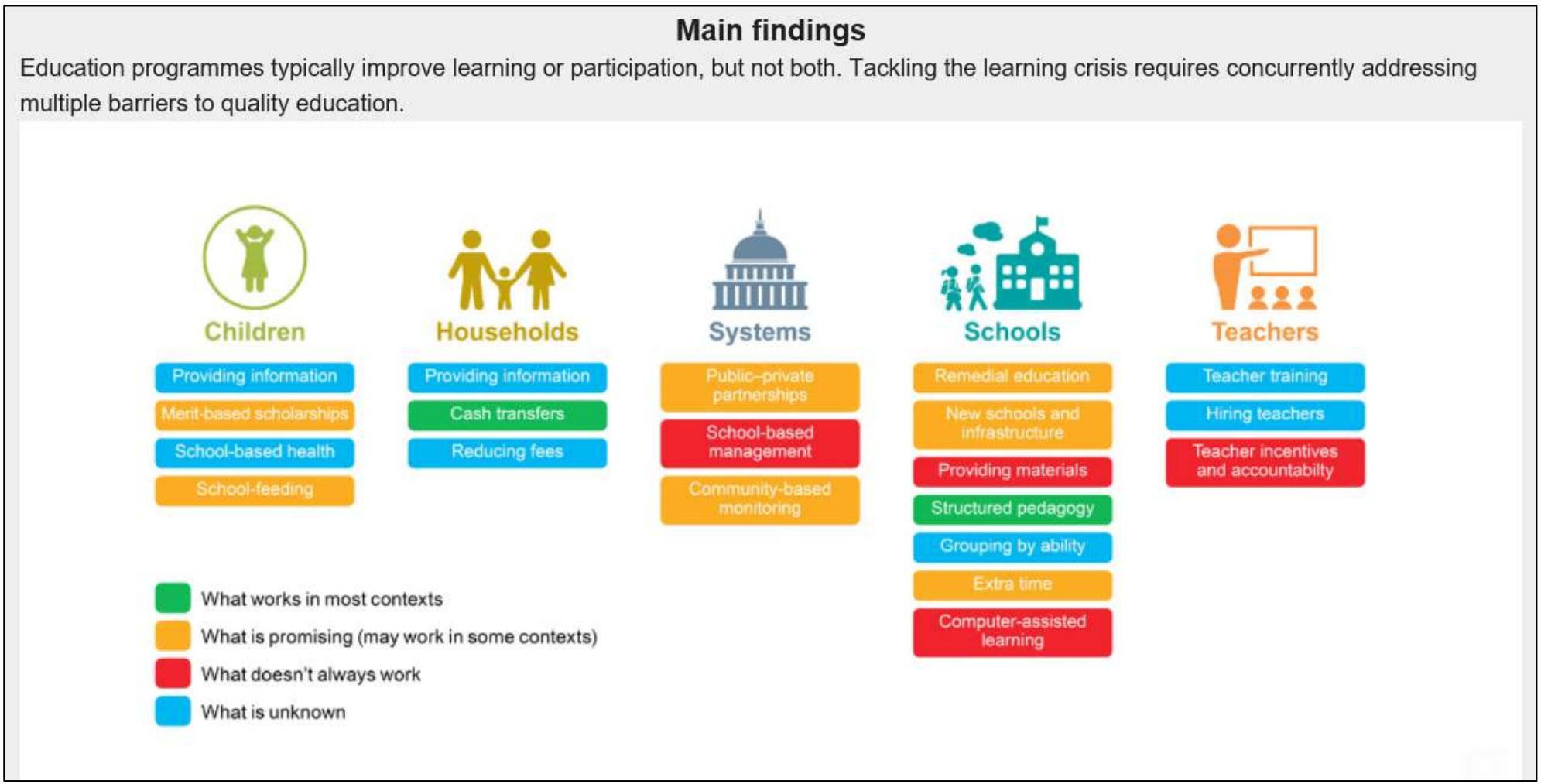
What kind of research should be going on in education when…?

- The place to find an uneducated child in in school?
- There is wide variation across countries in achieving the basics in primary school (literacy conditional on achievement varies from zero to 1)
- The average tested middle school/15 year old in school is one to two standard deviations below OECD—and Vietnam
- The (statistical) “elite” are getting a globally mediocre (at best) education?

Outline

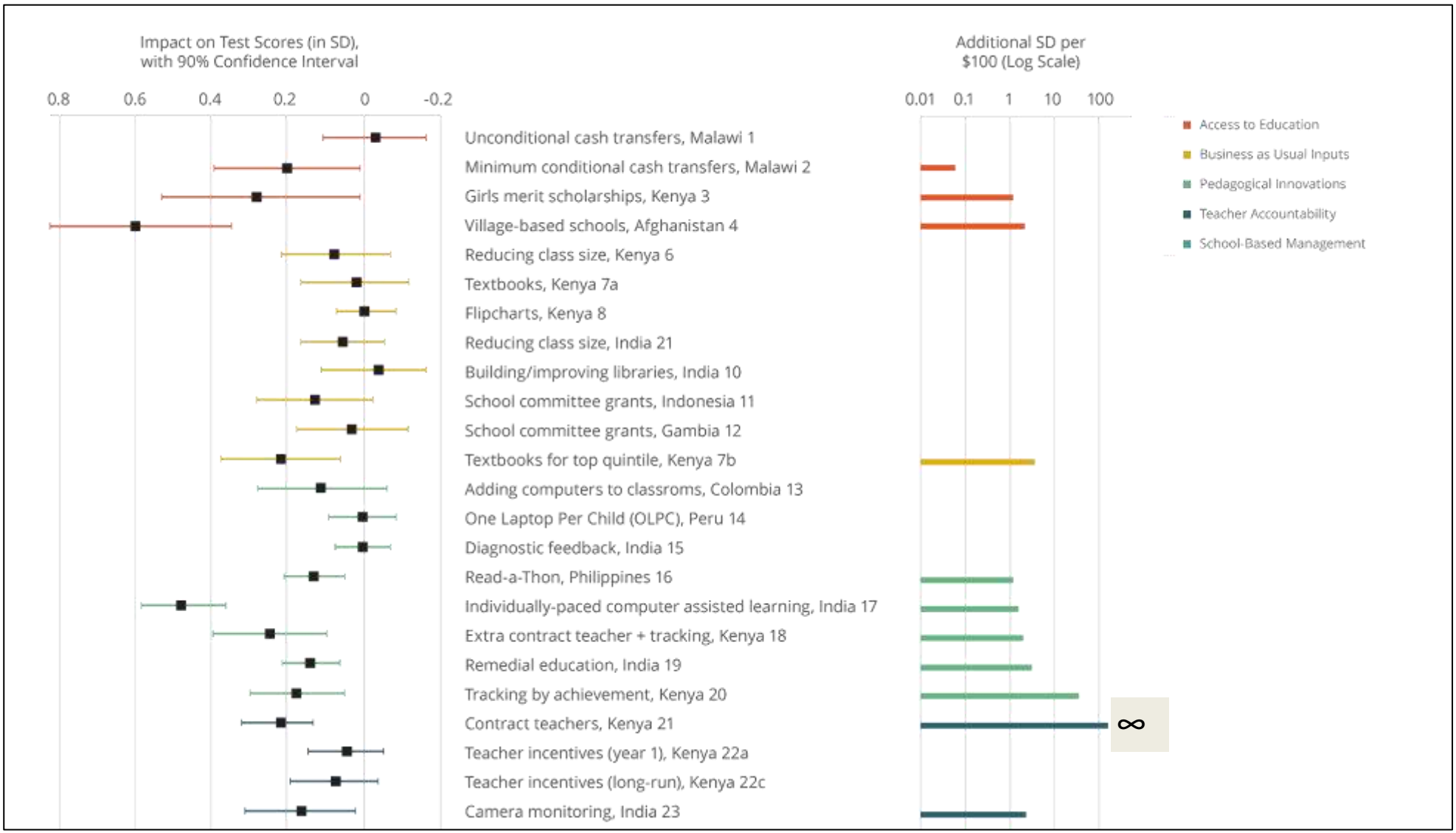
- Some motivation about learning in developing countries
- The ROI criteria for research
- Tentative Implications for Practitioners & Researchers

Currently, academic scholarship mostly focuses finding the marginal returns of specific interventions



Source: 3ie Systematic Review "The Impact of Education Programs on Learning and School Participation in Low and Middle Income Countries" September 2016.

“Cutting-edge” scholarship attempts to (re)introduce cost effectiveness with RCT estimates of impact



Note: Graph replication of original source; ideally results were be displayed based on the cost effectiveness of the bounds of the confidence interval.

Source: JPAL Education: Increasing Test Score Performance Cost Effectiveness Analysis

The previous graph should already alert you that something is *deeply* wrong

- Suppose we adopted the standard *positive* model of producer behavior that the producer maximizes subject to constraints. The two implications are:
  - Marginal product per dollar is equalized across all inputs
  - These are all equal to "lamda" or the shadow price of the relaxation of the constraint
- What are we to make of a table, by economists, that shows that many inputs (e.g. class size) have zero marginal product and some have *infinite* (!?) marginal product per dollar and this is unremarked on as a feature)?
- What are we to make of a field that knew *twenty years ago* (Filmer and Pritchett 1997) that marginal product per dollar in producing learning was not equalized across inputs *by orders of magnitude* and then spent the next twenty years doing expensive scholarship producing more estimates of the marginal product per dollar of various inputs on the self-negating premise this was "policy relevant"

The *return on investment* (ROI) in research should be judged based on anticipated net present value of the change in outcomes induced

	Potential Research Investment	Research ROI	Explanation
Option 1	• **$2M in 2006** running an RCT on an **NGO-run contract teachers program in Western Kenya (**Duflo et al 2012**)**	• Limited	• It was already known there were high marginal returns to contract teachers from *dozens* of experiences (Murgai and Pritchett 2006) but also already **known but scalability was limited** as *every single one was reversed.* The attempt to "scale up" the "rigorous evidence" about the infinitely cost effective "intervention" of contract teachers across Kenya failed to produce impact (Bold et al 2013)—exactly as expected.
Option 2	• **$2M** researching **successful political strategies** to adapt teacher **performance management** as part of teacher pay increases	• High	• This is a risky area of research as no clean simple RCT is possible, but **could save governments billions of successful**. - For example, in Indonesia teacher's pay was doubled in an effort to improve teacher performance. No learning improvements occurred and the government spent billions (Ree et al. 2015)

Research more commonly occurs in Option 1, but expected value (ROI) is higher for Option 2

Note: ROI is Return on Investment

There are seven criteria for NPV / ROI

	Seven Criteria for NPV/ROI	Definition
Potential Impact	1 Marginal Return per Dollar	• The expected increase in learning outcomes per dollar at LATE (Local Average Treatment Effect)
	2 Scope	• The likelihood that the local average treatment effect hold constant or diminish rapidly with the intensity of the intervention (e.g. how 'local is the LATE'?)
	3 Duration	• The expected timeframe that the effects will last for
Ability to Replicate	4 External Validity	• The likelihood of seeing the same results across different contexts (e.g. geography, population, time)
	5 Design Fragility	• The fragility/robustness of the program design to replication (e.g. How sensitive are the outcomes to variations in program design?)
Ability to Scale	6 • Political Support	• The likelihood of govt adoption, support, or opposition; ability to access a sustainable flow of public resources, including ability / willingness to cover costs
	7 • Organizational Delivery Capability & Match	• The likelihood that the implementing organization has the capability to deliver the program. Based on an assessment of consistency of the program delivery with the deep structure and organizational mission. Also includes funding or revenue model

Note: *Or another outcome measure; LATE is the Local Average Treatment Effect

1

Marginal Return Per Dollar (LATE)

EXAMPLE: MARGINAL RETURN

Ratio of test score gain per dollar in Portuguese and Mathematics for various inputs relative to teacher salary (=1), average estimates from Northeast Brazil		
		Average (across 2nd and 4th grades and subjects)
Material inputs		
	Textbook usage	**17.7**
	Writing materials	**34.9**
	Software*	**19.4**
Infrastructure inputs		
	Hardware*	**7.7**

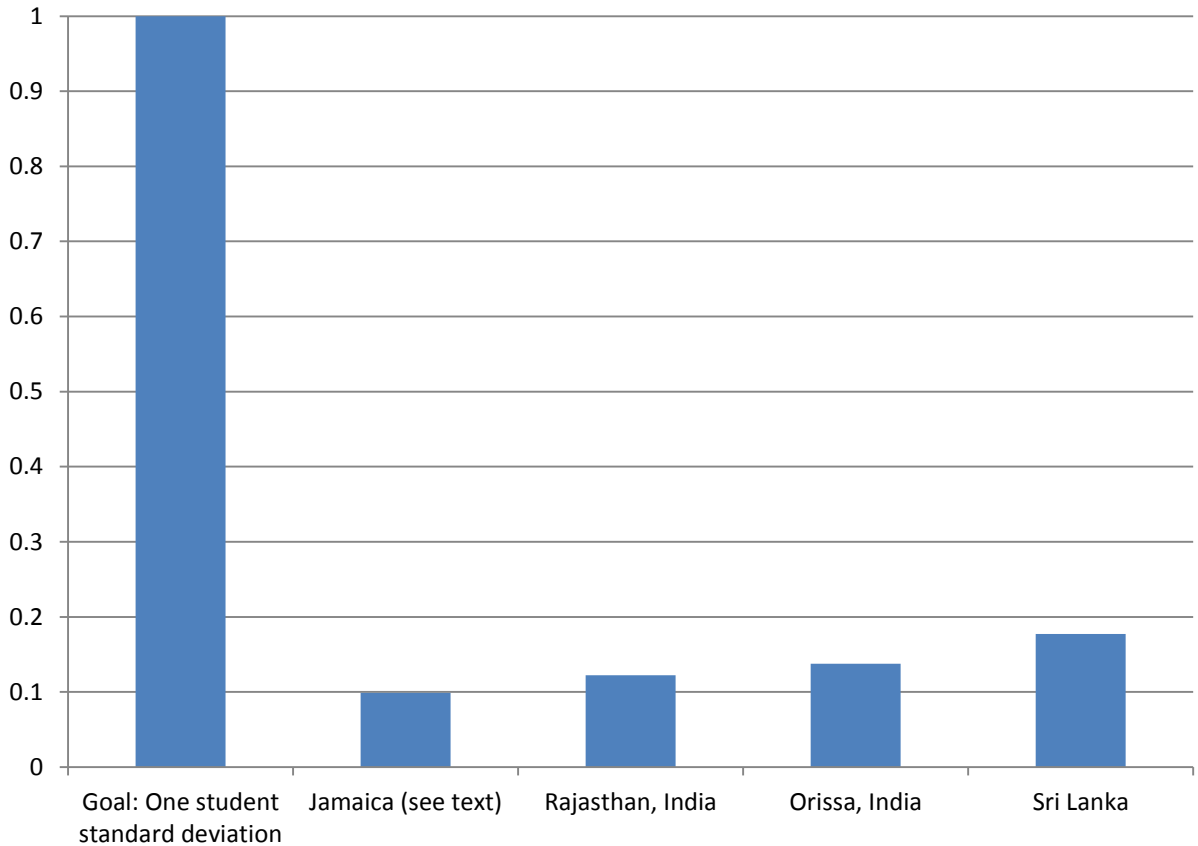
* Hardware: water, bookcase, teacher table, pupil chair, pupil desk, two classrooms, large room, director's room, kitchen, toilet, store cupboard Software: writing material, chalk, notebook, pencil, eraser, crayons, textbook usage, Source: Harbison and Haunushek 1992

RESEARCH IMPLICATIONS

- There are high returns to research of demonstrating *either* that:
  - There are high impact per dollar of an intervention—there is *too little* being spent on input/action X
  - There are low impact per dollar returns of actions on which there is (lots of) money being spent—there is *too much* being spent on input/action X

1

But we have known for donkey years that in developing country systems BAU applications of "thin" inputs have low impact on learning—with little impact on policy (more on that later)



Estimated effect size impact of "input fantasy" of increasing every input to its maximum level at estimated impact of inputs on learning

$$\Delta Scores\ input\ fantasy = \sum \beta_i * (X_i - X_{\max(\beta=0)})$$

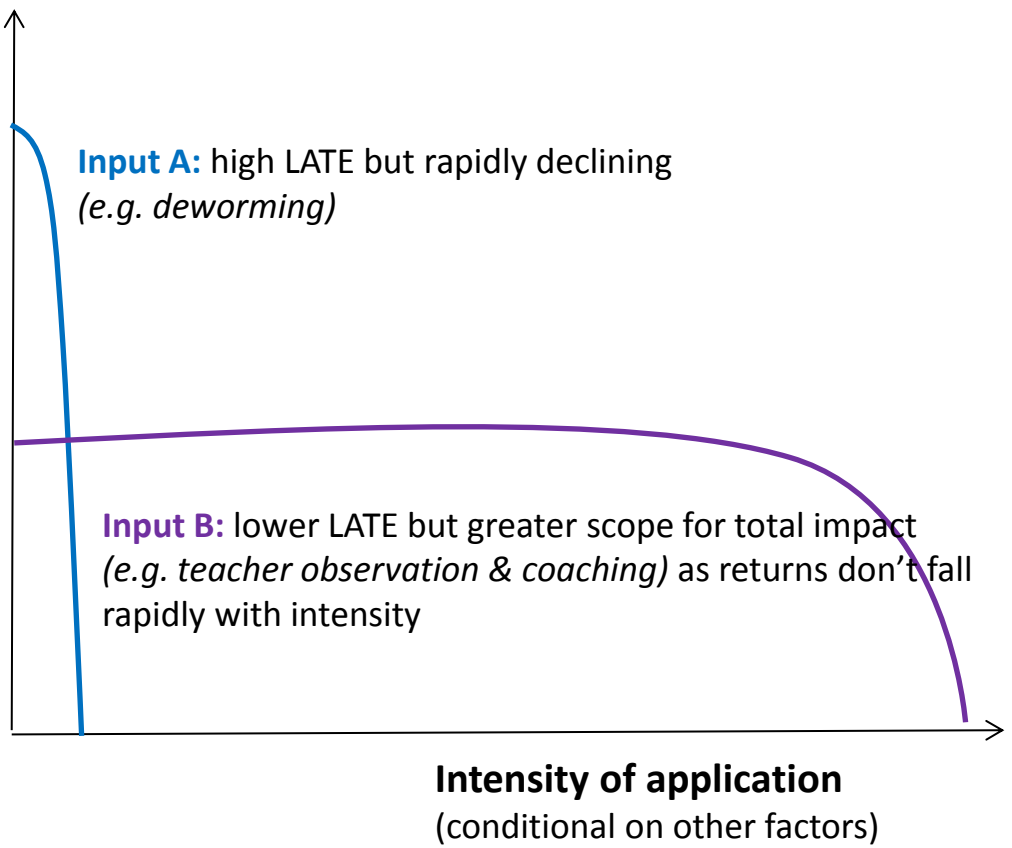
Source: Schooling Ain't Learning by Lant Pritchett

2

Scope for expansion—how "local" is the LATE?

EXAMPLE: SCOPE

Marginal product
(true LATE)



RESEARCH IMPLICATIONS

- **Baseline analysis** to understand current system in developing countries and developed countries (e.g. funding per pupil, number of books, class size, teacher attendance)
- **Scope calculations**
- **Proposed categorization** of education research based on the 'scope multiplier'
- **Examples:**
  - Glewwe and Jacoby (1994) show that **fixing leaking roofs on middle schools in Ghana** has high marginal impact per dollar—but limited scope for total gain as marginal product falls to zero once the roof doesn't leak.

Duration

- Open question about the persistence of "treatment" impacts on learning.
- Some studies suggest rapid depreciation of learning impacts so that rather than persistence, or even amplification of learning impacts, the long-run impacts are much smaller than short-run.

Program External Validity

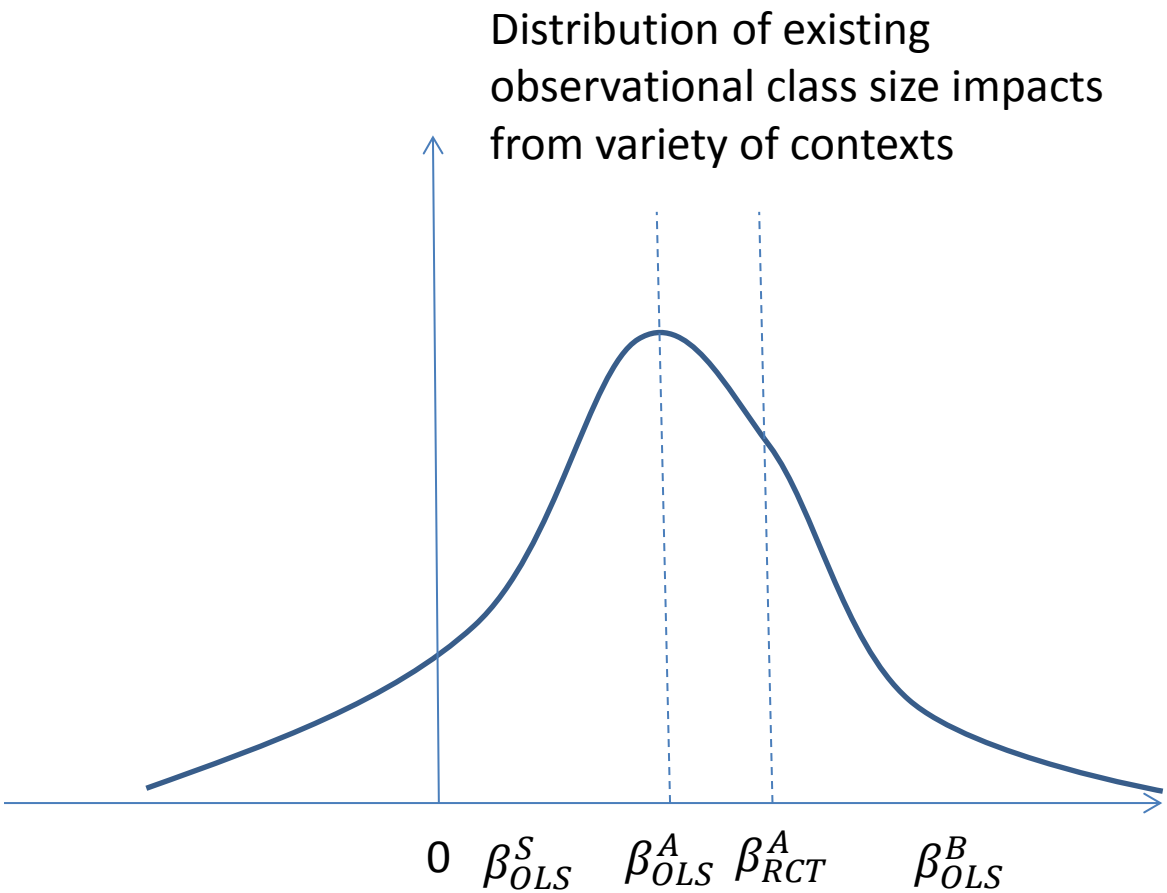
EXAMPLE TEXTBOOK PROVISION: FOUR DIFFERENT RCTS FIND NO IMPACT OF TEXTBOOK PROVISION ON (AVERAGE) STUDENT LEARNING, SO WE CONCLUDE "TEXTBOOKS DON'T WORK?" ON THE BASIS OF A SYSTEMATIC REVIEW OF THE RIGOROUS EVIDENCE?

Study	Location	Conjectured reason for lack of impact
Glewwe, Kremer, Moulin 2009	Kenya	Textbooks were too difficult for fourth graders, only top quintile benefitted
Sarbarwal et al 2014	Sierra Leone	Teachers didn't use textbooks for fear of "using them up"
Mbiti and Muralidhran 2015	Tanzania	School grants used for textbooks has no impact without (high powered) teacher incentives
Das et al 2013	India	Fungibility of parental expenditure meant only "unexpected" textbooks mattered, otherwise reduced expenditures by parents

External Validity: Three points

- There *cannot* be external validity when there is heterogeneous evidence about impact from observational data
- There empirically *shouldn't* be homogeneity of empirical results in many cases
- Fortunately, there *isn't* external validity
- And, the external validity problem is worse than that

When there is heterogeneous evidence from existing observational studies there *cannot* be external validity



Suppose you do an RCT in context A and recover a "rigorous" estimate of impact in context A.

By the assumption of heterogeneity of the observational (OLS) estimates of impact there is a context with OLS estimate smaller than A (call is S) and bigger than A (call it B).

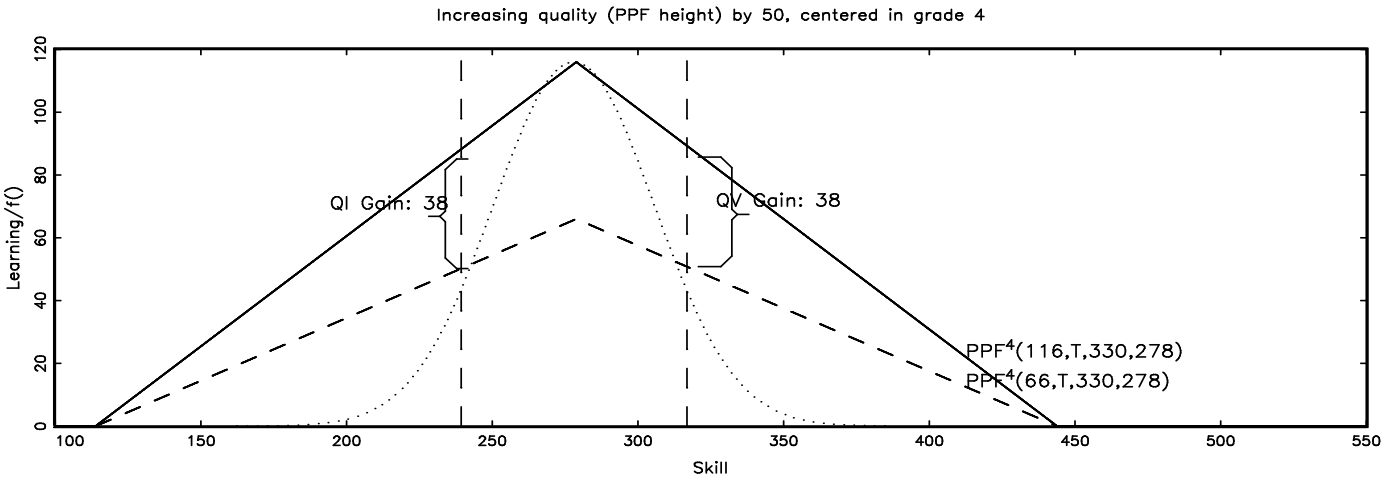
How should your expectation of the impact in context B change conditional on the findings in context A?

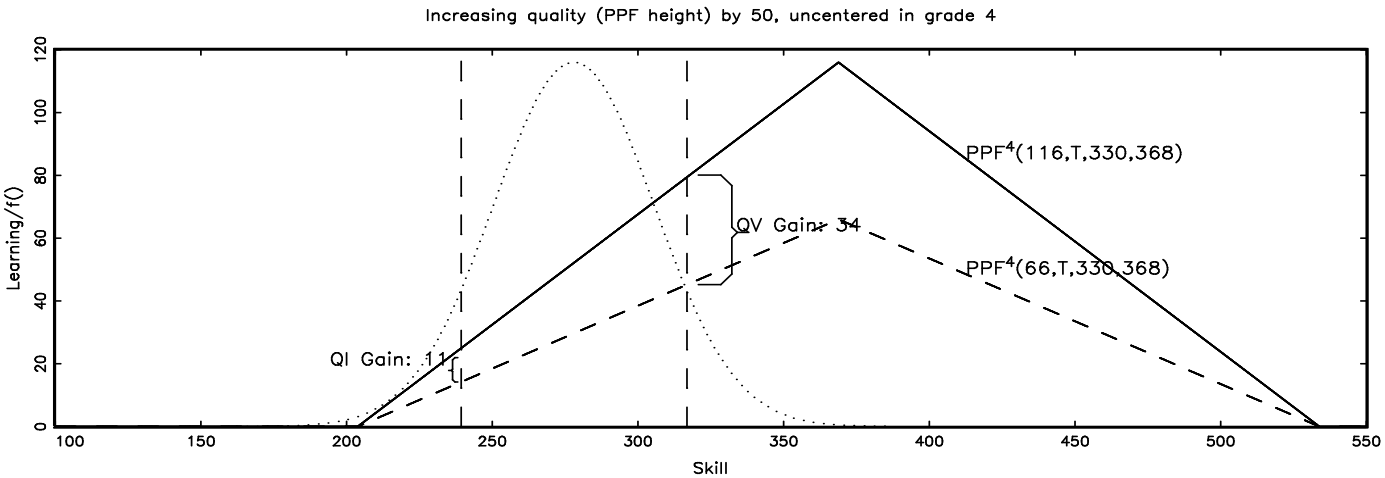
There is no coherent answer to this question as both the impact and the bias are determined by parameter sets:

$$\beta^{i}_{OLS} = \beta^{i}_{RCT}(\vartheta) + Bias^{i}(\theta)$$

Pritchett and Sandefur 2014

External validity of impact effects should not be present in a number of plausible scenarios: “uncentered” teaching





Suppose that get the most out of instruction at a “sweet spot” of existing knowledge and this declines linearly.

Then for a given PPF there is more learning if it is centered on the distribution of student skill.

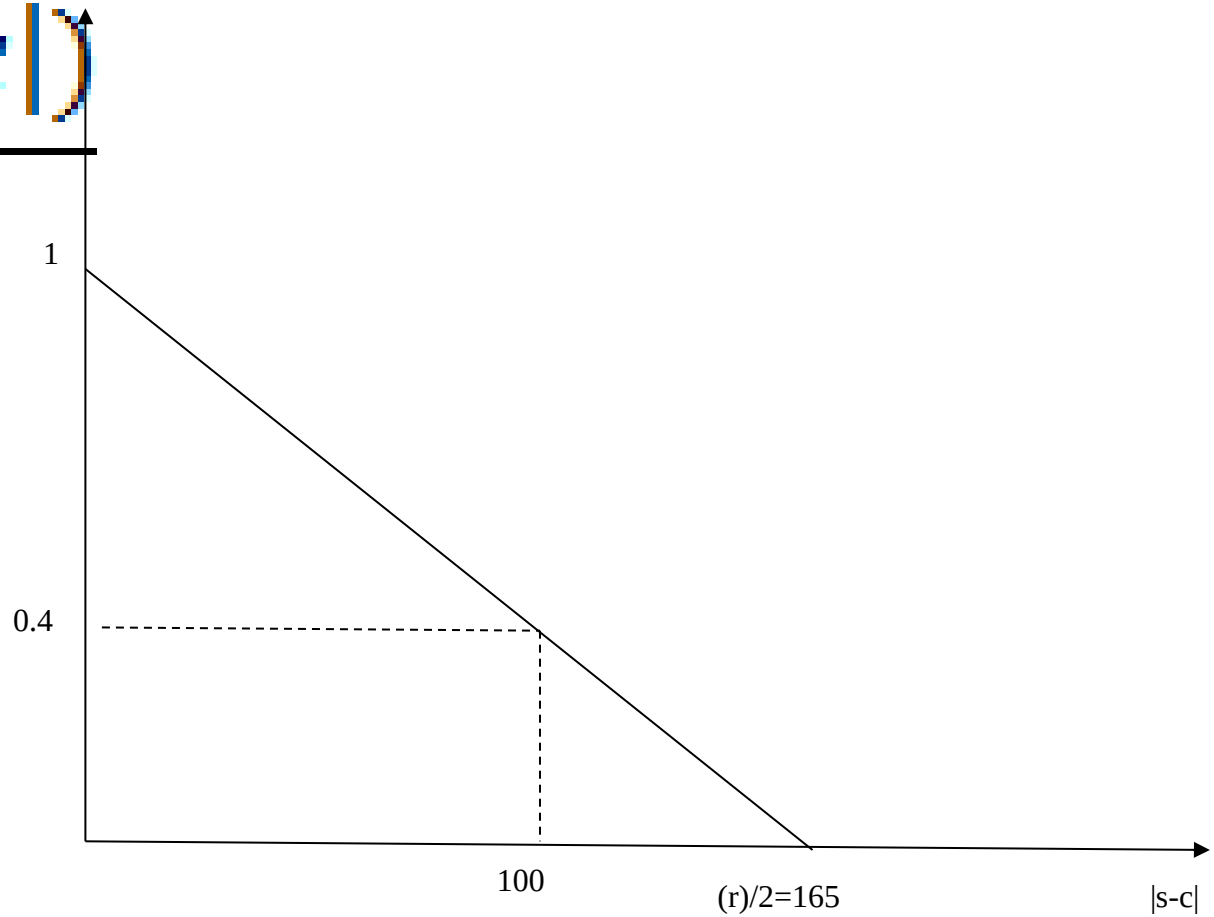
So the impact of the *exact same* upward shift in the PPF is bigger for a centered than uncentered instruction.

Beatty and Pritchett 2012

4

The learning impact *of the exact same shift in PPF* depends on the location of the PPF (Pedagogical Production Frontier) relative to center of distribution of student learning

$$\frac{dL(|s-c|)}{dh}$$

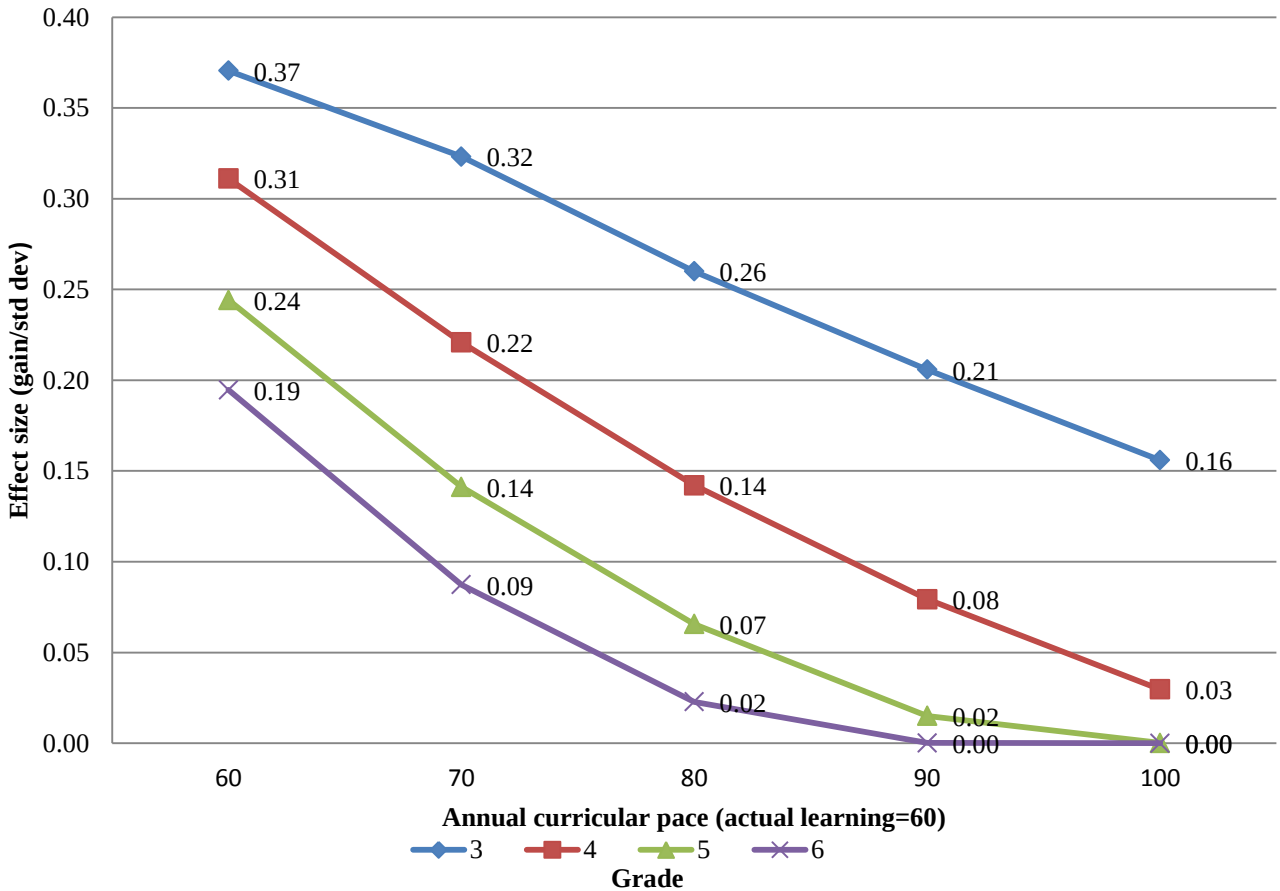


Beatty and Pritchett 2012

4

With uncentered learning any experiment uncovers a mix of increase in PPF (pedagogical production function) and curricular mismatch—the *exact same experiment* can produces estimates from .37 (huge) to 0 (nothing) effect sizes

Figure 15: The impact on student learning (in effect sizes) of increasing the PPF height by 10 in each grade varies by curricular pace

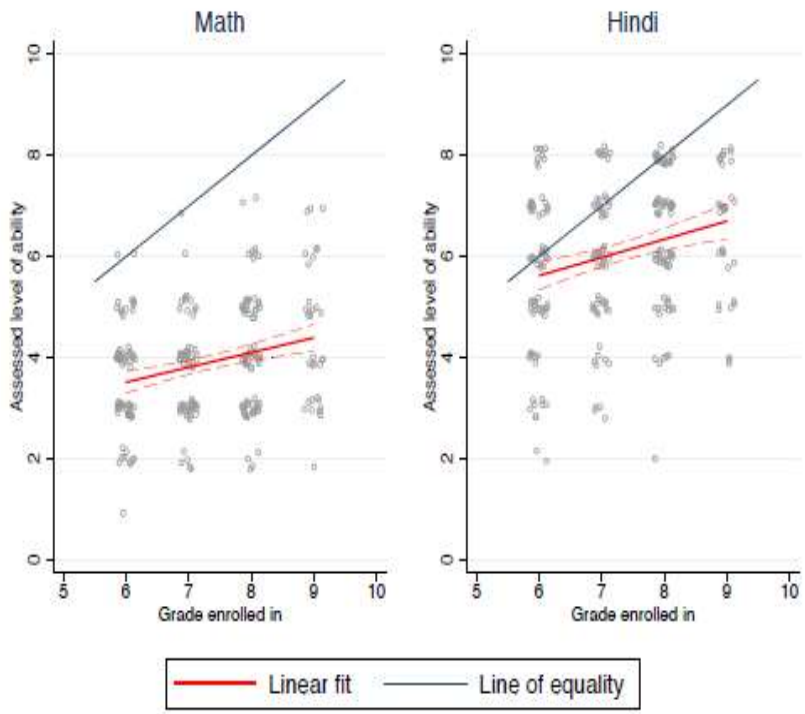


Beatty and Pritchett 2012

And, it is always good to be right….mismatch and lack of learning in upper grades in India

The problem

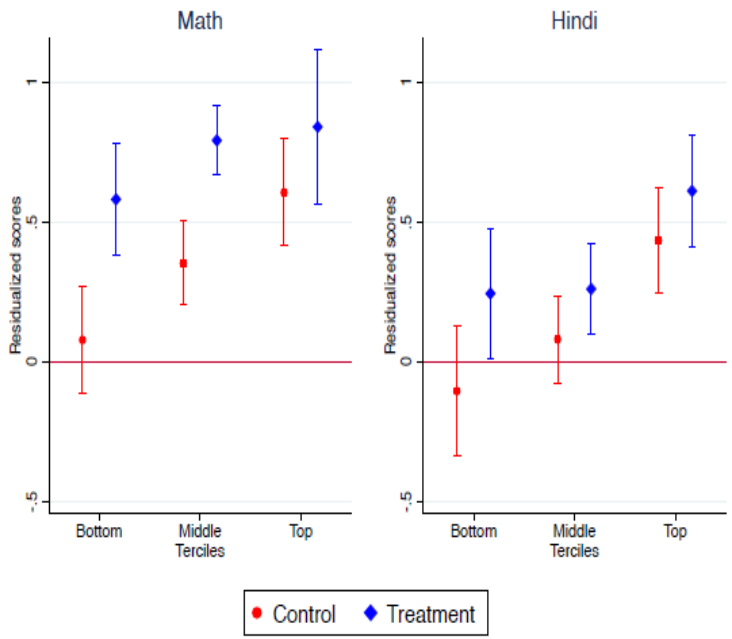
Low and dispersed achievement, mismatch with curriculum



Muralidharan, Singh and Ganimian (2016)

Results from out-of-school model

Unlike business-as-usual, the intervention taught *all* students



Control group kids in the bottom tercile learned *nothing* in Math and Hindi and individualized instruction had massive impact

4

Given that there cannot be and should not be, it is actually kind of reassuring there *isn't* external validity of program impact

- Vivalt (2016) showing the stunning variance across estimated impacts of RCTs
- Pritchett and Sandefur (2015) show in a non-education example the MSE of prediction using OLS of own context is smaller than using RCTs from other contexts.
- Evans and Popova (2015) show that the "systematic reviews" of the EPF literature come to very different conclusions.
- Even when there is "external validity" of the empirical answer (zero) there isn't to the causal explanations (see above).

And, external validity problem ROI is worse than you think….

- Most of the concern about external validity has equated “context” with “place” (like a country).
- If that is the case then the response is “just do one for each country.”
- But, without a theory or set of “invariance laws” then we don’t know what “context” is that produces empirical differences and these could change over time in the same country.
- Suppose a country had an “overambitious curriculum” problem—then every late grade PPF improvement would show small impact—but if they fixed the overambitious curriculum issue then the “true” impact of *every* PPF intervention would shift.
- Or, teacher incentives. It might be with de-motivated teachers nothing works and with motivated teachers lots of things work.
- So “just do one for each country” is a fatuous response to problems with external validity.

Design Fragility

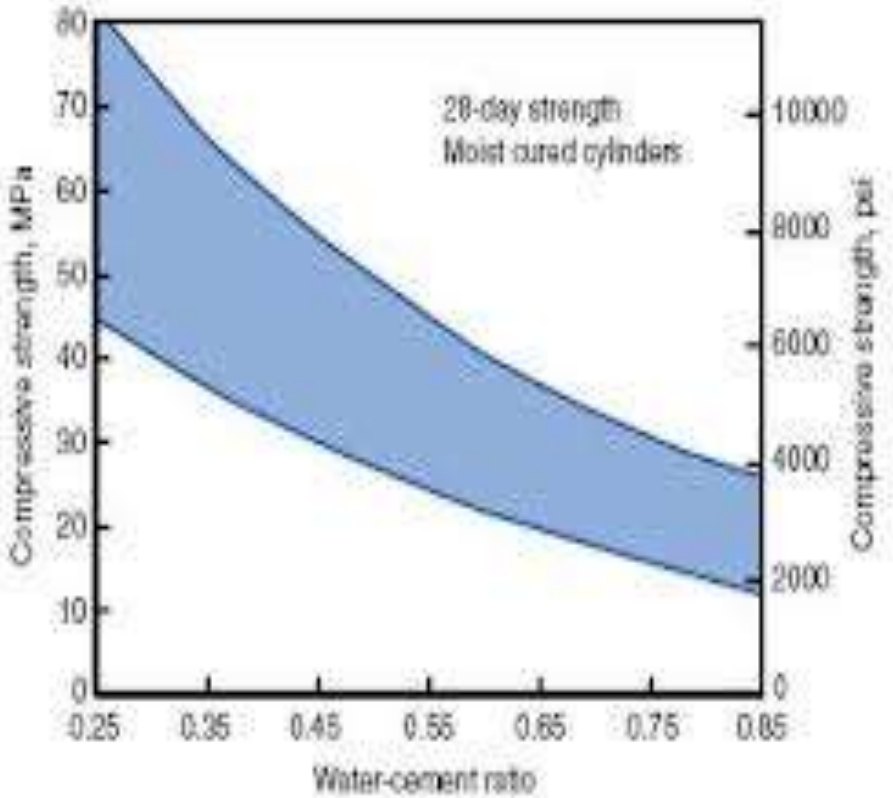
- Even within a given context if the response surface (or “fitness function”) is non-linear and interactive over the design space then the estimated program impacts are “fragile” as opposed to “robust” to design and hence the study results cannot be extrapolated across designs.
- Nearly all existing studies are reported at a level of granularity that lacks “construct” validity

Visualization of construct validity and external validity

- A “design space” delineates all of the variations in design subject to control by the implementers of the project (factors that determine project success but not within the scope of project/program/policy design are “external”)
- A fitness function/response surface/objective function shows the gain in impacts/outcomes from a given design

Interactive effects and produce rugged response surfaces

Concrete is stronger when poured drier…



…only if it is adequately compacted

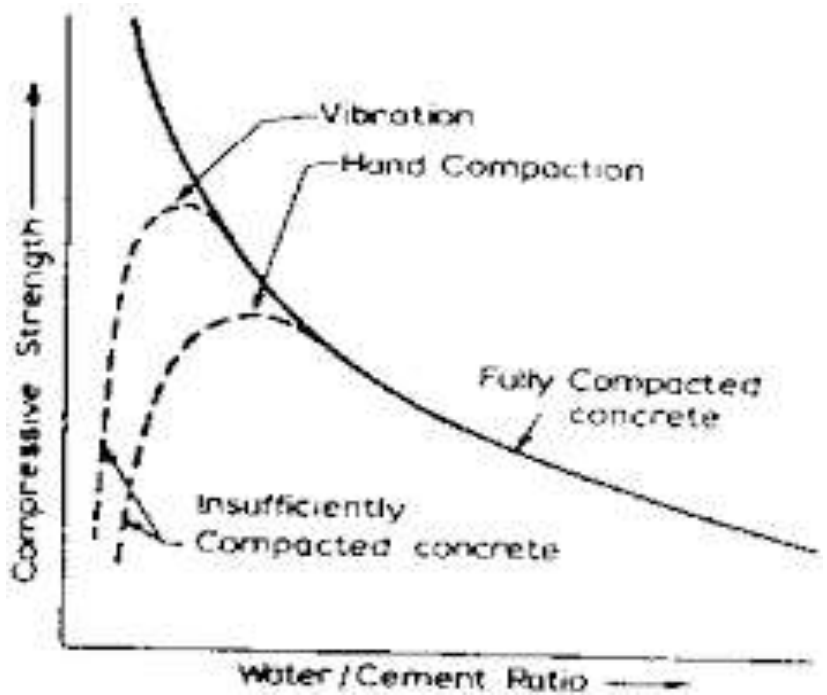
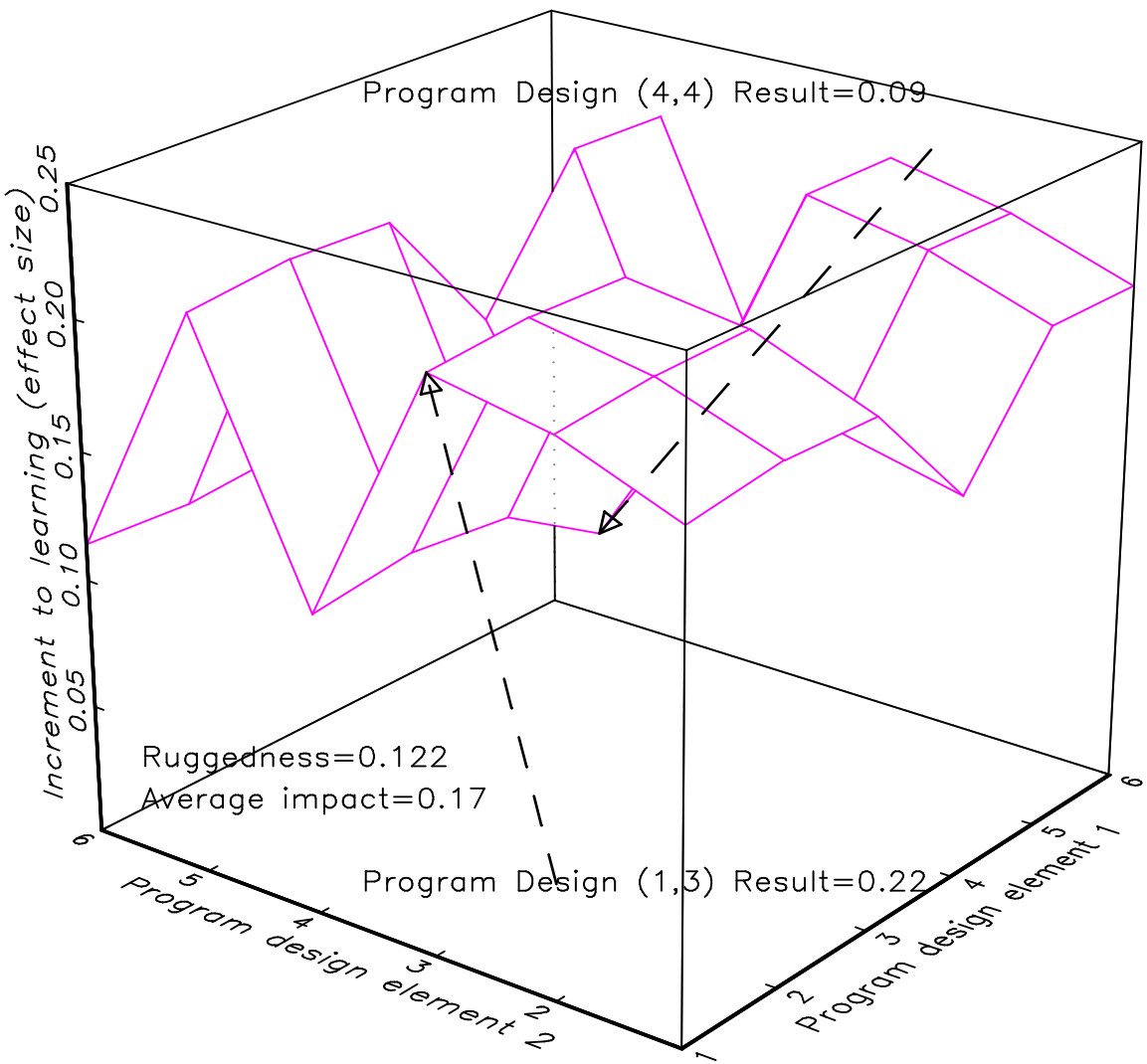


Illustration of a rugged fitness function/response surface

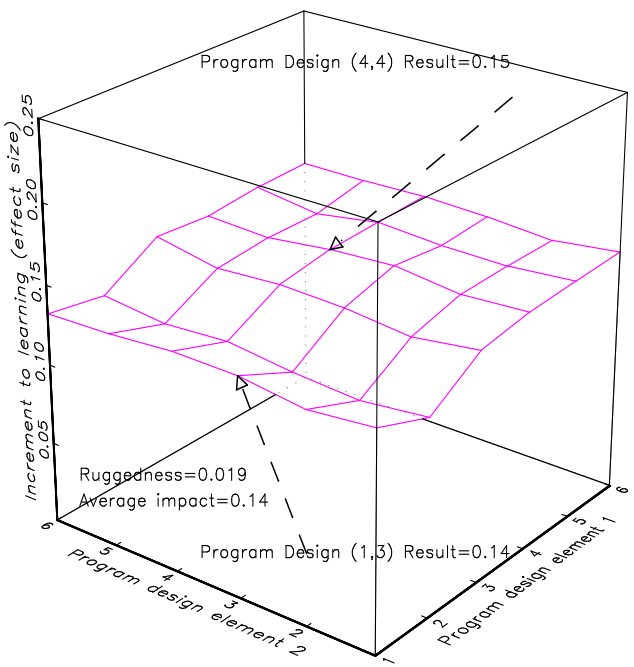


Visually distinguishing “External Validity” from “Design Fragility”

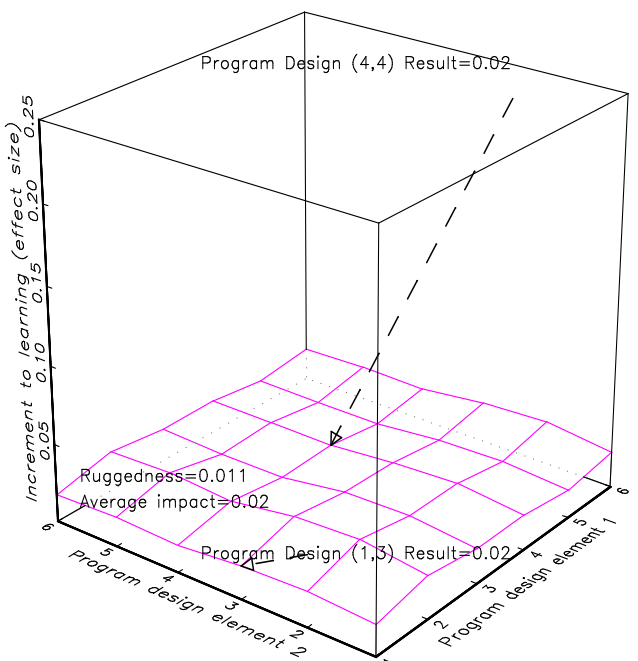
- Four *entirely hypothetical* graphs of possible effect sizes of a *class* of program (e.g. “provision of textbooks”, “ICT in classrooms”, “reduction of class size”, “performance pay”)
- I define “Design Fragility” to be whether there is large outcome variance across *instances* of a *class* depending on program design (or interaction with context)
- Pure “external validity” is whether from “context” to “context” the fitness function is similarly shaped and located

“Pure” external validity

Response surface in context A—design doesn’t matter much, all works

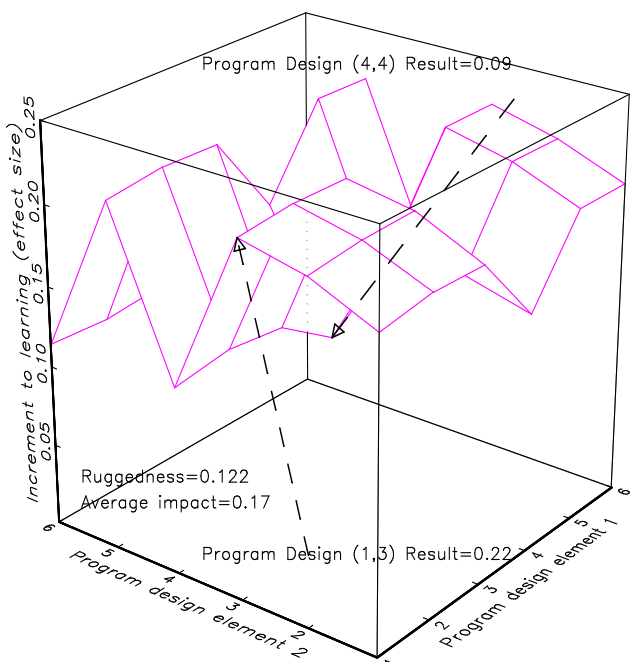


Response surface in context B—design doesn’t matter much, nothing works

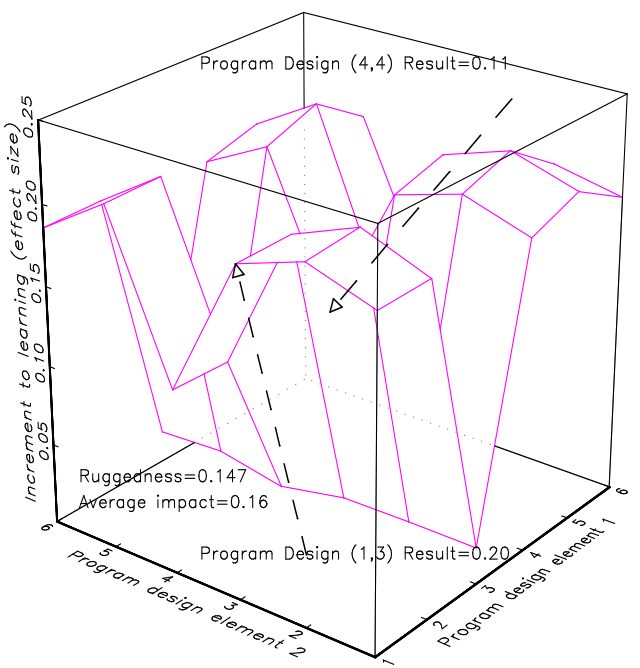


Construct validity: Rugged fitness functions imply different designs produce different results

One "class" of program ("textbook provision")

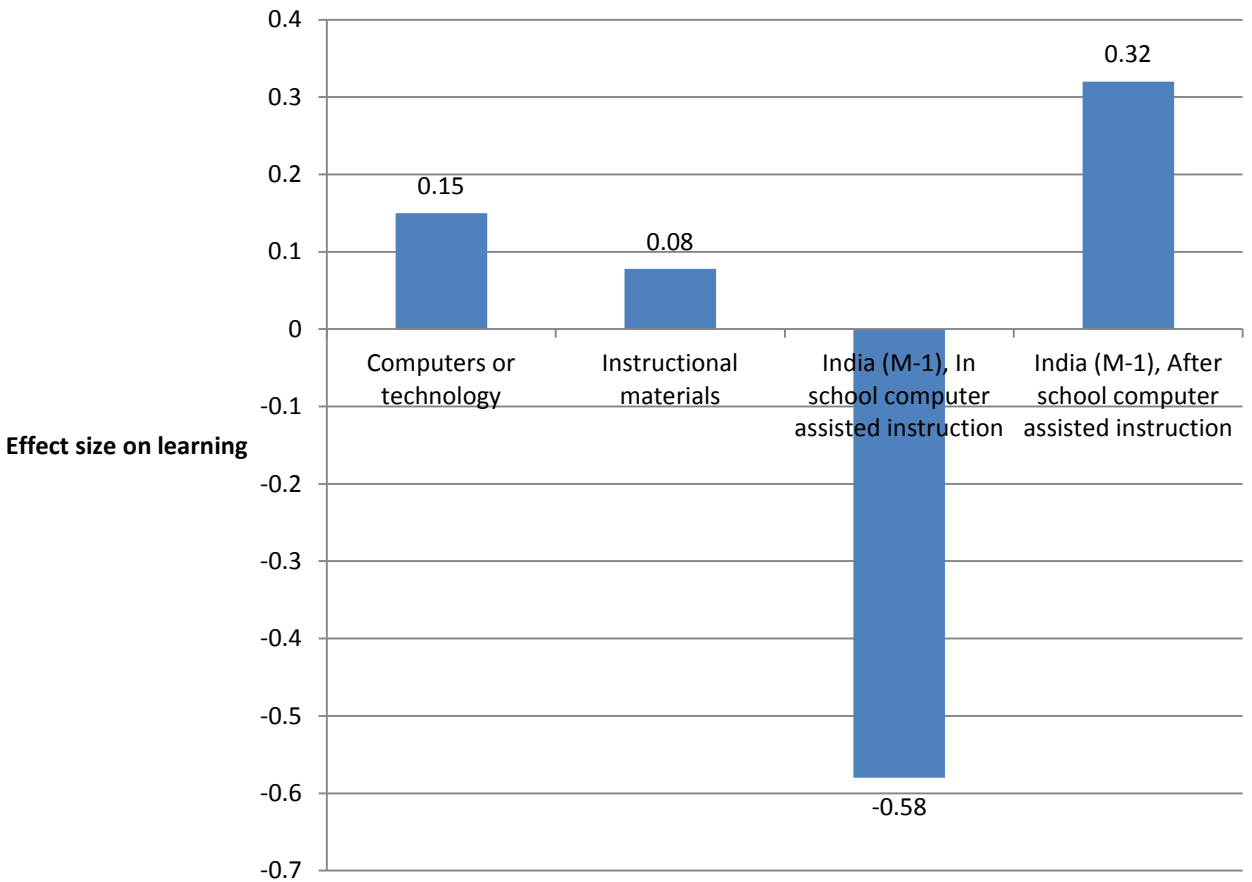


A different class of program ("teacher training")



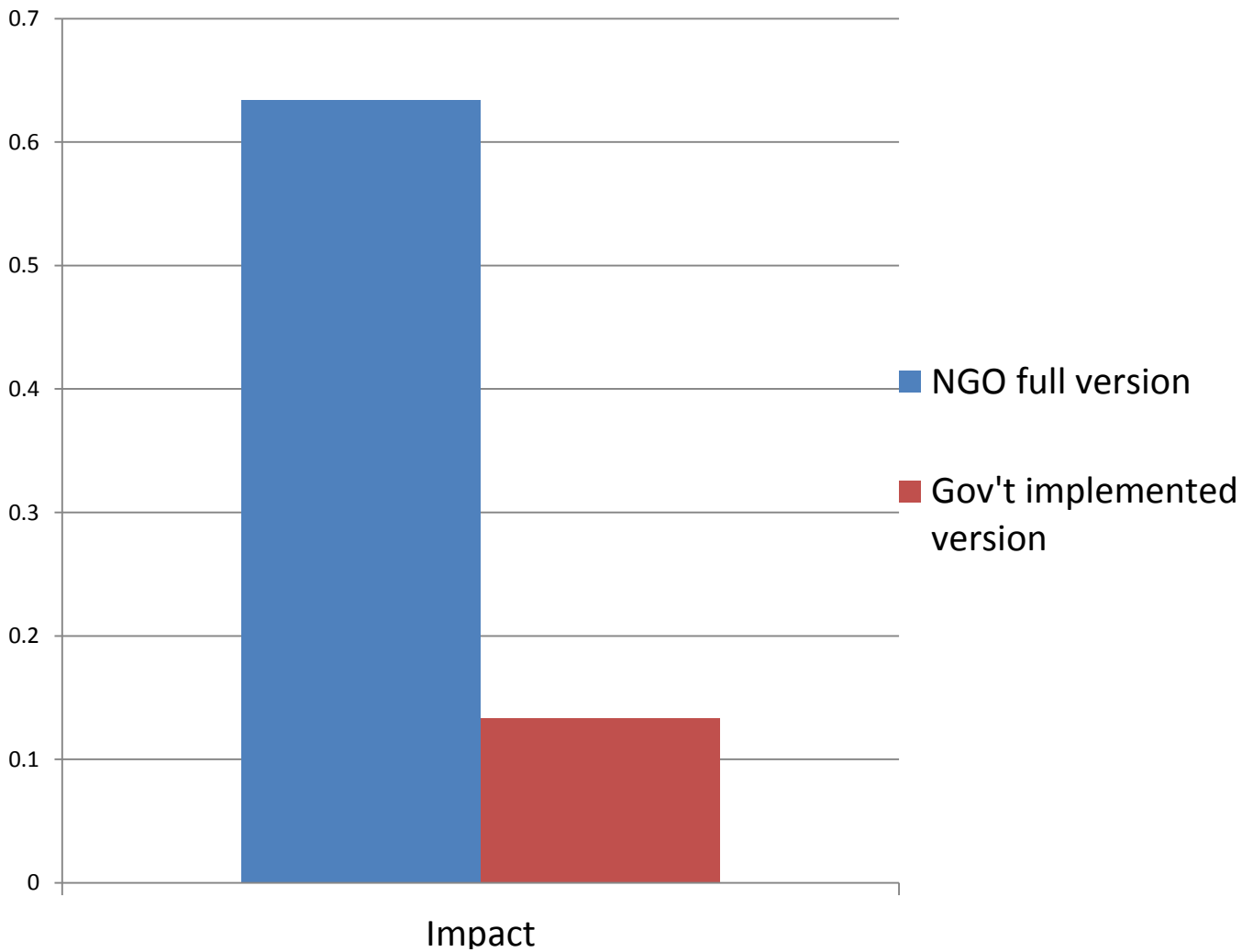
5

Existing "systematic reviews" that compare across classes of projects where there is design fragility produce gibberish as the "within class" variance is huge compared to "across class"



A "systematic review" (McCwan 2015) showed that "ICT" interventions were on average better than "Instructional Materials" interventions—but the across class difference in averages was .07 and two different elements of the ICT average—*which were treatment arms of the same experiment*—differed by .90 effect size. It would seem what *exactly* you do within ICT matters more (by an order of magnitude) than ICT vs instructional materials

Reading program implemented by NGO has massive impact but a modestly different variant had impacts 1/6 as large



5

Vivalt's results suggest a huge amount of the observed variation in impacts is *within papers*—so not "external validity" but "design fragility"

Table 7: Variability across RCT studies for intervention-outcome pairs

(1) Intervention	(2) Outcome	(3) CV(SMD$_i$)	(4) Within paper CV	(5) I^2	(6) Number studies
Conditional Cash Transfers	Enrollment Rate	0.83	0.968	1.00	37
HIV/AIDS Education	Use of contraception	3.12	6.97	0.51	10
Micronutrients	Hemoglobin	1.44	0.731	1.00	46
Median (51 intervention/outcome pairs)		**1.77**		**0.99**	**7 (per pair)**

Source: (Vivalt, 2016), Appendix C, Table 12.

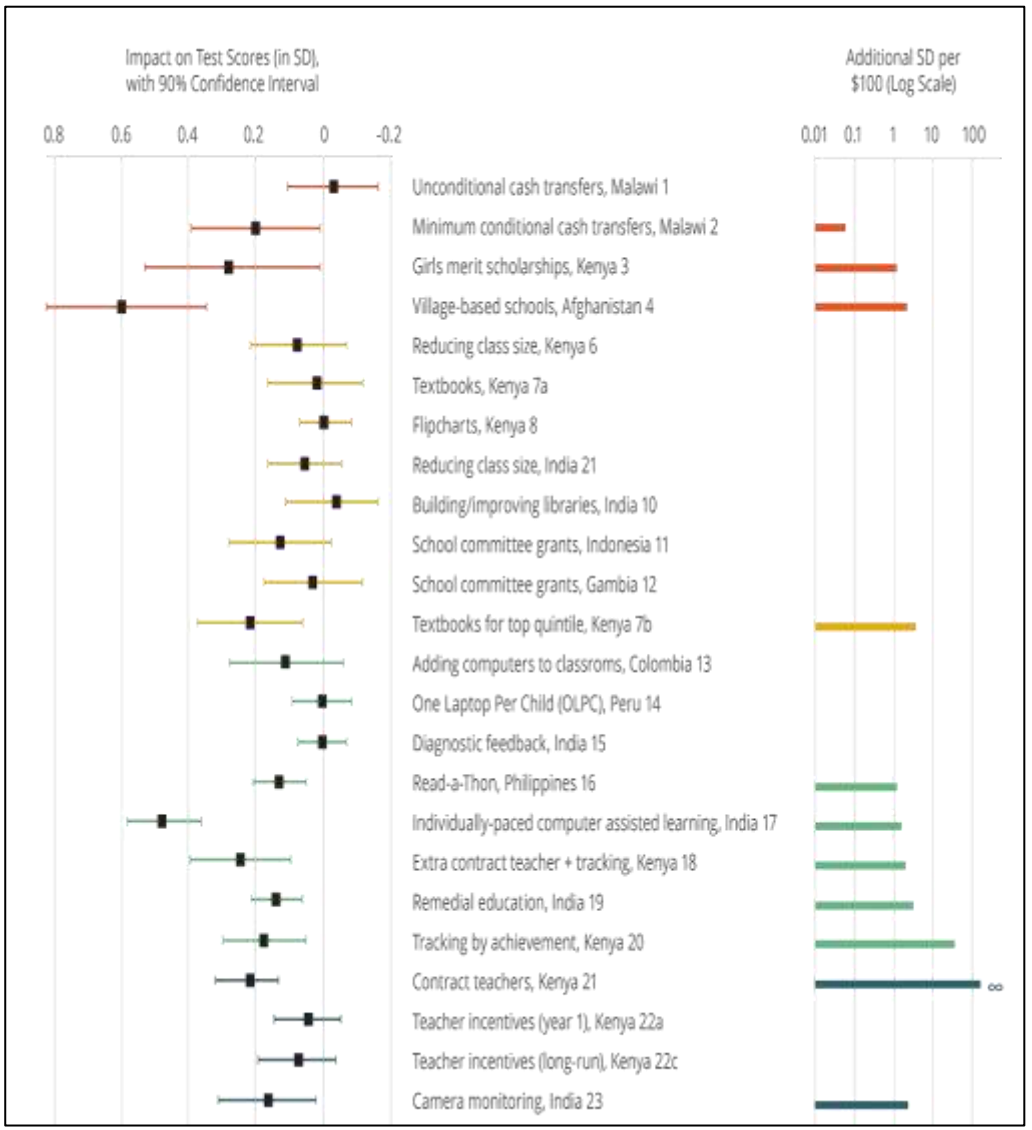
Positive model of program *adoption*

- The ROI on research has to have a positive model of policy/program/project adoption conditional on the information revealed by the research.

- A very simple point: You cannot use as the positive model that your findings reject as the positive mode for believing your “policy recommendations” will be adopted.

- That if you assume your “policy maker” is trying to maximize one objective function when really he/she has a different objective function then research that is conditional on the one has limited NPV as the “recommendation” will not get adopted

Normative as Positive is Silly but the default model



Suppose for a moment we believed these results had external validity (they don't) and were not subject to design fragility (they are)...are they "policy relevant"? "'

Well, suppose they were all from the same "context" (country). Then they would reject *for sure, by orders and orders of magnitude* the chooser of inputs was choosing on the basis of cost-effectiveness.

So the "policy recommendation" or "finding" is the ordering of inputs by cost effectiveness? But we just rejected that the policy maker was interested in that.

Note: Graph replication of original source; ideally results would be displayed based on the cost effectiveness of the bounds of the confidence interval.

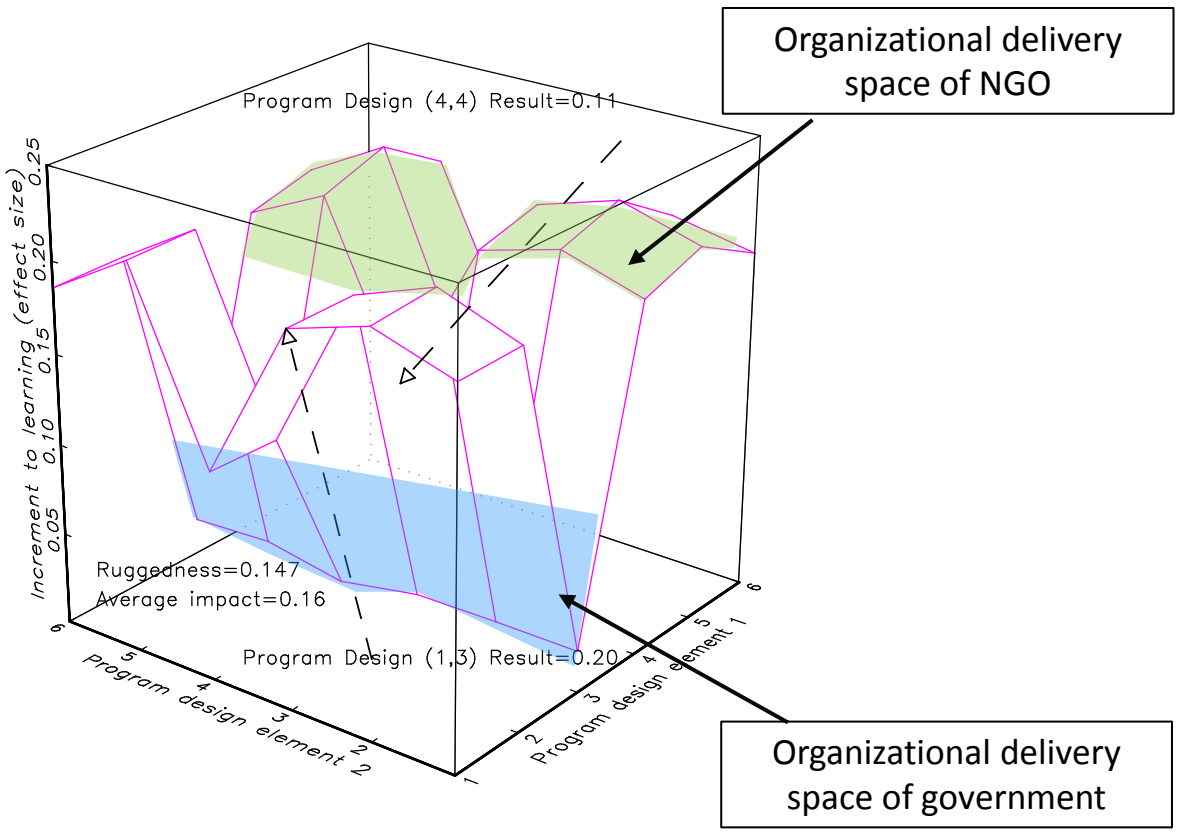
Source: JPAL Education: Increasing Test Score Performance Cost Effectiveness Analysis

Organizational delivery and capability and “match”

- Organizations (both private sector and public sector) have limited ability to change their *nature* and what they are *capable* of implementing changes only slowly.
- This means a finding that a point in the design space “works” even if it has external validity and construct validity and could be adopted might not be feasible for implementation across different organizations.

7 Organization Delivery Capability & Match

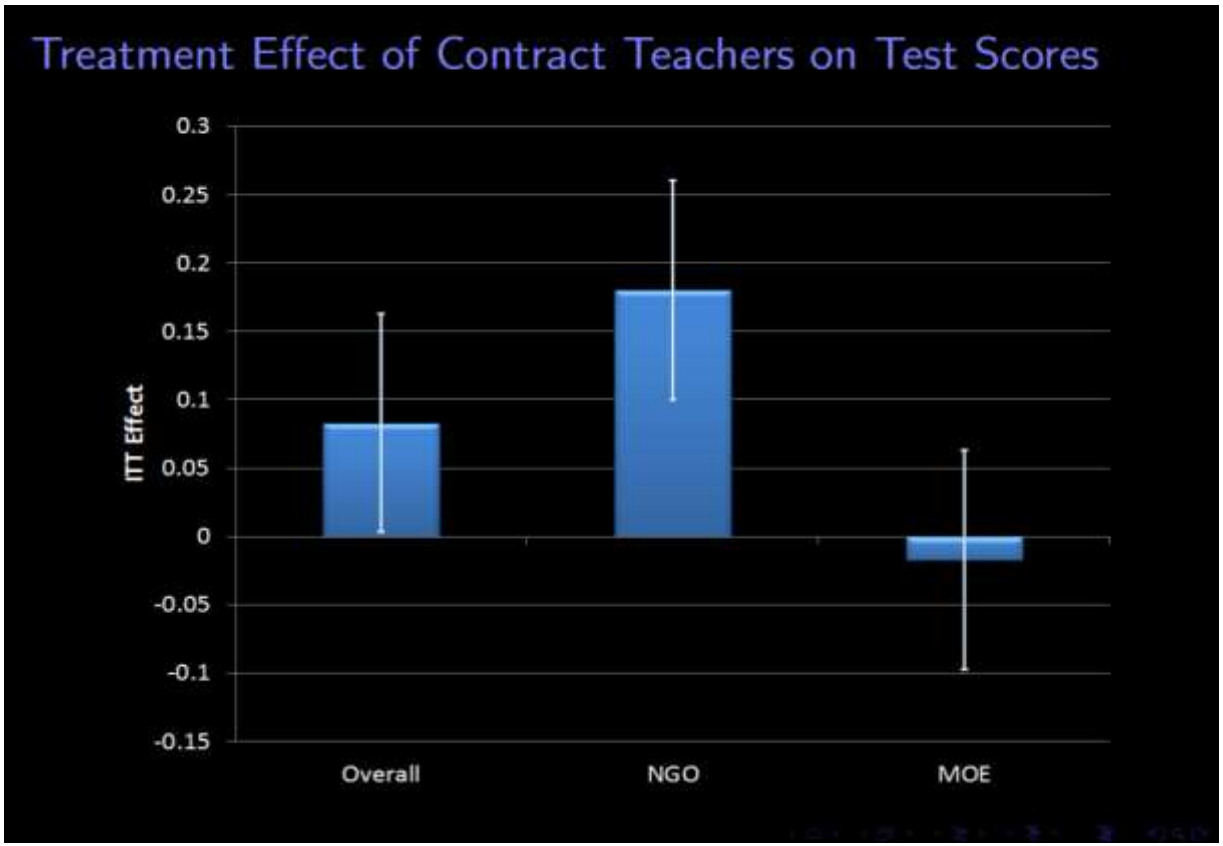
EXAMPLE: DELIVERY CAPABILITY



RESEARCH IMPLICATIONS

- **PPP opportunities / outcomes based funding** for proven organizations (e.g. Bridge Academies International-Liberia; Collaboration Schools – Western Cape)
- **Organizational systems analysis;** Qualitative examples and records of **positive deviance** with replicable components (which states, districts, sub counties have higher than average results in this area? What can be learned?)
- **Case studies of 'failures' / limited results of scale up** – diagnostic of where exactly implementation broke down and if break down was uniform (e.g. Bold, Kenya)

Exact same program of "contract teachers" as shown by rigorous evidence to "work"—but the results did not scale



Source: Bold et al 2013

The key question is whether the government could have implemented the program with fidelity.

My argument is that the program, as it involved giving discretion to "communities" over teacher employment status *could not* be implemented by the government given the very deep constraints of civil service employment, the courts, and how Weberian organizations work—not just "politics"

Most piecemeal funding of RCTs in education have not been, and cannot be justified has having acceptable NPV

	7 Criteria	Common problems with RCT approaches
Potential Impact	• Marginal Return per Dollar	• It does do well the one thing it does well
	• Scope	• The "interventions" identified often have tiny scope (e.g. deworming as an education intervention)
	• Duration	• The impacts on scores often do not persist.
Ability to Replicate	• External Validity	• Other than CCTs on enrollment there are no findings with external validity—and we cannot expect them. This means the cost of each RCT has to be only amortized over its own "context"—and no one knows what that is in space or over time.
	• Design Fragility	• The interventions often are "fragile" and variants of high impact interventions have low impact (and likely vice versa). Four ways textbooks failed but the conclusion cannot be "textbooks are irrelevant to learning"
Ability to Scale	• Political Support	• Often research is funded with no consideration of the likelihood the "recommendation" could be adopted (e.g. "contract teachers" "performance pay") in anything like the current politics.
	• Organizational Delivery Capability & Match	• Many RCTs are organized as 'field experiments" or the implementation is handled by an NGO and the viability of scaled implementation is not considered.

Agenda

- Some motivation about learning in developing countries
- The ROI Criteria
- Tentative Implications for Practitioners & Researchers (if we get to it)

Within research there are two categories: academic scholarship and organizational learning

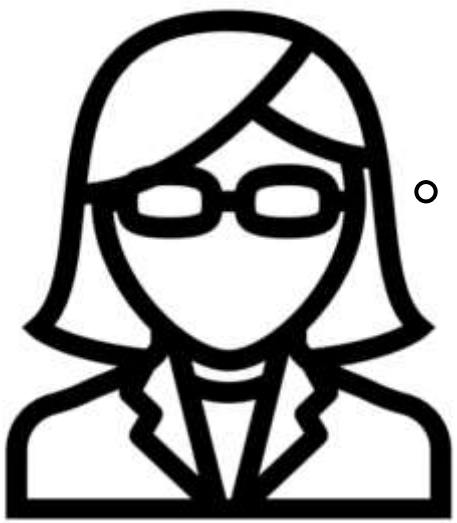
ACADEMIC SCHOLARSHIP

- Produced within an academic context, often by professors or PhD students
- Scholarship is discipline based
- Peer-reviewed publication and tenure is the objective of scholarship

ORGANIZATIONAL LEARNING

- Internal or externally contracted work to enable better program implementation
- Not confined to disciplinary boundaries
- Working papers and consulting reports (e.g. gray literature) as well as direct working contexts (e.g. technical assistance) that transmit knowledge, both codified and tacit
- The objective is to improve performance

Implication: A Practitioner may use our framework to ask herself two questions to identify high ROI research

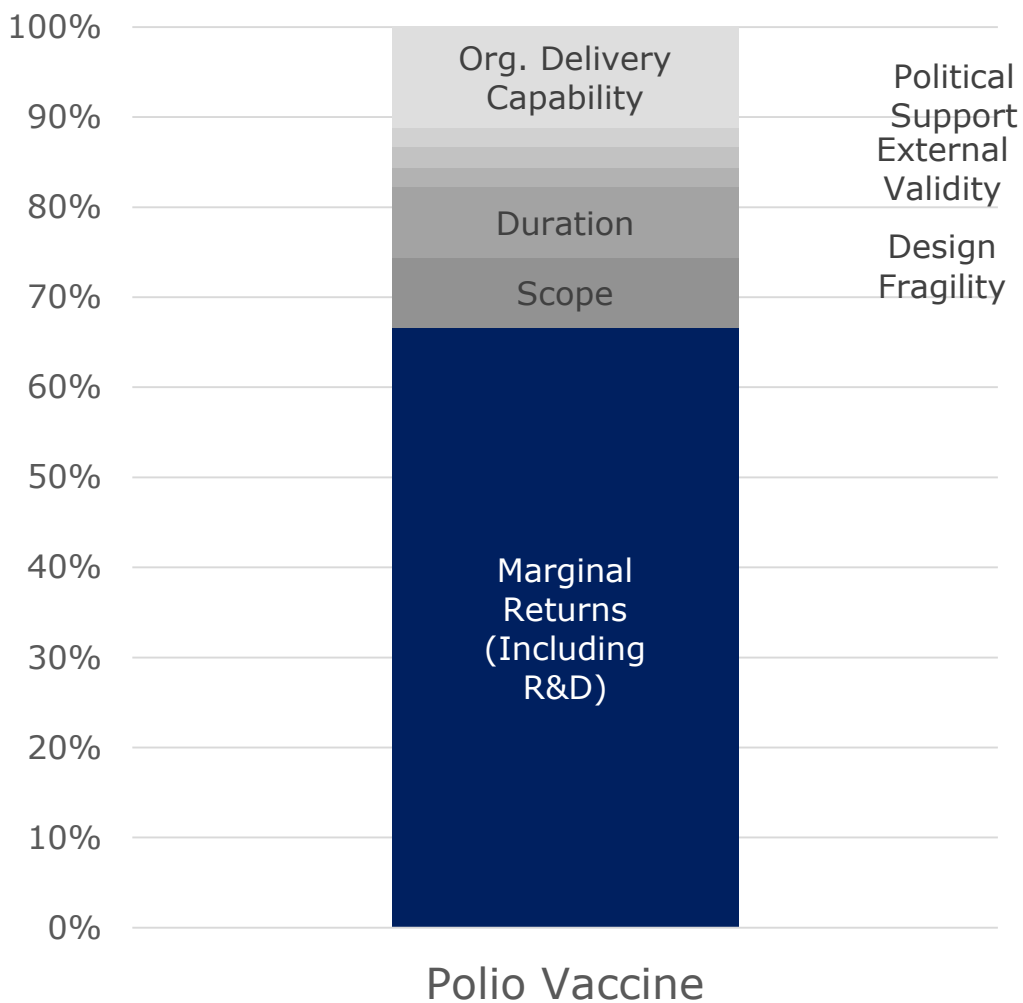


[Classic Scholarship Question] What is something that **could** be done (by someone) but we **don't know about its impact on outcomes**?

[Classic Org. Learning Question] What has **high impact**, but we **don't know** how to **get adopted**, **replicate or scale?**

Arguments for RCTs in development often rely on the analogy of the "gold standard" of drug trials—which I argue is a badly flawed comparison

RELATIVE RESEARCH INVESTMENT

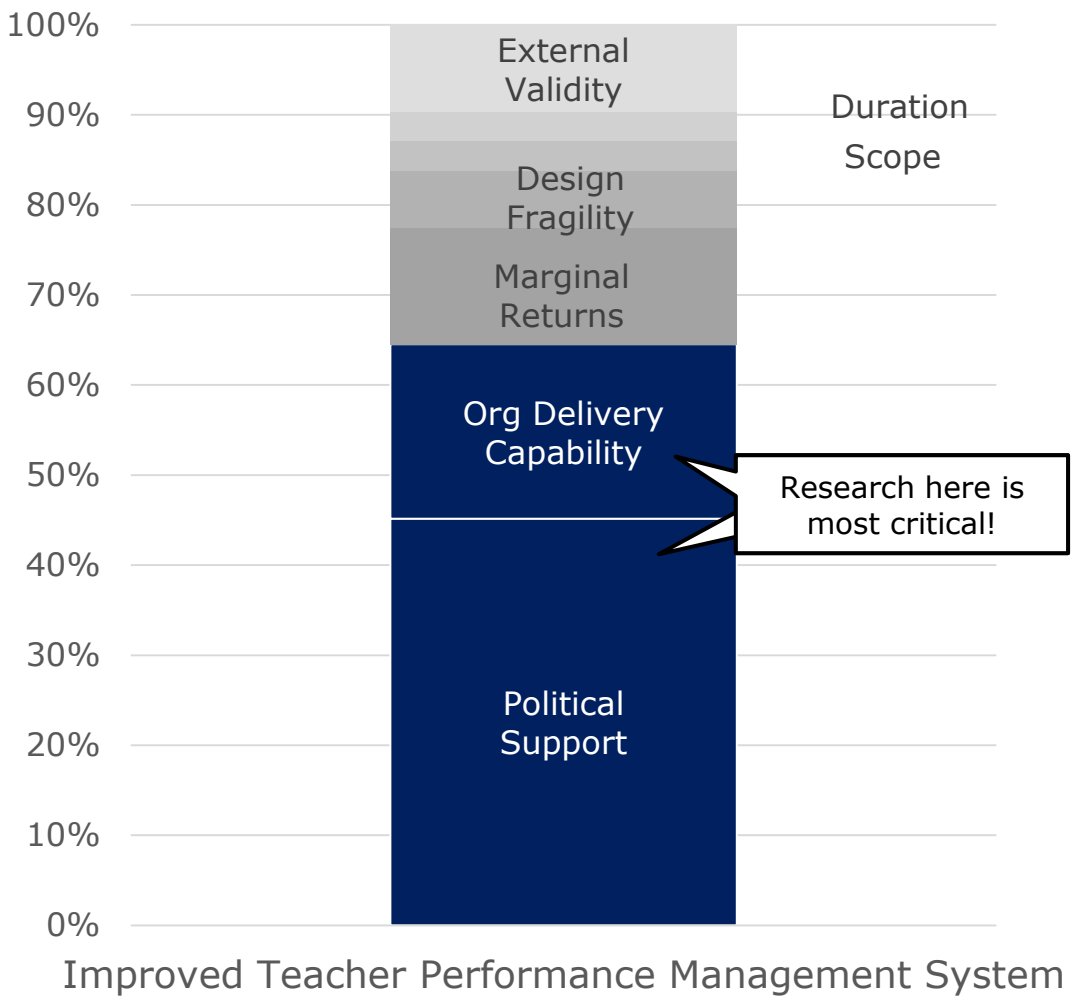


EXPLICATION

- **Marginal Returns:** When creating a new vaccine, the vast majority of investment occurs to do R&D and run RCTs to validate effectiveness; populations are also prioritized with highest disease burden to maximize cost effectiveness
- **Scope:** Optimal number of treatment times and sequencing is often determined after product is established
- **Duration:** Longitudinal studies establish length of effect
- **Design Fragility:** This is very clearly specified in all its detail because there is no "robustness"
- **External Validity:** Limited concern as vaccines generally have similar implications across geographies, time, and populations
- **Political Support:** Vaccines encounter little political resistance – introducing new medicine to populations is politically favorable
- **Organizational Delivery Capability:** Organizations (hospitals, clinics) are generally aligned to delivery vaccines; parallel systems made be built (e.g. Donor run vaccine drives)

~~In education the question of "what would~~ have high impact if done?" is secondary to "what can be done that would have impact?"

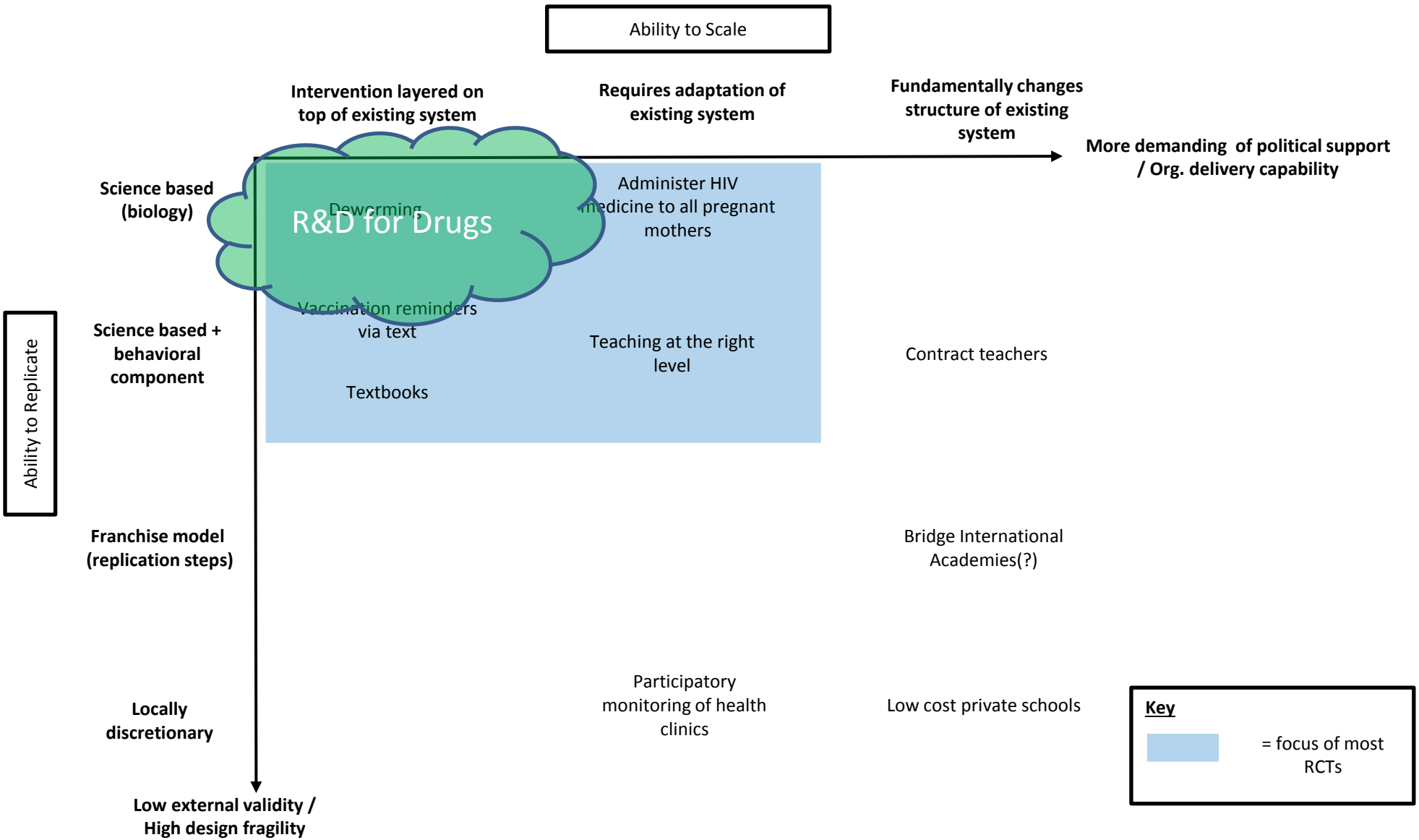
RELATIVE RESEARCH INVESTMENT



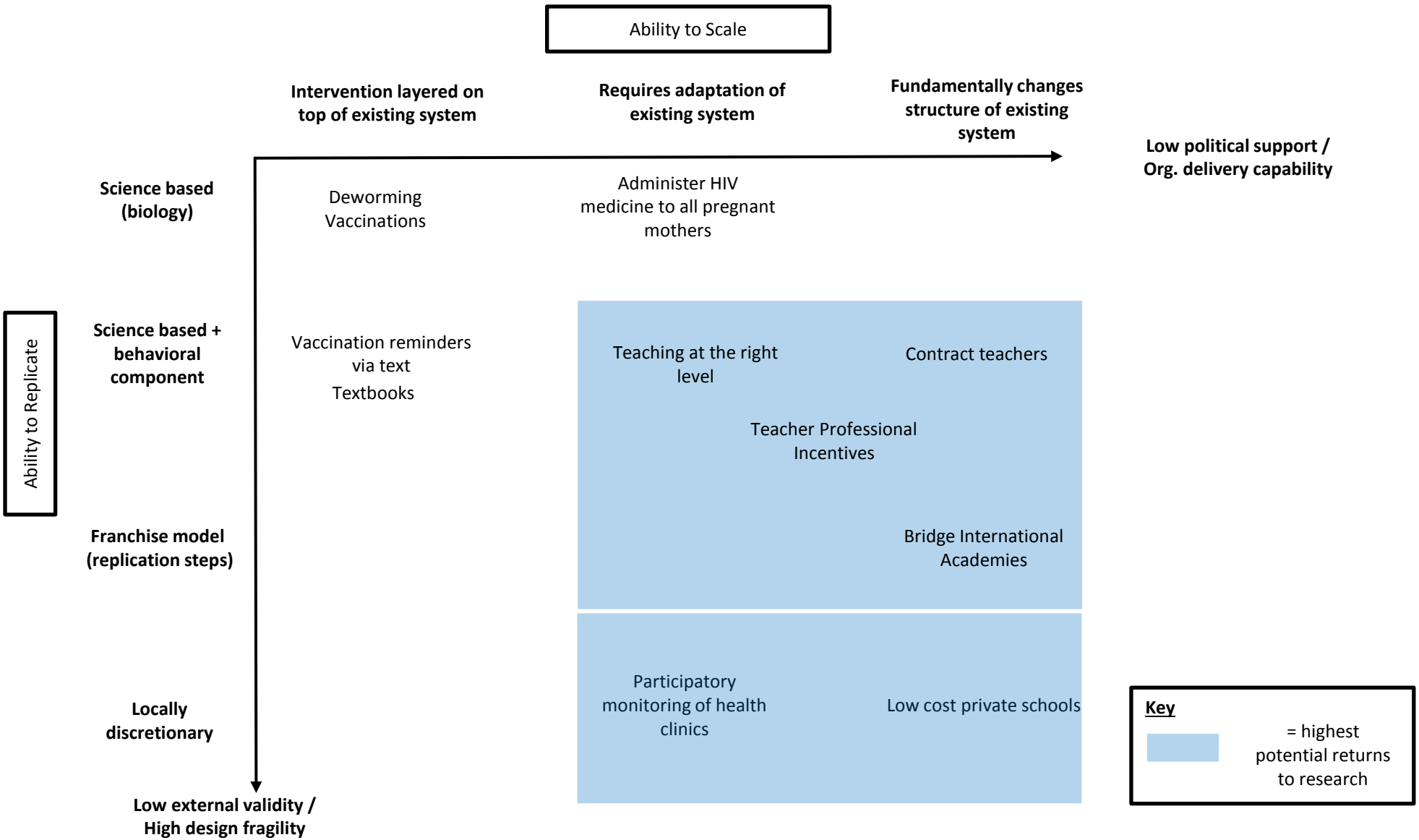
INTERPRETATION

- **Political Support:** A feasible performance management system cannot be designed independent of an analysis of the political economy, union strength, political interests etc.
- **Org Delivery Capability:** The design of such a system must account for capacity of the Ministry of Education (e.g. frequency of evaluation, independence of evaluation staff)
- **Marginal Returns:** Assessment of marginal returns to performance management systems should only occur for policies that are politically adoptable and able to be delivered by the organization
- **Scope:** Different iterations of marginal returns will likely test the question of support
- **Duration:** If successful, a reform performance management system should permanently re-arrange incentive structures
- **Design Fragility:** Feasibility of policy delivery should be assessed between state and district governments as differences may arise
- **External Validity:** An analysis of the unique cultural and political features that enabled a change are critical to explore scale outside the country of interest

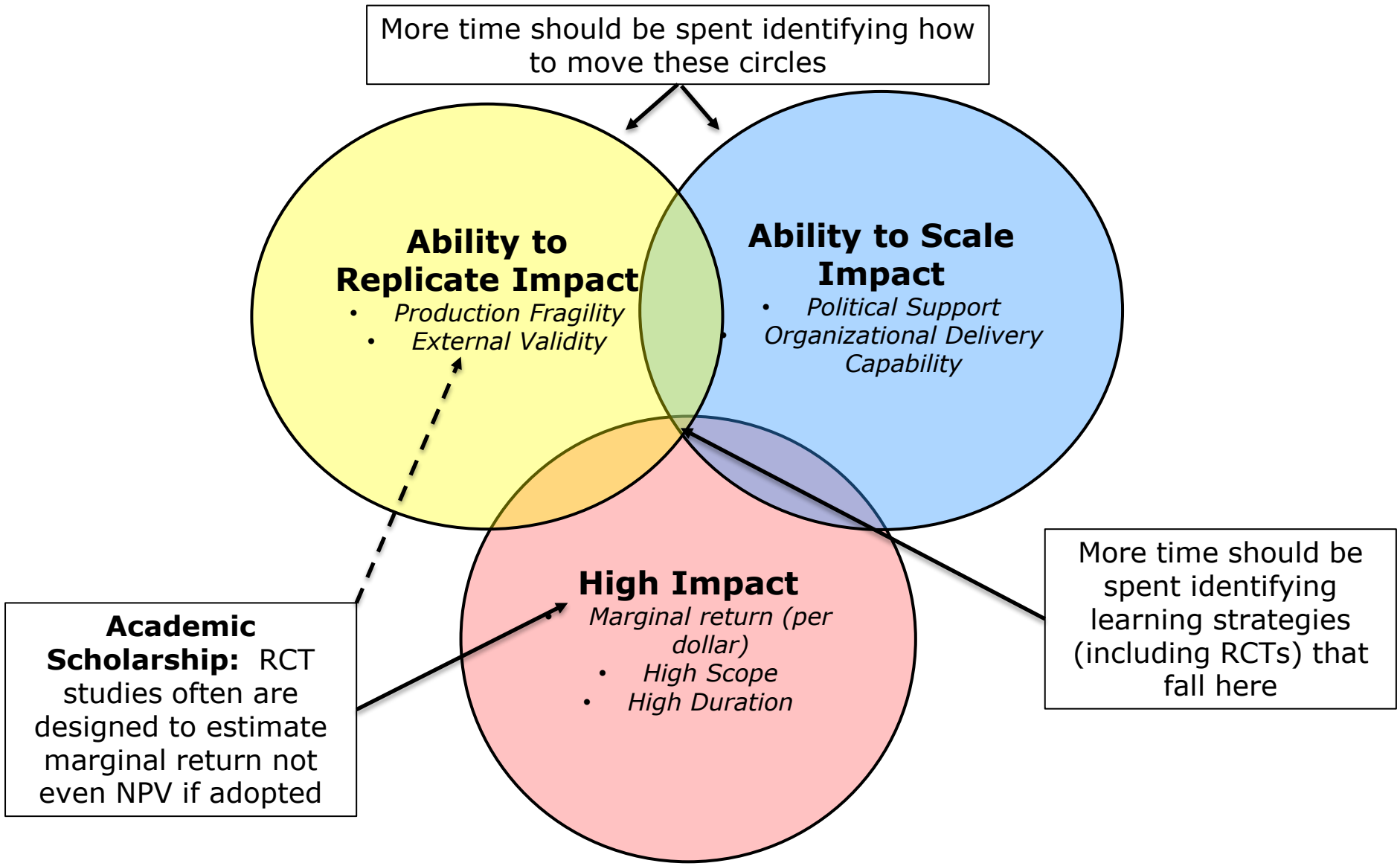
The implications: research has historically occurred in areas that are thought to be similar to drugs in health



However high impact interventions are likely outside that space



Academic scholarship primarily focuses on a small subset of the research questions that affect performance



Funders: Funders may choose to invest in one section, or along the value chain

Potential Impact	Ability to Replicate	Ability to Scale