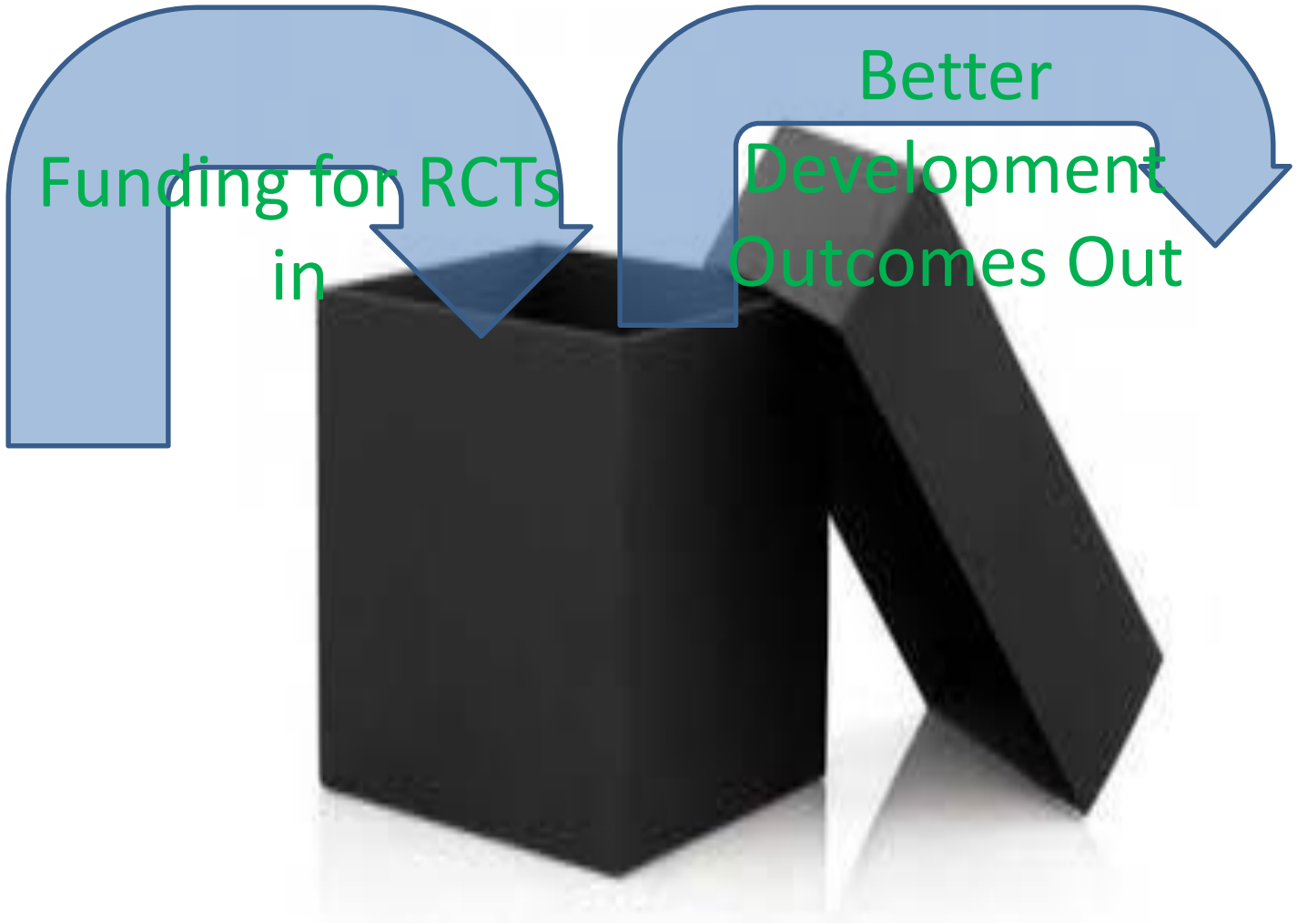
The Debate about RCTs in Development is Over: We Won. They Lost.

Lant Pritchett

February 21st, 2018

DRI/NYU

The important claims were about how to get published in economic journals but about *impact on development outcomes*—what was in the black box?



Outline

- I am going to build a conceptual and visual framework for analyzing the complete "logframe" or "theory of change" of RCTs as a development tactic/strategy
- With that conceptual/visual apparatus I am going to show the "first generation" RCT claims were false about five key points and that "second generation" practice have conceded most of these points

Preliminary I: Design Space and Response Surface

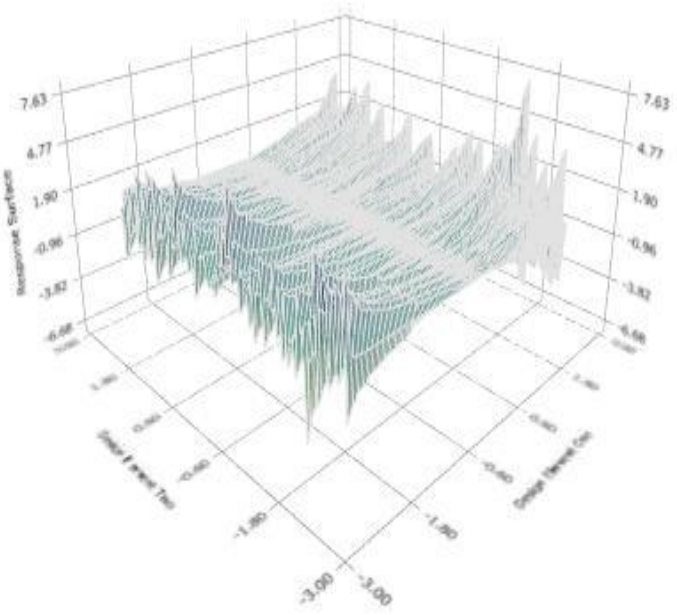
- *Design space* is all of the possible ways a given *class* of interventions can be designed—e.g. names like “teacher training” or “conditional cash transfers” or “livelihoods” or “microcredit” or “vocational training” designate *classes* defined by a design space.
- Any actual project/program/policy is an *instance* of that class where the *instance* is an element of the design space (e.g. *this* CCT gives how much of what to who on what conditions (etc.))

Even a super simple class of program, like a "CCT" has many design elements

Table 4: Design Space for CCT projects, illustrated with three specific CCT projects

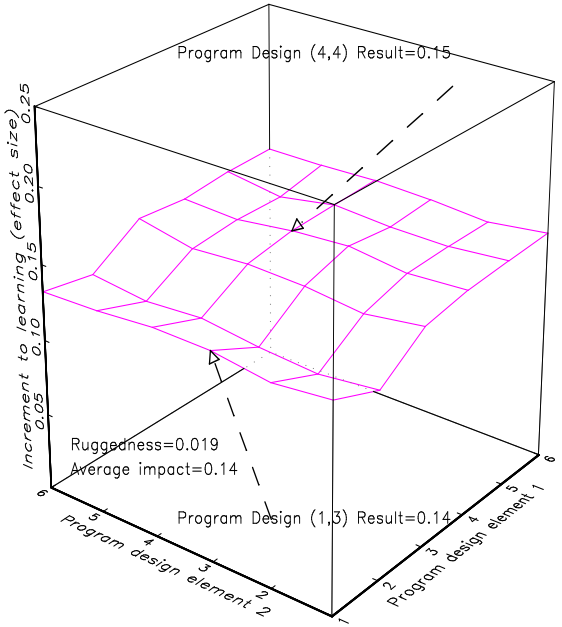
Dimension of design space of a CCT	PROGRESA, Mexico (Oportunidades)	Red de Protección Social, Nicaragua	Malawi
Who is eligible?	Poor households (census + socioeconomic data to compute an index)	Poor households (geographical targeting)	District with high poverty and HIV prevalence.
To whom in the household is the transfer paid?	Exclusively to mothers	Child's caregiver (primarily mother) + incentive to teacher	Household and girl
Any education component to the CCT?	Yes – attendance in school	Yes – attendance in school	Yes – attendance in school
What are the ages of children for school attendance?	Children in grades 3-9, ages 8-17	Children in grades 1–4, aged 7–13 enrolled in primary school	Unmarried girls and drop outs between ages of 13-22
What is the magnitude of the education transfer/grant?	90 – 335 Pesos. Depends on age and gender (.i.e. labor force income, likelihood of dropping out and other factors).	C$240 for school attendance. C$275 for school material support per child per year.	Tuition + $5-15 stipend. Share between parent ($4-10) and girl ($1-5) was randomly assigned.
How frequently is the transfer paid?	Every 2 months	Every 2 months	Every month
Any component of progress in school a condition?	No	Grade promotion at end of the year.	No
Any health component of the CCT?	Yes – health and nutrition	Yes - health	Yes – collect health information
Who is eligible for the health transfer?	Pregnant and lactating mothers of children (0-5)	Children aged 0–5	Same girls
What health activities are required?	Mandatory visits to public health clinics	Visit health clinics, weight gain, vaccinations	Report sexual history in household survey (self-report)
Who certifies compliance with health conditions?	Nurse or doctor verifies in the monitoring system. Data is sent to government every 2 months which triggers food support.	Forms sent to clinic and then fed into management information system.	

A response surface or fitness function is the mapping from the design space to an outcome of interest

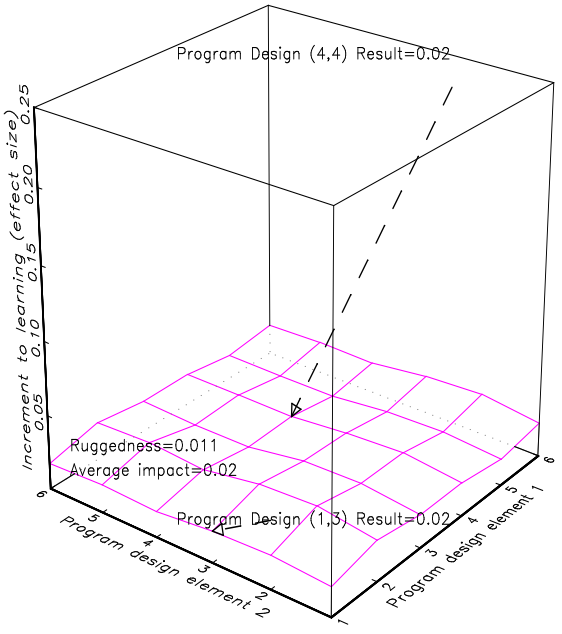


“Pure” external validity

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

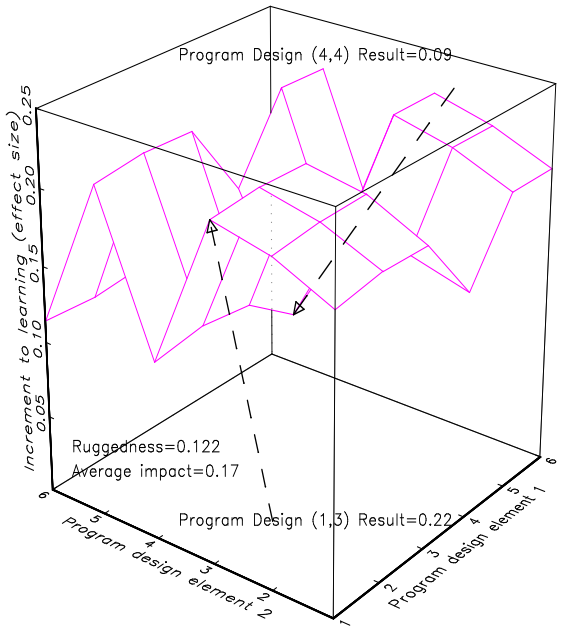


Response surface in context B—design doesn’t matter much, nothing works

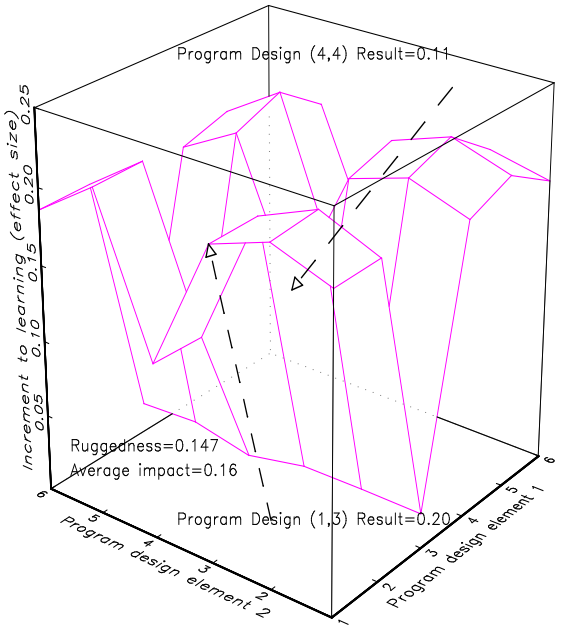


Construct validity: Rugged fitness functions imply different designs produce different results

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



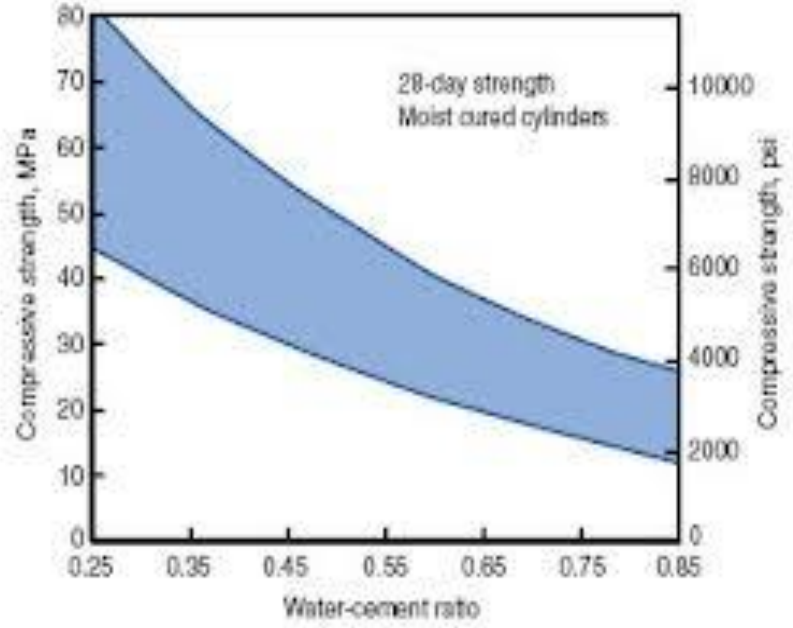
A different class of program ("teacher training")

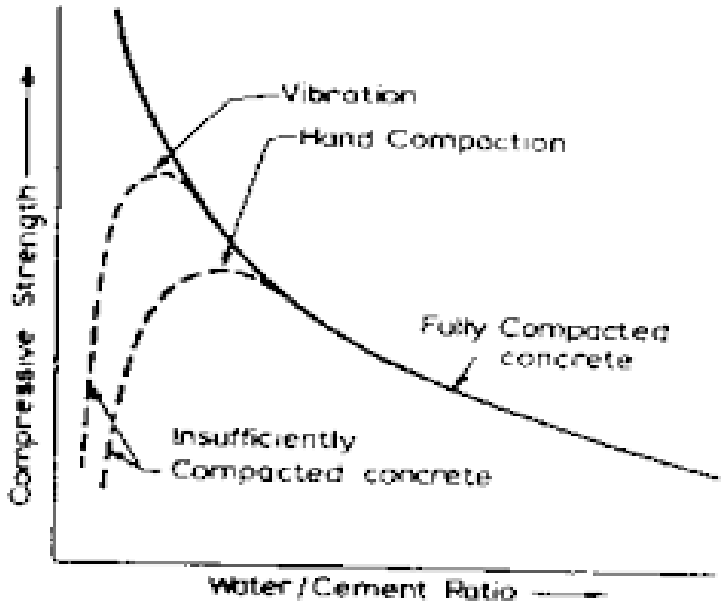


A concrete analogy: interactive effects and produce rugged response surfaces

Concrete is stronger when poured drier…

…only if it is adequately compacted when it is poured dry





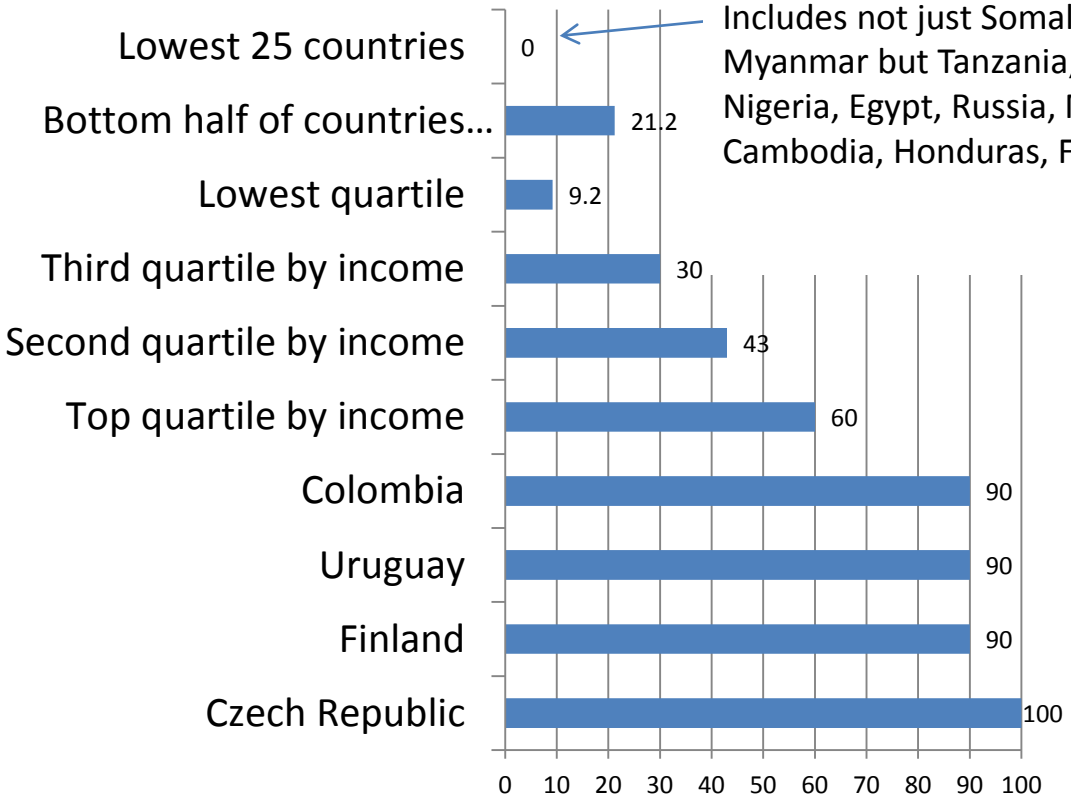
If design space is: water/cement ratio and compaction then RCTs varying the water/cement ratio will recover very different results along the “fully compacted” design (first graph, solid line in second graph) versus other degrees of compaction

Preliminary 2: Organizational capability for policy implementation

- A policy is a mapping from states of the world to actions by agents (with an objective)
- An organization's *capability for policy implementation* is the capability to induce its agents to correctly assess and act on states of the world in ways that promote the (stated) objectives of the policy.
- Achieving success from various policies requires different (in quantity and potentially in kind) degrees of capability for policy implementation

An example where every country has the **same** policy but outcomes span the possible range: *all* differences are due implementation

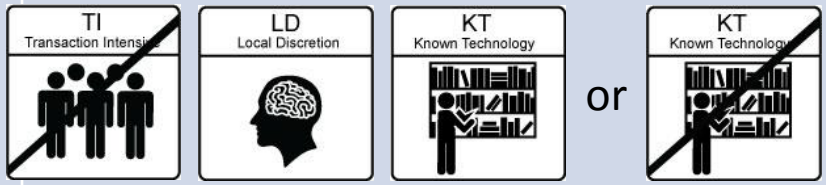
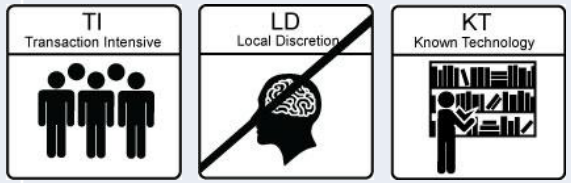
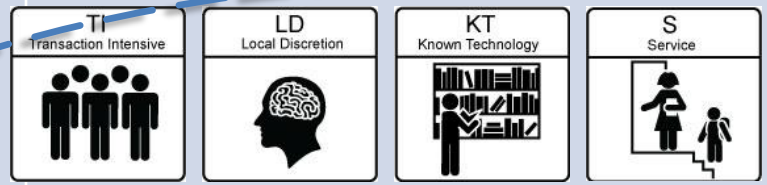
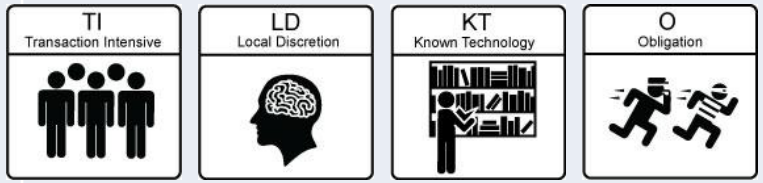
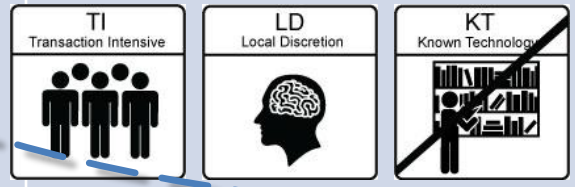
Percent of 10 misaddressed letters coming back to USA within 90 days (all countries agree to return within 30 days)



Includes not just Somalia and Myanmar but Tanzania, Ghana, Nigeria, Egypt, Russia, Mongolia, Cambodia, Honduras, Fiji, etc.

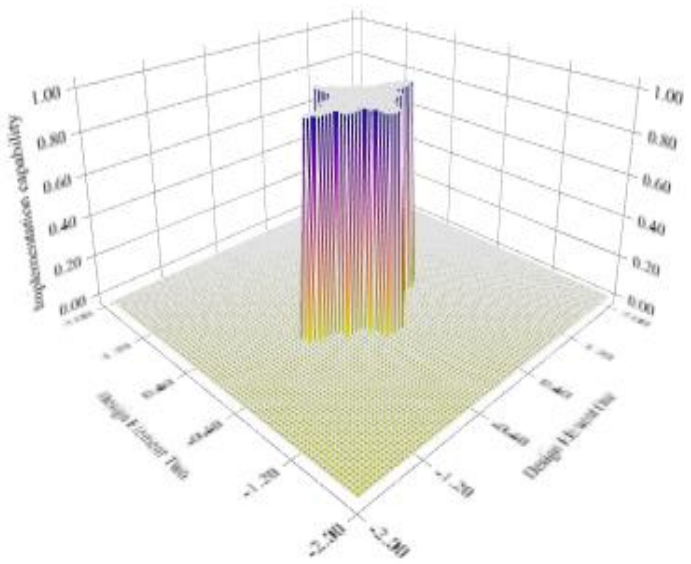
Source: Chong et al 2014

21st Century is about state capability implementation intensive challenges

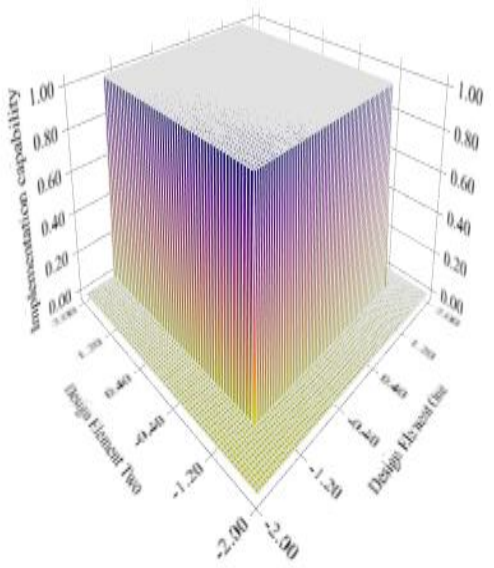
		Health	Finance
Policy making		Iodization of Salt	Monetary policy
Logistics		Vaccinations	Payment systems
Implementation intensive service delivery		Curative care	Loans
Implementation intensive imposition of obligations		Regulation of private providers	Regulation of private providers
Wicked hard		Preventative health	Equity financing of start-ups

Mappings of organization capability to replicate a policy/program/project with fidelity over the design space

Limited implementation capability



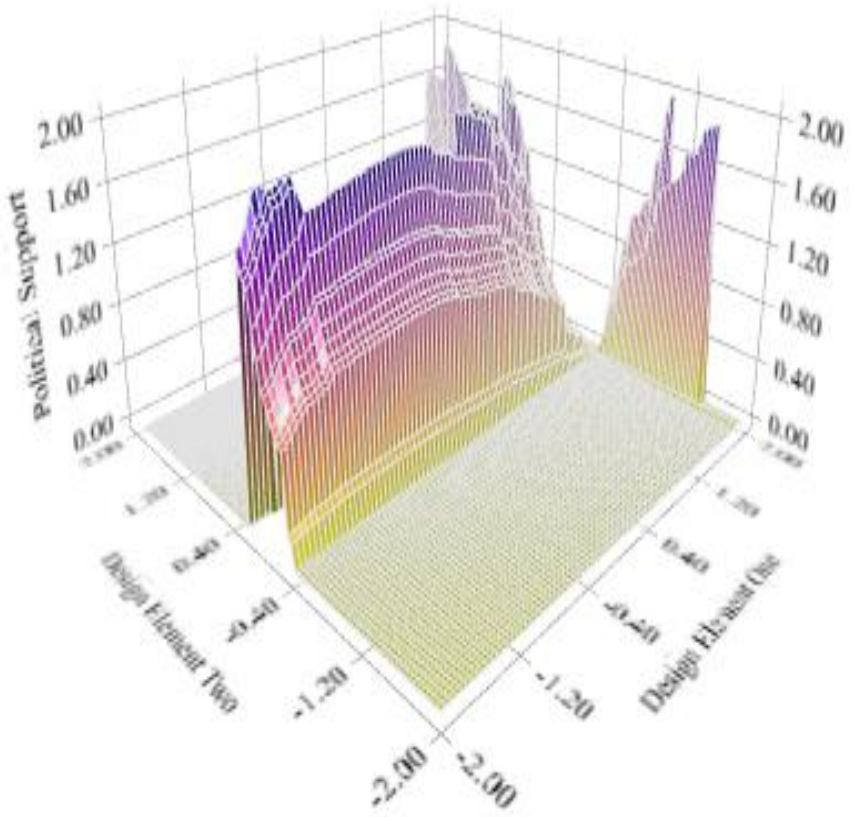
Lots of implementation capability



Preliminary number 3: Political support

- For a variety of reasons (both benign and non-benign) the support may be different for different elements of the design space
- So there is also a surface over the design space of those that can “generate and maintain sufficient political coalitions to sustain authorization for implementation”

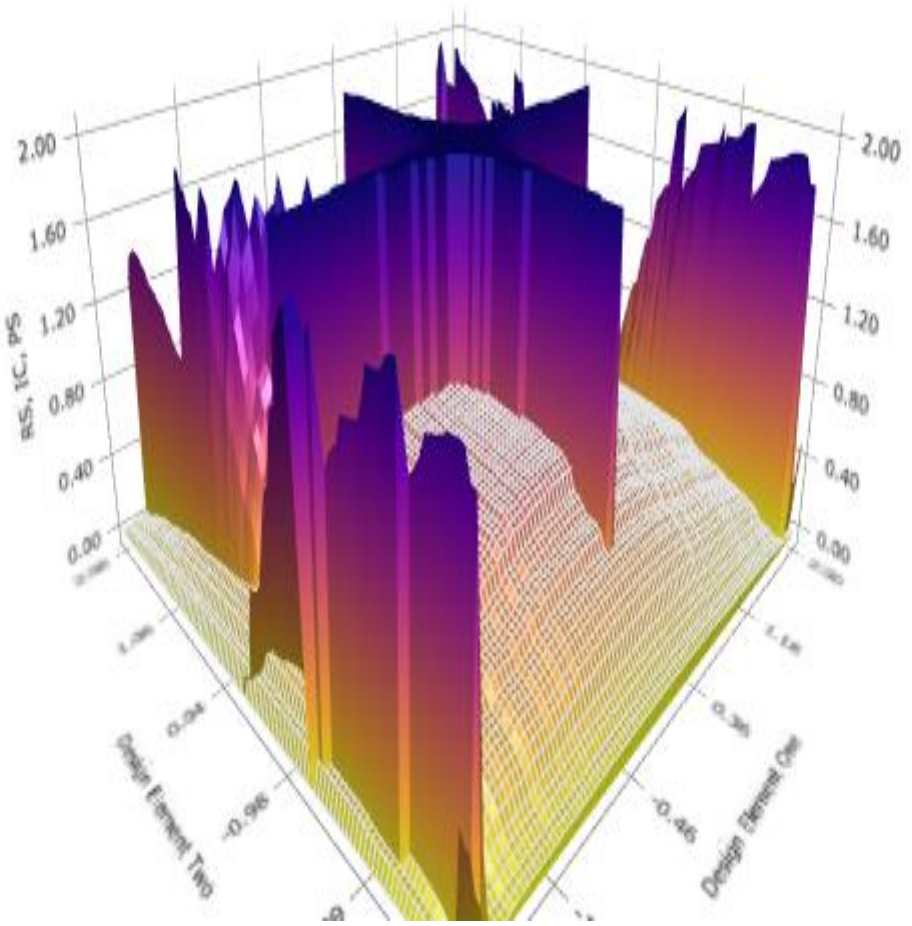
Political support surface



In order to increase well-being a Policy/Program/Project has to meet the Trinity

- *Instrumentally correct:* the design has to be such that, if it were implemented with fidelity it would lead to higher levels of well-being for the intended beneficiaries.
- *Administratively feasible:* The responsible organization has to be able to implement with reasonable fidelity the P/P/P with the resources made available to it.
- *Politically supportable:* One has to create and sustain a political coalition with sufficient power to authorize the P/P/P

Response Surface with Implementation and Politics



The first generation *randomistas* claim

Significantly more funding of rigorous *independent impact evaluations* using techniques of randomized control trials will lead, not just to more academic papers with firmer results, but to actual significant improvements in the development process (policies, programs, projects) that will lead to higher human well-being.

Put another way

The kinds and types of knowledge that can, in principle and in practice, be generated by applying RCT techniques via independent impact evaluations to development projects/programs/policies are a key binding constraint on development practice (e.g. has a very high Lagrangian) and hence greater investments in RCTs will lead *pari passu* to significantly higher levels of human well-being cost effectively (relative to other available investments)

The "RCT as IIE" or *randomista 1.0* logframe for development impact has six *necessary* steps and five of the six are false

The knowledge about the response surface over P/P/P acquired through RCTs ….	
…can be generated about highly consequential actions	False. National development is a four fold transformation at *ontologically* aggregate process and individuated interventions are second order.
…leads to feasible large scale interventions	False. Efficacy of P/P/P is mostly limited by low organizational capability for implementation not knowledge of the response surface.
…either is in regions of political support and/or changes political support sufficient to authorize action	False. RCT knowledge has no special traction on political decision making.
… is of sufficient construct validity to guide action	False. Response surfaces are rugged over super high dimensional design spaces.
…is of sufficient external validity to be "amortized" and made cost effective	False. The external validity of RCT evidence is in many/most key instances is l
…is superior to other evaluation methods.	True.

	2018: Debate over. *Every* point to non-RCT advocates.
Topics important for development	National Development leads to better well being. National development is ontologically a social process (markets, politics, organizations, institutions). RCTs have focused on topics that account for roughly zero of the observed variation in human development outcomes.
Organizational capability and learning	Organizations doing any non-logistical activity (and most even of those) cannot be beaten into doing better by evidence from "independent" outsiders.
Political economy	There is massive evidence that governments do not implement many many many projects/proposals/programs that are cost effective and do spend budget on items known to be not cost effective. The NAP model of a benign SWF planner hampered by lack of rigorous evidence on effectiveness whose behavior an RCT will change is complete wack nonsense.
Construct validity	RCTs examine an instance (or small numbers of treatment arms) which, in a rugged response surface over a high dimensional design space reveals next to nothing. Simple iterative methods dominant RCTs in finding good policy designs.
External validity	External validity (a) logically incoherent when existing evidence has variance, (b) RCTs worse predictors of impact than OLS, (c) reviews show massive variance. If experiments were the hallmark of science alchemists would win Nobel prizes.





Two points before we even start

- The RCT movement's claims about development impact were always *faith based,* not evidence based—even in the weakest sense of evidence (e.g. causal empiricism or practitioner experience or case studies).
- That is, while the case for RCTs as superior estimates of the response surface over a specific action was powerful and correct—the rest of the logframe or theory of change or causal pathway from RCT to impact was never actually made.

	RCT 2.0 –"learned from experience" and conceded on all key points and hence changed the practice of doing RCTs from "independent impact evaluation" to more MeE (Monitoring experiential learning, and impact Evaluation) approaches
Topics important for development	Still stuck on this point.
Organizational capability and learning	"Crawl the design space"—worth with local partners in the *design* phase and build implementation feedback loops to build towards effective interventions and capability simultaneously. This gives up on the notion of "independent" evaluation as now the "intervenors" and "evaluators" are the same people. This gives up on the priority of "impact evaluation" (from outputs to outcomes involving causal claims about impact on beneficiaries) to "efficacy"—helping organizations get from inputs to activities to outputs.
Political economy	Working with governments on the generation and use of "evidence" as a broader issue than just doing an RCT. Conceded on the "special" role of RCTs.
Construct validity	Completely conceded. The specifics of program design have to be worked out instance by instance in an iterative way.
External validity	Completely conceded. Evaluation costs have to be amortized over the specific project as there cannot be claims of generalizable knowledge.

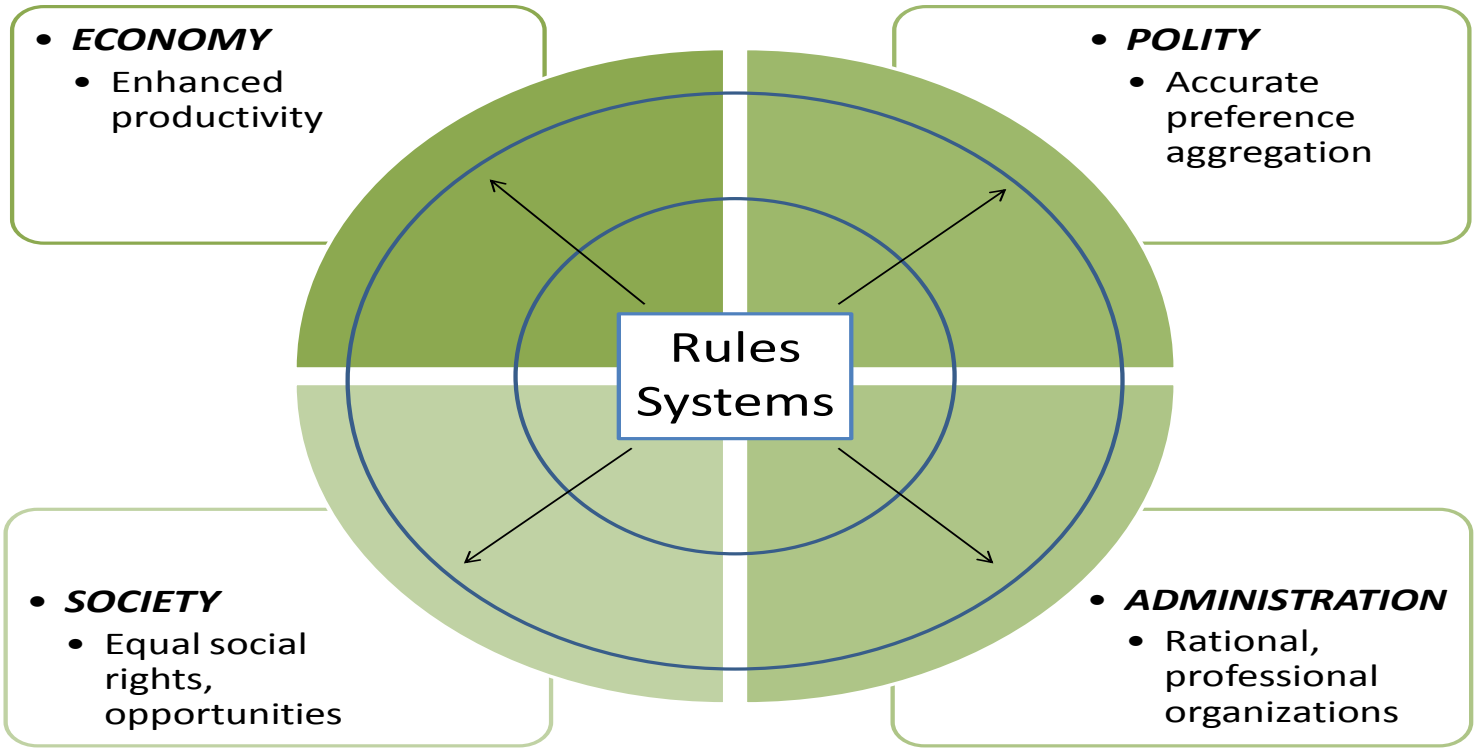
The rest of the slides are making in detail and with examples the five arguments—each of the five arguments would take a 90 minute seminar (or more) to do it justice.

Can RCTs add useful information on the big questions about development—those most consequential for human well-being?

No.

“National Development” is a four-fold transformation of ‘rules-systems’ and social capabilities (with complex interacting pieces)

Figure 1: Development as a four-fold modernization process



Source: Pritchett 2009 “Is India a Flailing State?” http://dash.harvard.edu/handle/1/4449106

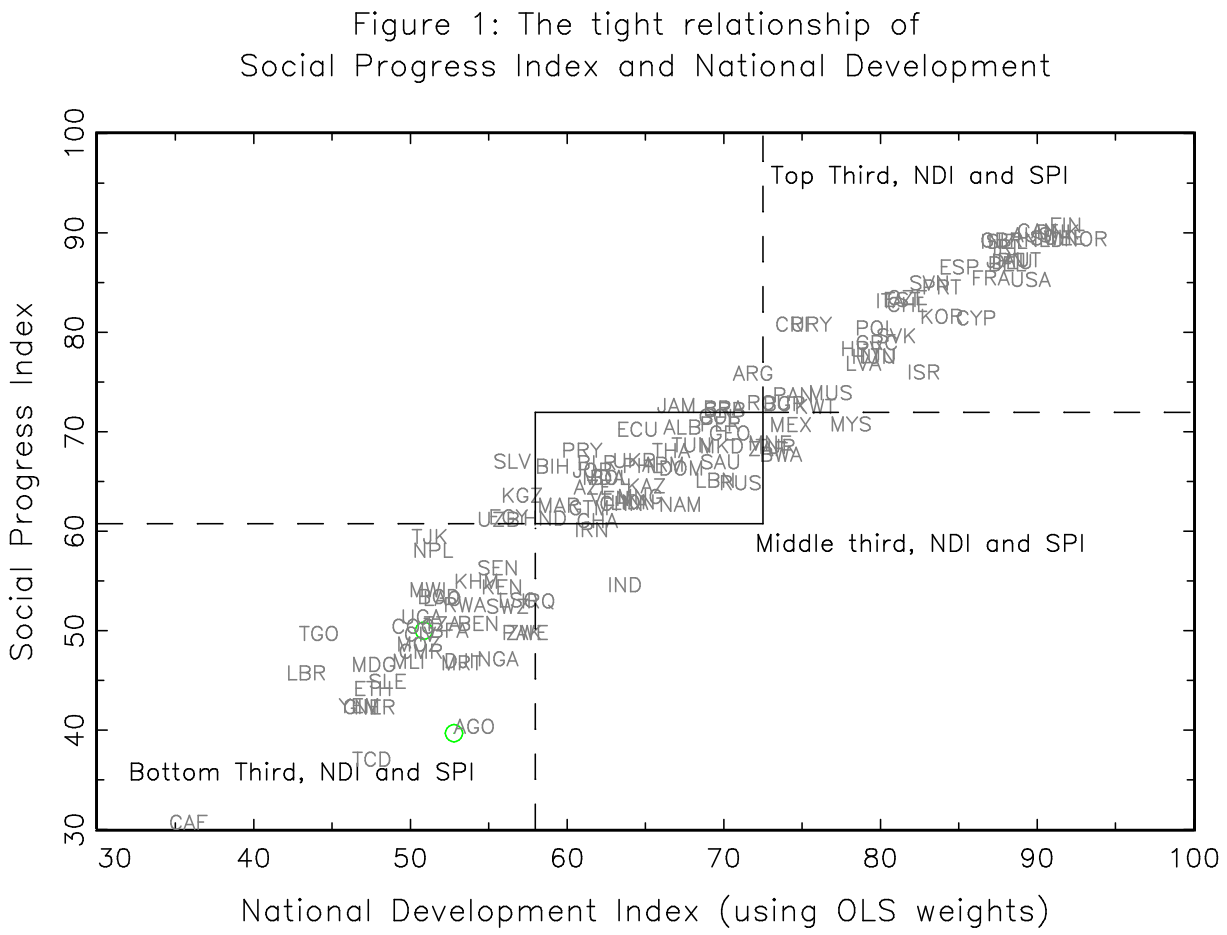
Two big, related but distinct, definitions of “development”

- *National development* is *ontologically* a social process and is an inter-related set of transformations of *group* dynamics—”the market” is a social phenomena, “institutions” are a social phenomena, “organizations” are a social phenomena—not reducible to aggregations of individuals
- *Human Development* are measures of well-being that are *ontologically* individualized (and for which aggregation is possible, but secondary)

The *normative* objective is *human development* (by some metric) the *instrumental means* to that objective is *national development.*

Frogs are frogs and ***development*** is about becoming *more* of the thing you ontologically are, not changing your nature.

Turns out, national development and individual indicators are tightly related



Source: [Turns out Development Does Bring Development]

Anything than a super high R2 of “national development” and any measure of human well being would be pretty unusual

- Economy—is the available resources to devote to problems.
- Responsive polity—is whether the state is responsive to problems articulated by citizens
- Administrative capability—is whether organizations can accomplish goals.
- The relationship of this to any truly universal and high priority human need *has to be* very high.
- Policies, programs, projects and their design and the creation and application of knowledge to problems is completely and total *endogenous* to national development. We should expect the “exogenous” component to be small and identifying and solving pressing problems is what high functioning systems *do.*

As a subset of national development, Lucas was right: growth, growth, growth.

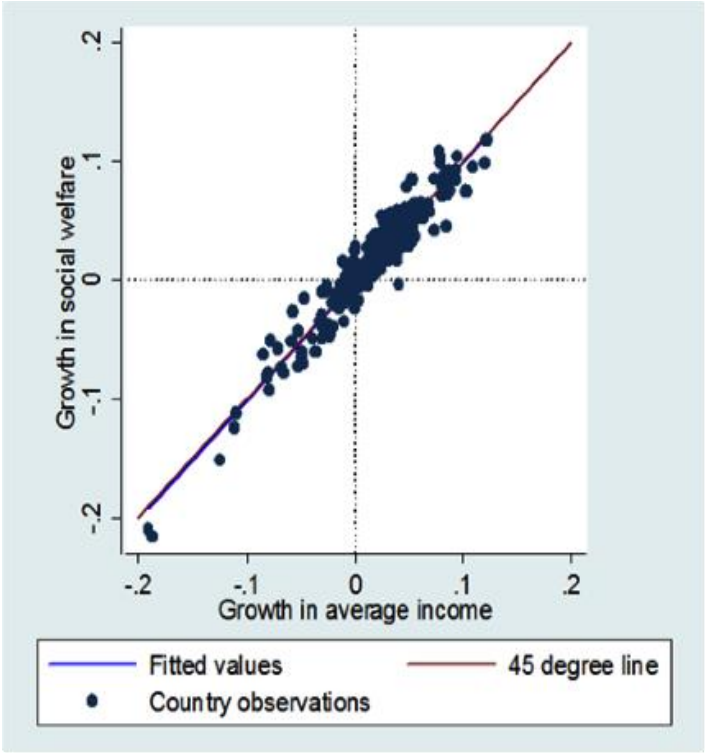
Is there some action a government of India could take that would lead the Indian economy to grow like Indonesia's or Egypt's? If so, what, exactly? If not, what is it about the "nature of India" that makes it so? ***The consequences for human welfare involved in questions like these are simply staggering****: Once one starts to think about them, it is hard to think about anything else.*

(Lucas 1988)

Turns out, there was something that could be done, it was done in the early 1990s (whatever it was) and the net NPV contribution of India's growth accelerations (relative to BAU growth) has been 3.5 *trillion* dollars by 2010.

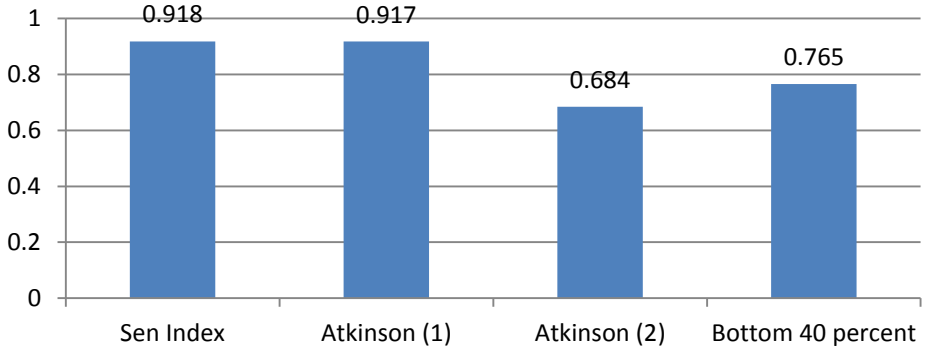
Empirically the growth of incomes of "the poor" (however defined) or inclusive growth is pretty much the growth of average incomes (plus minus a bit)

Sen index



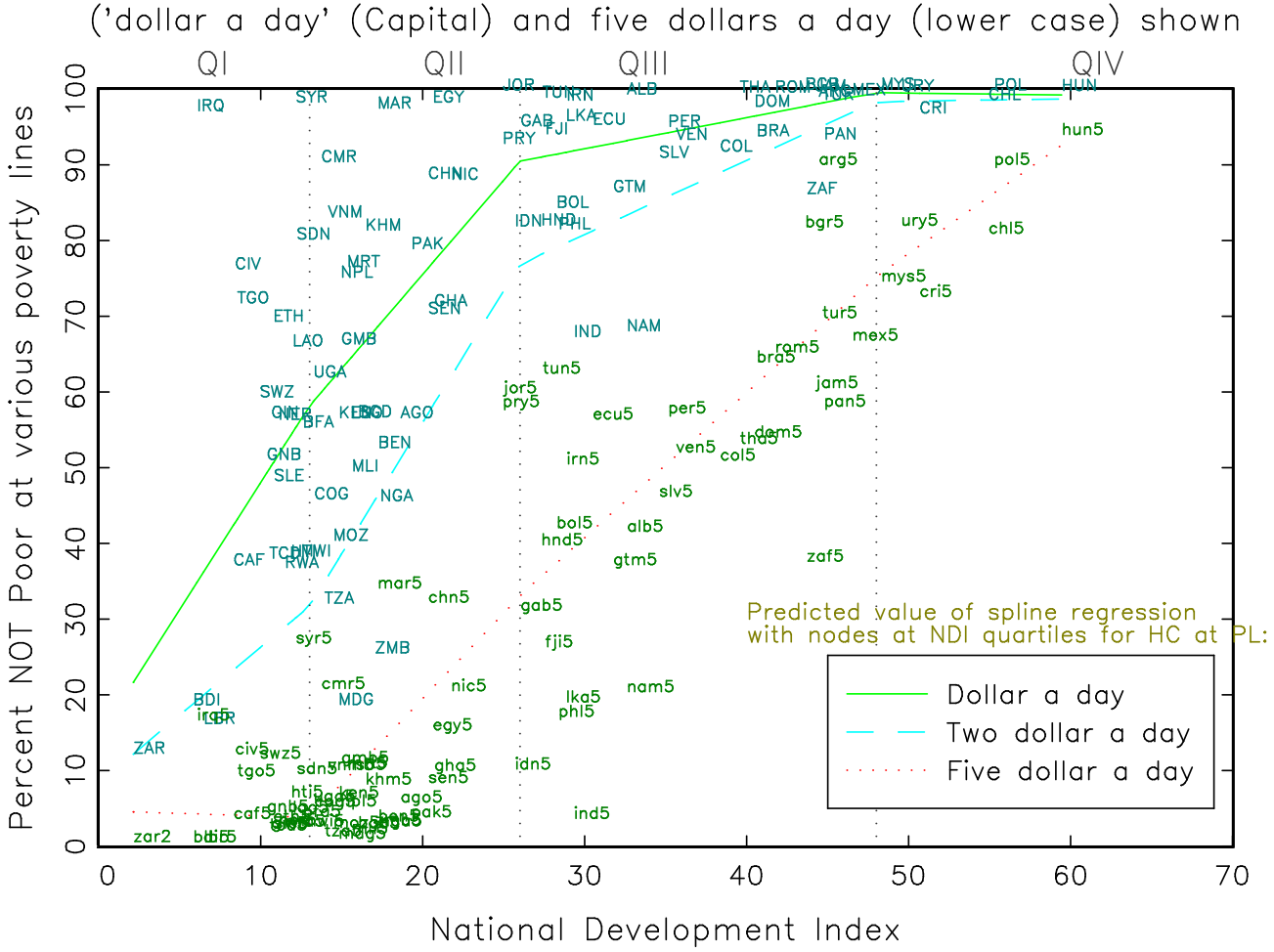
http://www.voxeu.org/article/growth-inequality-and-social-welfare

Variance in performance on SWF across countries due to average growth in income

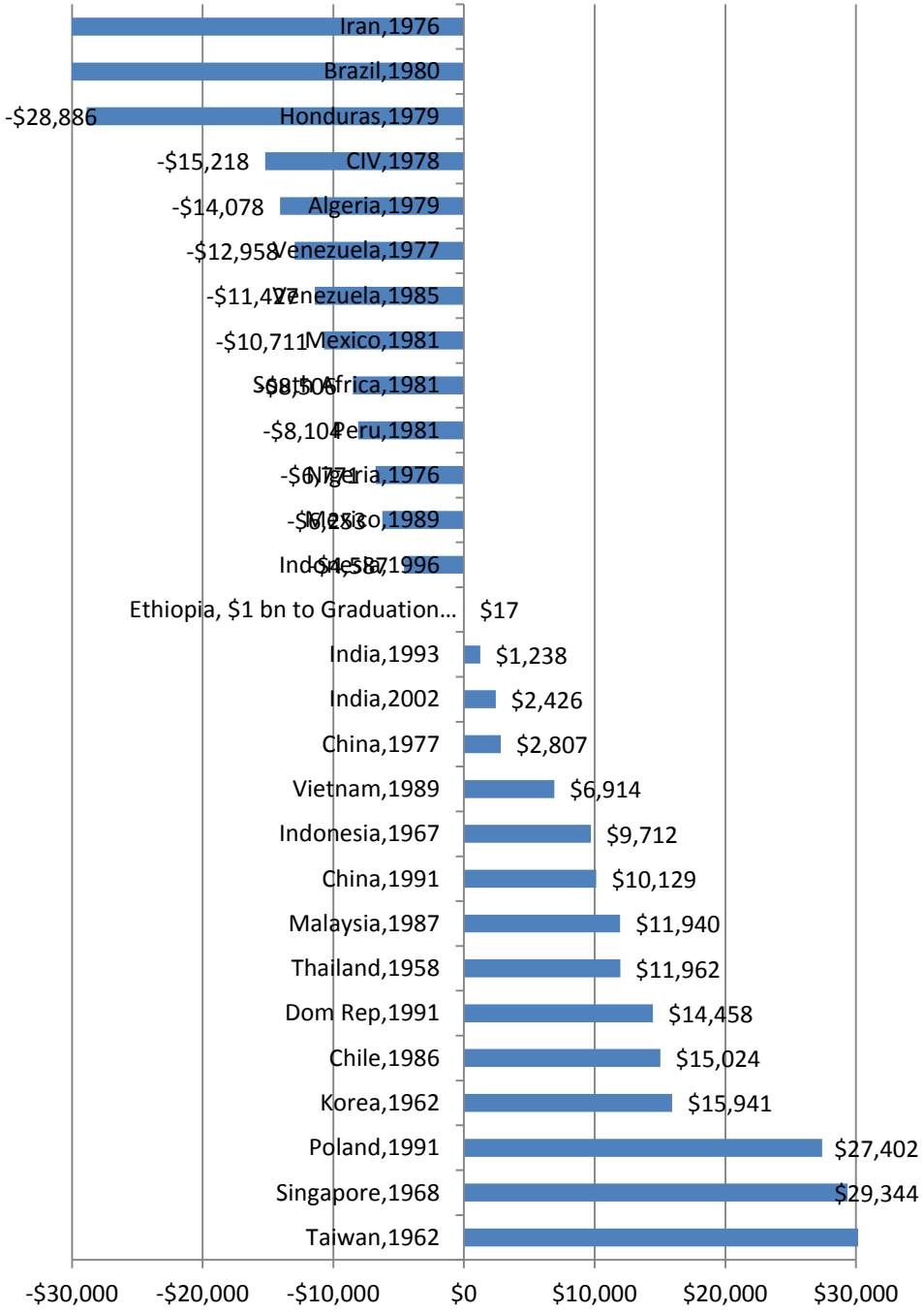


Source: Dollar, Kleineberg, Kraay (2014), table 2, panel A (all spells)

National development is *strongly* related to poverty—and the relationship is linear all the way up national development for 5 dollar a day poverty



Source: [Pritchett and Kenny 2013]



The gains (and losses) in NPV per person in the economy from big growth accelerations (or decelerations) are *orders of magnitude* larger than the per person NPV of the best available development programs.

The recently reported NPV per person gains from the Graduation approach done by BRAC in five countries *for the targeted households only* are on the order of $1700 per person (off spending of $1000 per person).

Say Ethiopia spends a US$ 1 billion to benefit 1 million people for benefits of 1.7 billion. In a country of 100 million people this is U$17 per person in NPV. This is two *orders of magnitude* (100 times) less than India's 2002 growth acceleration, three orders of magnitude less than Brazil's 1980 slow down.

Source: Estimates of gains/losses adapted Pritchett et al 2016.

The 20 interventions on which there have been sufficient rigorous impact evaluations to make comparisons about generalizability (Vivalt 2014)	Done more in more developed countries than less developed economies? (e.g. Denmark more than Malawi)	Done more today than historically in developed economies? (e.g. Denmark today more than in 1870)	Done more in rapidly progressing countries than stagnant countries? (e.g. More in Korea than Ghana)	Country's progress accelerates/ decelerates when a country does more/less of it? (e.g. More in China post 1978 than pre 1978)
Conditional cash transfers				
Deworming				
Improved stoves				
Treated bed nets				
Microfinance				
Safe water storage				
Scholarships				
School meals				
Unconditional cash transfers				
Water treatment				
Contract teachers				
Financial literacy training				
HIV education				
Irrigation				
Micro health insurance				
Micro nutrient supplementation				
Mobile phone based reminders				

What the RCT agenda has mostly been working on (by availability to do a review) doesn't pass a simple four part "smell test" for being important to development

Both of these are important, but they are not the same agenda

Promoting national development

- {what is here}

Mitigating the consequences of a lack of development on human well being

- {what is here}

One controversial claim

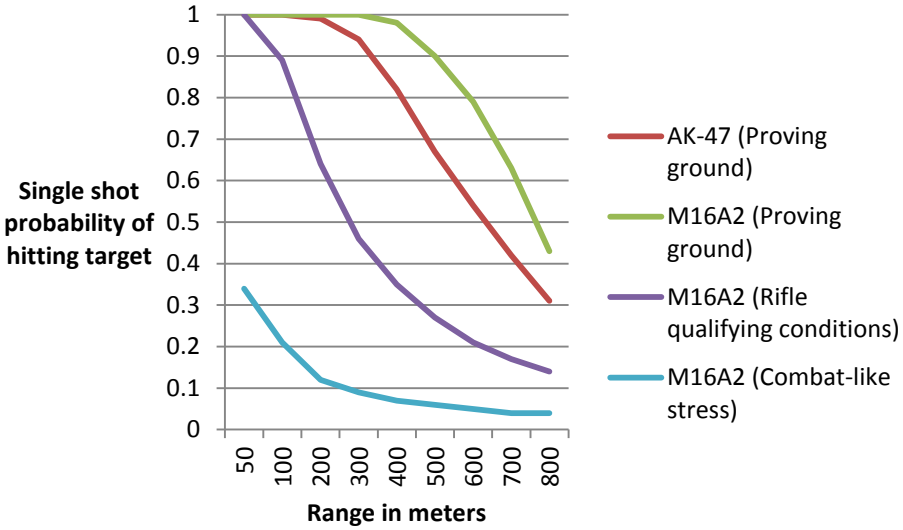
RCTs and its movement have mostly been part of a systematic effort of the rich countries to "define development down" and move the development agenda away from the interests and concerns of the governments and citizens of the "South" towards a very restrictive, low-bar, foreign assistance agenda, capable of generating political support in the West and hence obsessed with narrow attribution versus success. RCTs only make sense as a important element of a low-bar or "kinky" development approach.

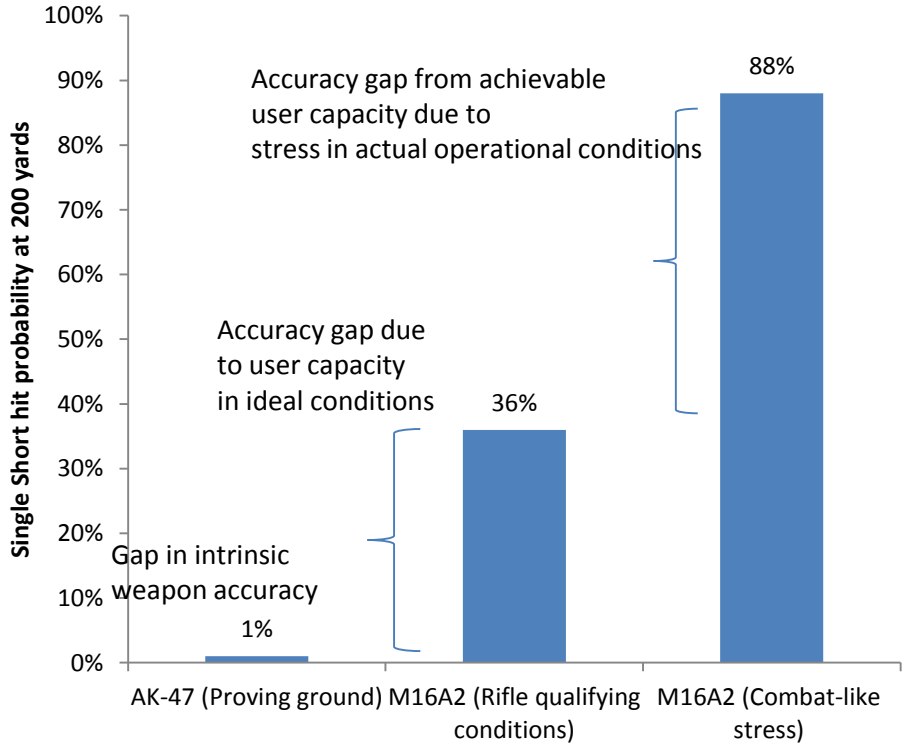
"Innovations" like the Socket ball are illustrations of the delusions of kinky development

Is expanding the stock of knowledge about the response surface a key constraint to organizational effectiveness?

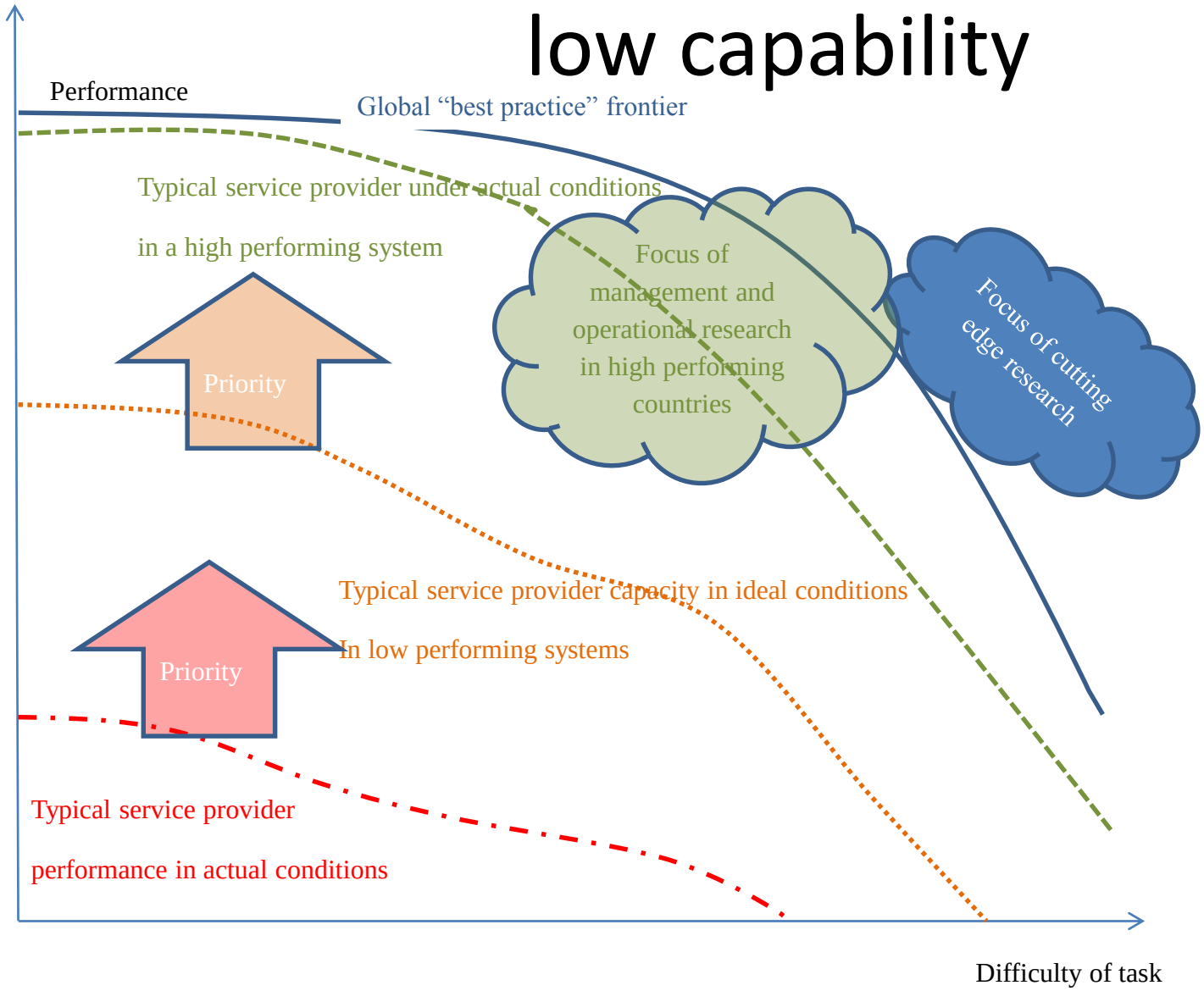
No. In most developing country settings efficacy is limited by organizational capability, not lack of knowledge about the response surface.

The AK47 is the less accurate weapon than the M16—why is it so popular?





Getting the basics done (at all) is a huge and pressing issue because of low capability

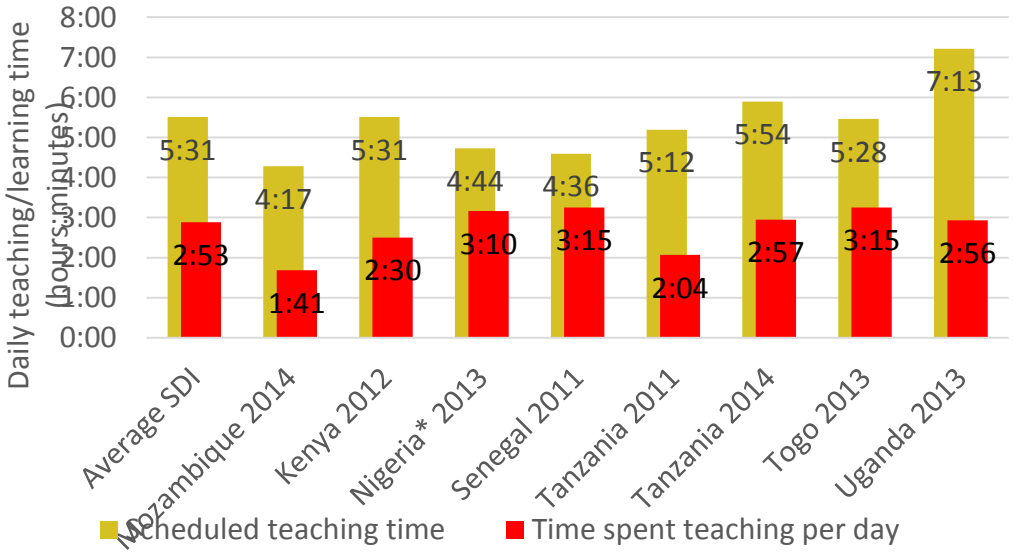


The *capacity* of individuals is often very low…and effort is low

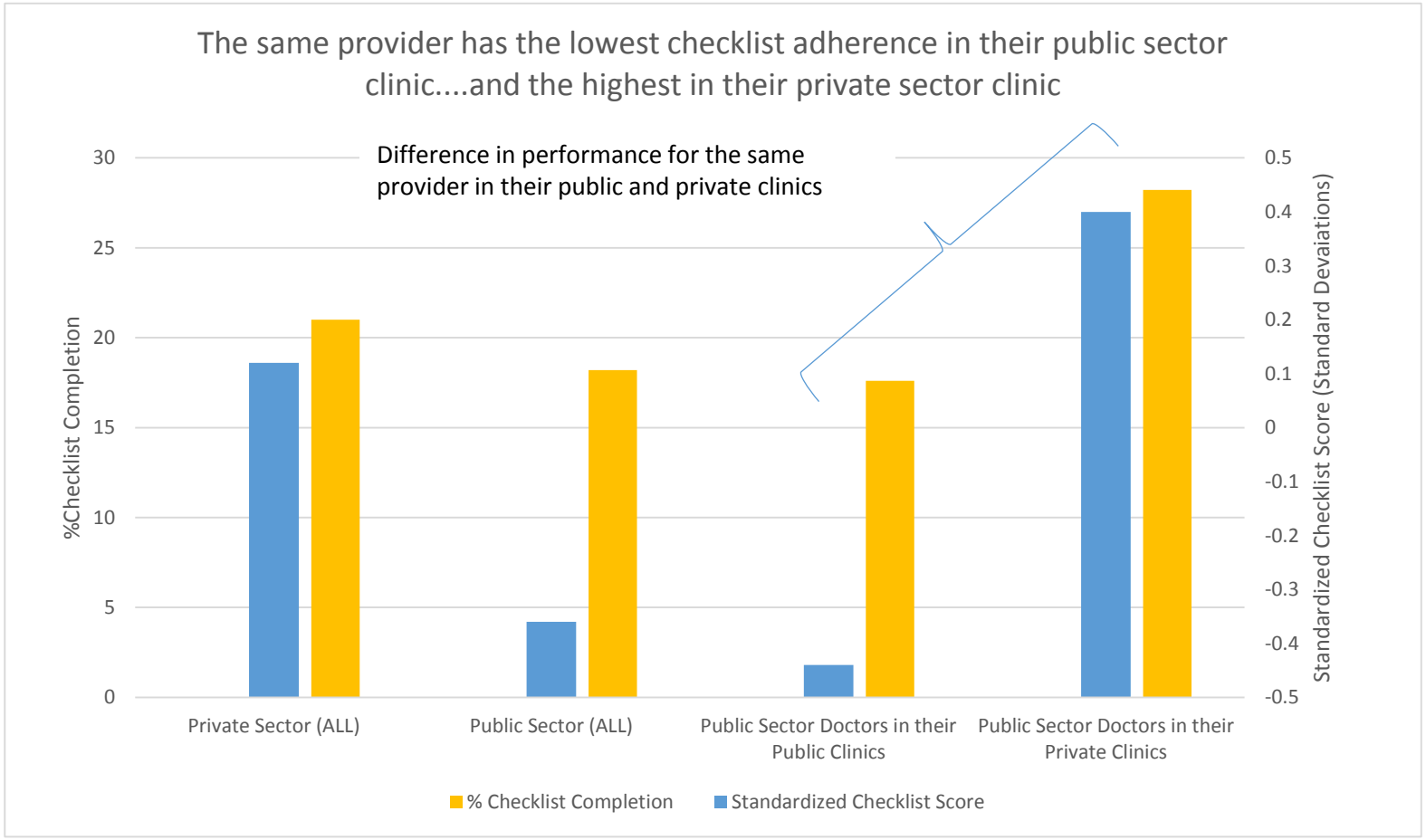
SDI Findings: Teacher skills in public schools

	Average SDI	Kenya 2012	Mozambique 2014	Nigeria* 2013	Tanzania 2014	Togo 2013	Uganda 2013
Minimum knowledge (At least 80% in language and mathematics)	12.7	34.8	0.3	2.4	15.6	0.9	10.1
Average test score (language, mathematics, and pedagogy); "Full marks" is 100.	42.0	55.6	26.9	30.5	46.6	33.9	43.3

* Nigeria is 4 States



The *worst* and *best* medical care in rural Madhya Pradesh came from the *same* people—many times the problem is not the *capacity of individuals* it is the capability of organizations

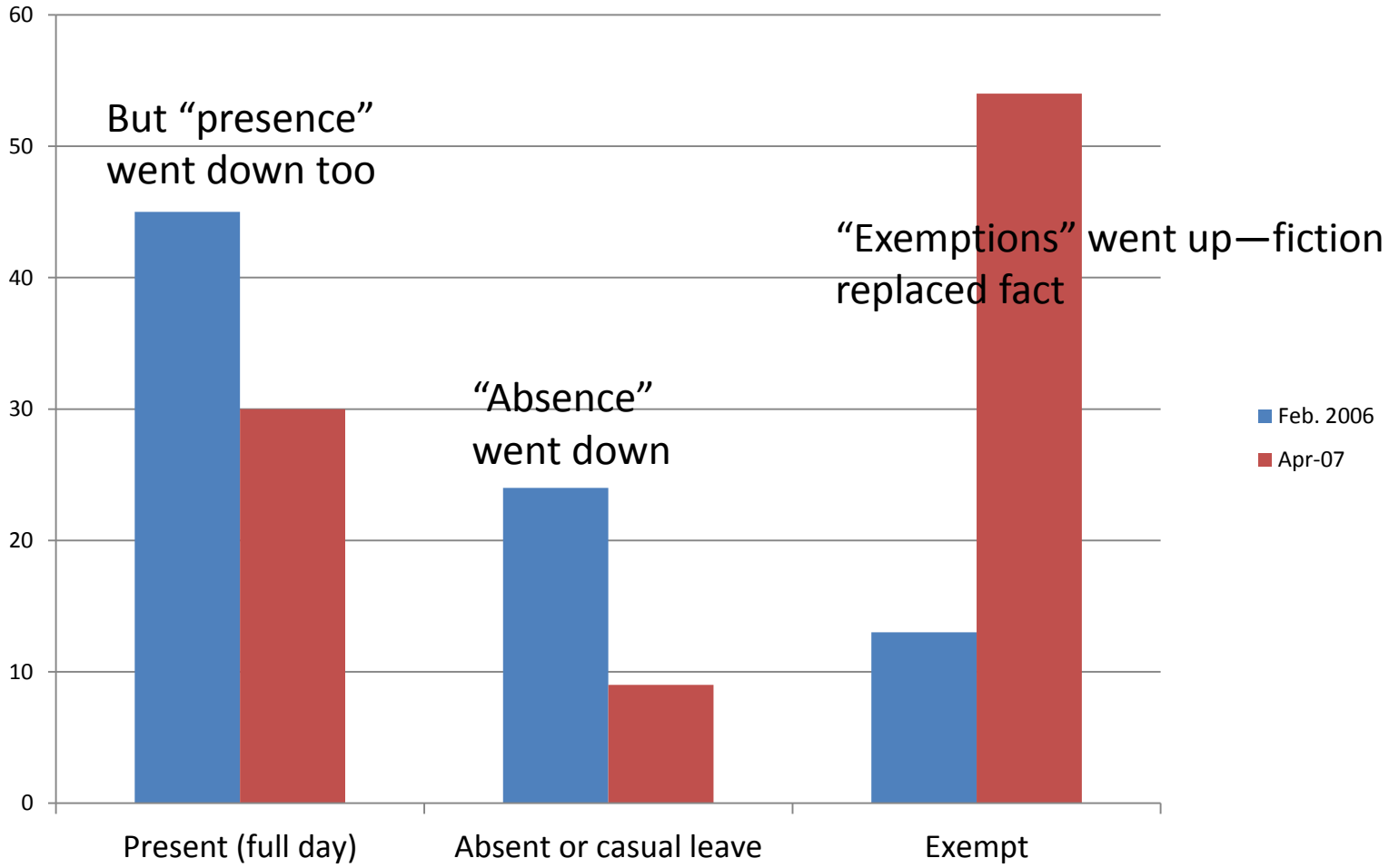


Source: Das et al forthcoming

What has been learned from lots of RCTs is that organizations cannot/will do what they are being asked to do—you cannot even do the "treatment"

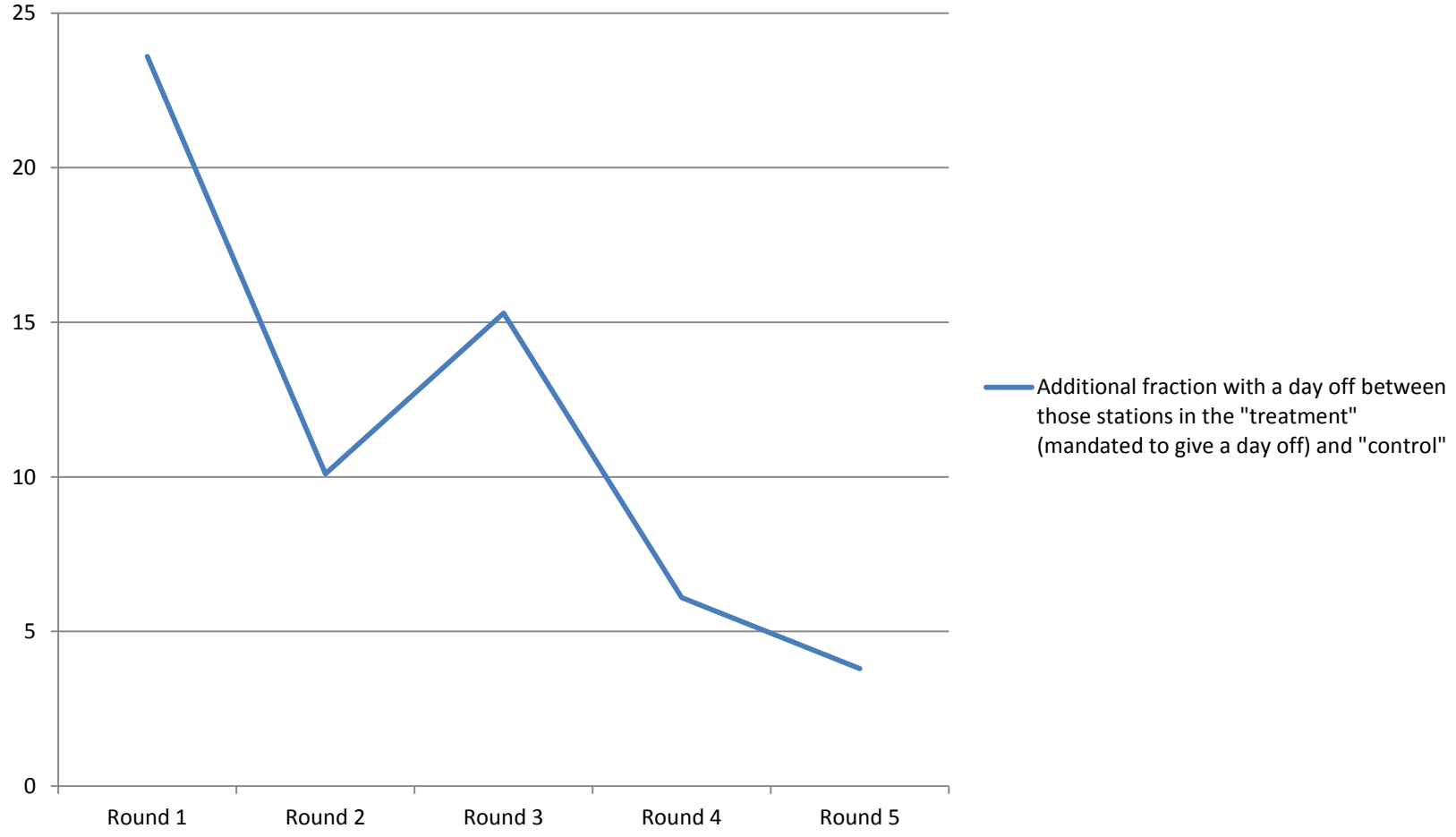
- Attendance of nurse-midwives in Rajasthan
- Attendance at health clinics in Karnataka
- Policing experiment in Rajasthan

During the course of the field experiment to motivate nurses
to attend their clinics in Rajasthan… they found they could not implement the "treatment"



Source: Banerjee et al 2008, **Putting Band-aids on a Corpse**, adapted from Figure 3

An experimental evaluation of attempts to improve police performance found that in important respects the police hierarchy did not control the routine police scheduling behavior

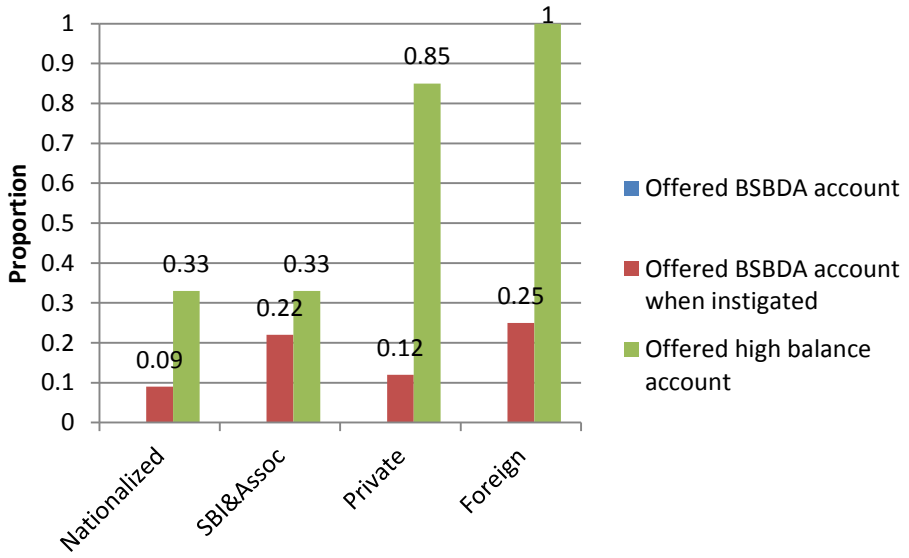


Source: Banerjee, Chattopadhyay, Duflo, Keniston, Singh 2012

Pushing the action into the private sector via regulatory mandates doesn't change the need for capability…

Features	Details
RBI policies	The BSBDA was subject to RBI policies on KYC/AML for opening of normal bank accounts.
Restrictions	The BSBDA would be considered a normal banking service available to all: banks were to impose no restrictions such as age, income, and amount criteria. However, the holders of a BSBDA would not be eligible for opening any other savings account in the same bank; they could have other deposit accounts, such as term/fixed deposits and recurring deposits.
Minimum facilities offered*	No minimum balance requirement. No initial deposit required for opening a BSBDA. The services available include deposit and withdrawal of cash at bank branch as well as ATM usage; receipt/credit of money through electronic payment channels; deposit/collection of cheques drawn by government agencies and departments. No limit on the number of deposits in a month. Account holders will be allowed a maximum of four withdrawals per month, including ATM withdrawals. Passbook and an ATM card or ATM-debit card will be provided free of charge. Chequebook would not be provided as a minimum facility. The minimum facilities would be provided free of charge, and no charge would be levied for non-operation or for activation of inoperative BSBDAs. Beyond the minimum facilities, banks could set their own pricing structure for fees and services.

*Source: Reserve Bank of India (2012a, 2012b)



(and no, one bar is not missing—offer of the legally mandated low cost account were literally zero in all types of banks.)

Source: Mowl et al 2014

And India has *above average* state capability on the standard measures…

	Very negative	Slow negative	Slow positive (with years to high capability)	Rapid	
Strong capability (SC>6.5)		BHR, BHS, BRN	CHL(0), SGP(0), KOR(0), QAT(0)	ARE(0)	
	8	0	3	4	1
Middle capability (4<SC<6.5)	MDA, GUY, IRN, PHL, LKA, MNG, ZAF, MAR, THA, NAM, TTO, ARG, CRI	PER, EGY, CHN, MEX, LBN, VNM, BRA, INDIA, JAM, SUR, PAN, CUB, TUN, JOR, OMN, MYS, KWT, ISR	KAZ(10820), GHA(4632), UKR(1216), ARM(1062), RUS(231), BWA(102), IDN(68), COL(56), TUR(55), DZA(55), ALB(42), SAU(28), URY(10), HRV(1)		
	45	13	18	14	0
Weak capability (2.5<SC<4)	GIN, VEN, MDG, LBY, PNG, KEN, NIC, GTM, SYR, DOM, PRY, SEN, GMB, BLR	MLI, CMR, MOZ, BFA, HND, ECU, BOL, PAK, MWI, GAB, AZE, SLV	UGA(6001), AGO(2738), TZA(371), BGD(244), ETH(103), ZMB(96)		
	32	14	12	6	0
Very weak capability (SC<2.5)	YEM, ZWE, CIV	SOM, HTI, PRK, NGA, COG, TGO, MMR	SDN(7270), SLE(333), ZAR(230), IRQ(92)	NER(66), GNB(61), LBR(33)	
	17	3	7	4	3
	102	30	40	28	4

The grand aggregated estimate of RCT findings is that the average government implemented RCT has zero impact (.199 less .163 in impact over std dev.)

Table 5: Regression of Effect Size on Study Characteristics

	(1) Effect size b/se	(2) Effect size b/se	(3) Effect size b/se	(4) Effect size b/se	(5) Effect size b/se
Number of observations (100,000s)	-0.018*** (0.01)			-0.019*** (0.00)	-0.015** (0.01)
Government-implemented		-0.163*** (0.06)			-0.150** (0.06)
Academic/NGO-implemented		-0.070 (0.04)			-0.075 (0.05)
RCT			0.049 (0.04)		
East Asia				-0.015 (0.03)	
Latin America				-0.006 (0.04)	
Middle East/North Africa				0.284** (0.11)	
South Asia				0.009 (0.04)	
Constant	0.120*** (0.00)	0.199*** (0.04)	0.080*** (0.03)	0.114*** (0.02)	0.201*** (0.04)
Observations	534	634	634	534	534
R^2	0.19	0.23	0.22	0.22	0.20

You cannot beat a turtle into moving

The head has to come out for the body to move

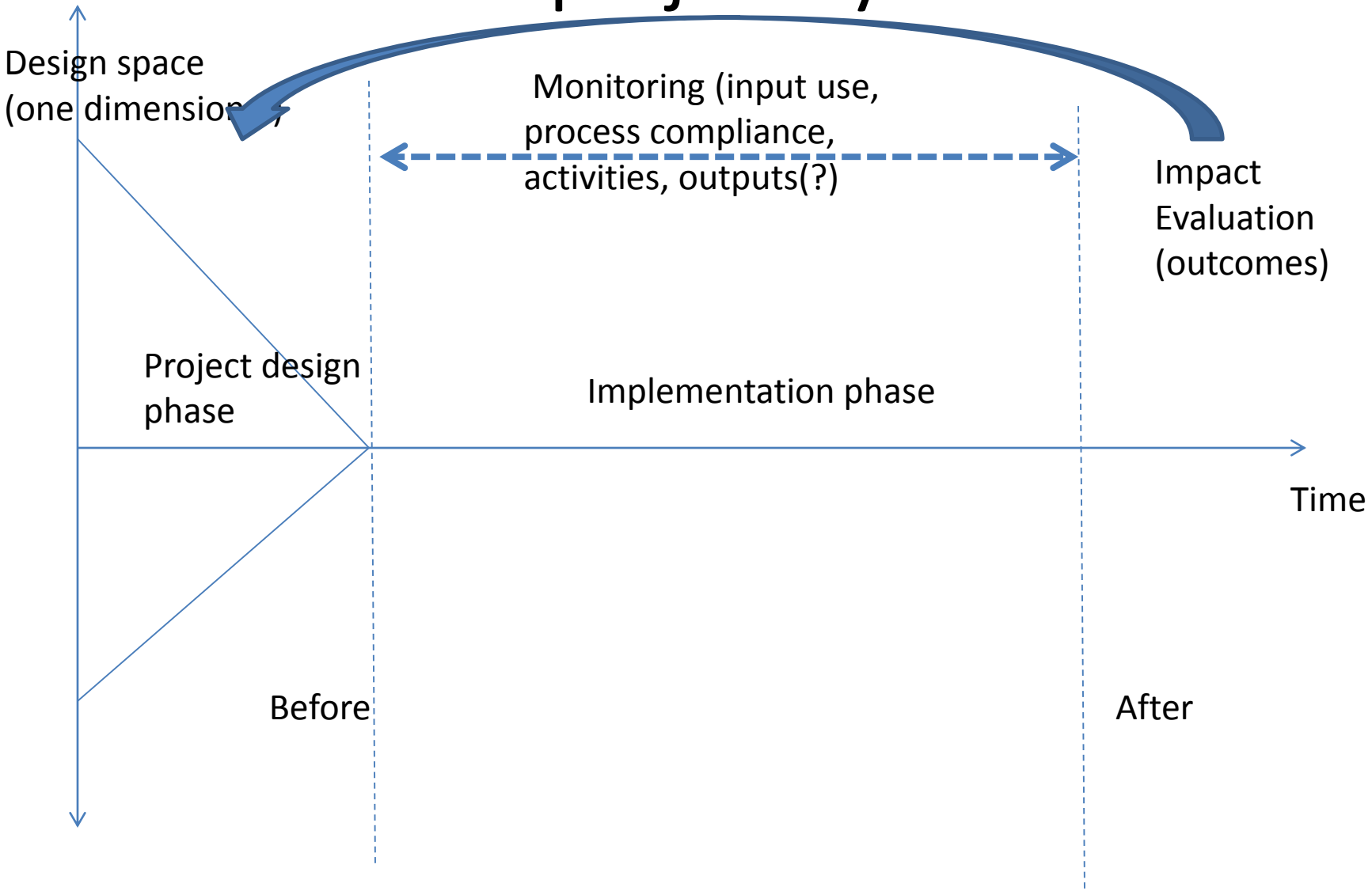
Organizations can survive external attack…by not moving

Trying to make things better by pushing “policy” (mapping) without tackling the underlying determinants of organization capability

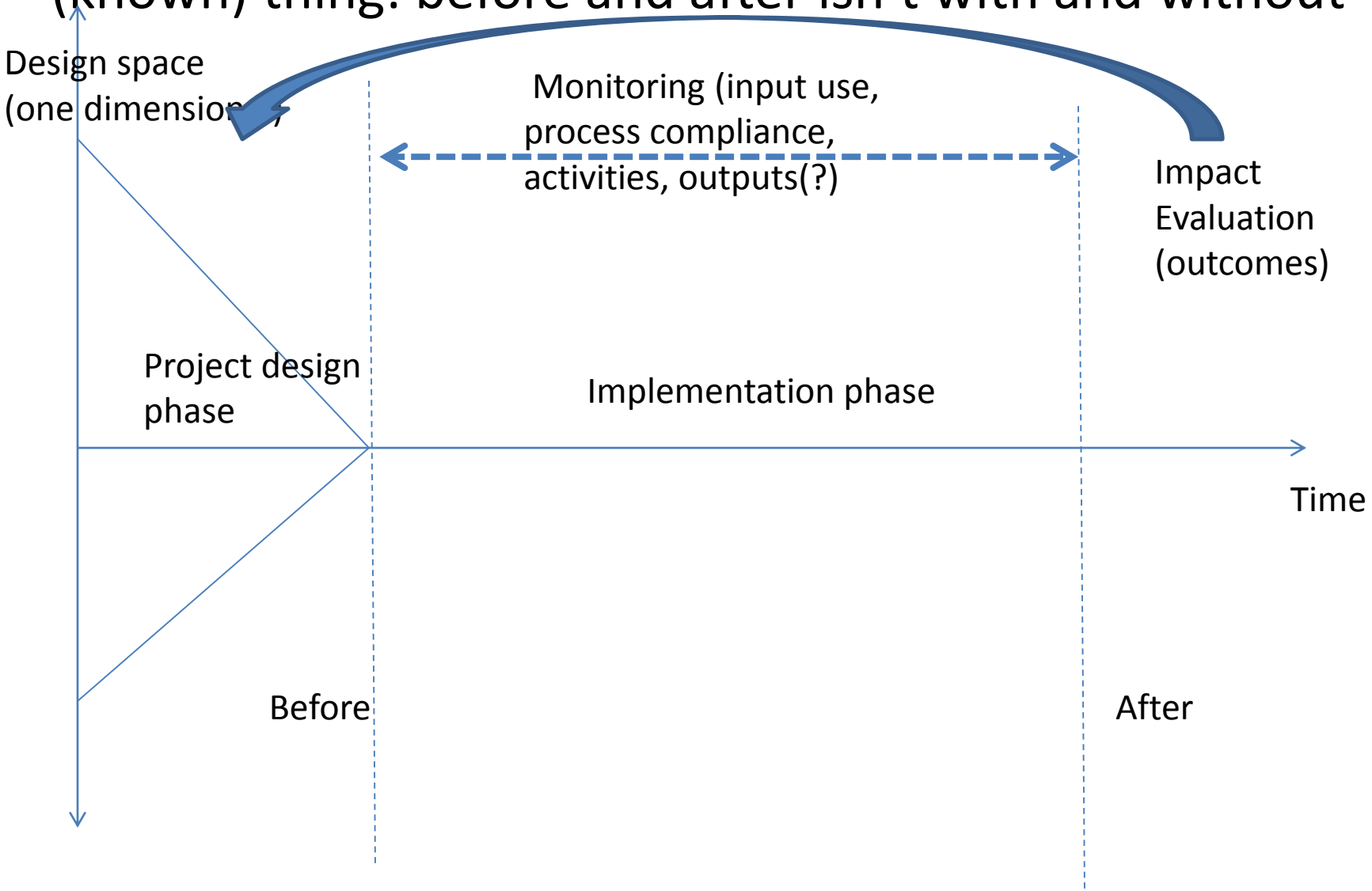
TABLE 3: PROJECT CONTENT CATEGORIZED ACCORDING TO WHETHER IT TARGETED REGULATIVE, CULTURAL-COGNITIVE OR NORMATIVE ELEMENTS

Regulative	Normative	Cutural-cognitive
Elements targeted at shaping behaviour through the threat of sanction. Behaviour constrained through extrinsic means.	Elements targeted at making desired behaviours socially appropriate and acceptable, causing agents to intrinsically "decide" in favour of specific actions.	Targeted at helping individuals structure and interpret the information that they receive in order to bias them towards specific choices, regardless of the incentives created by regulative and normative mechanisms.
E.g.Actvities focused on strenthening laws, shaming practices, regulatory bodies, control systems etc	E.g. activities targeting cultural beliefs, profession norms, ethical values and standards considered widely appropriate	E.g. (formal) info, education or guidance towards compliance with intl standards. (informal) Attempts to increase cognitive capacities (understanding, ability to interpret and apply practices).
92%	**3%**	**5%**

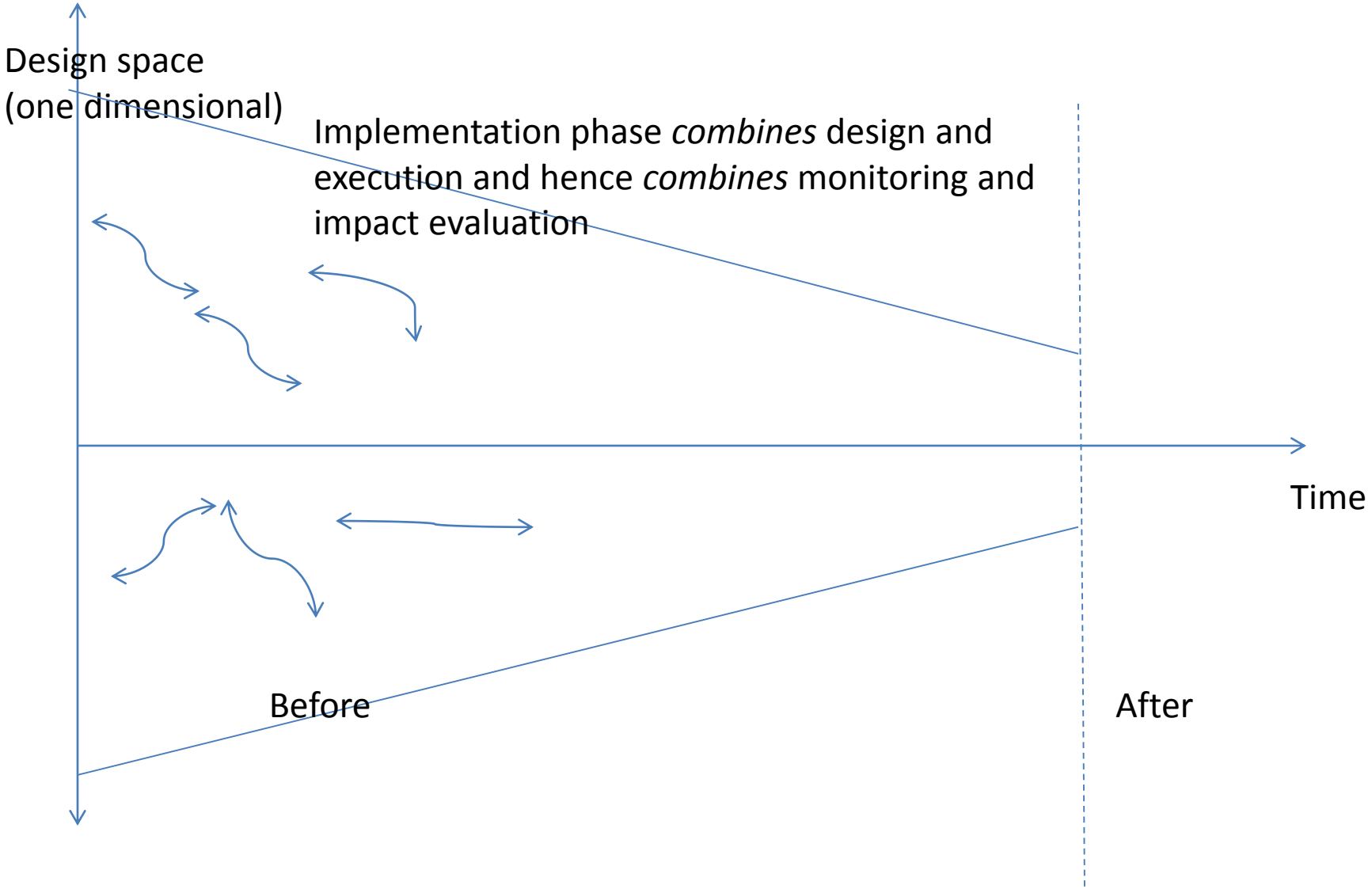
The typical (pre-adaptive) approach to the project cycle



To this model of learning in the project cycle the RCT movement brought was mono-maniacal about one (known) thing: before and after isn't with and without



What is radically different in “adaptive” approaches (like PDIA) versus the emphasis on “rigorous evidence”




1 EXPLORE
2 SYNTHESISE
3 DESIGN
4 TRY IDEAS
5
Monitoring experiential Evaluation (MeE)
6 PILOT
7 EVALUATE
8 SHARE (Ideas Box)
9 SCALE OUT
Monitoring experiential Evaluation (MeE)

Rapid feedback loops beat all hell out of rigorous when response surface is rugged—particularly with respect to robustness of conclusions

Table 5: Learning results varied across ruggedness of the fitness space

(1) Ruggedness parameter	(2) Ruggedness (absolute difference)	(3) Gain CDS over RCT (ratio to max less average)	(4) Percent excess of RCT over CDS standard deviation
.25	.020	.319	1.04
.5	.042	.445	1.19
1 (base case)	.074	.489	1.64
2	.094	.461	2.36
4	.103	.412	4.25

Source: Nadel and Pritchett 2016

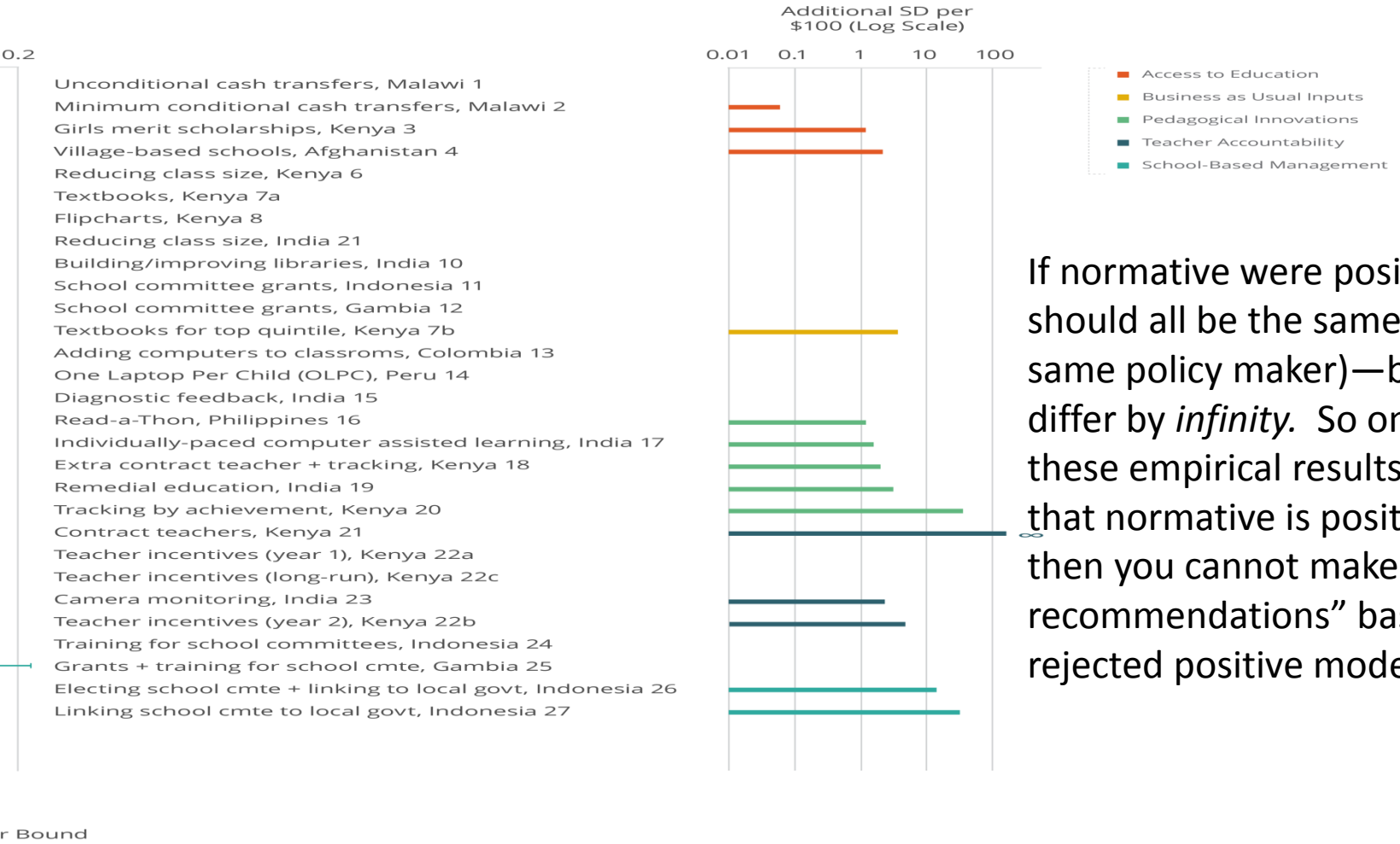
Does the generation of RCT knowledge significantly change the scope of what is politically feasible?

No.

Normative as Positive cannot be used if you just rejected the normative

- In 1997 Filmer and Pritchett wrote a paper saying:
  - If the "policymaker" were applying resources to maximize learning then the implication is that marginal product per dollar should be equalized across all uses.
  - The evidence rejects this hypothesis by order(s) of magnitude as measured marginal product per dollar varies massively—and systematically—within and across countries.
  - Conclusion: as a *positive* descriptive model of policy maker behavior we cannot assume she/he *not* maximizing learning per dollar of expenditure but is pursuing some other objective function.
  - Therefore examining marginal product per dollar and making "policy" recommendations based on that is a silly game and research needs to focus on the positive political economy of learning.

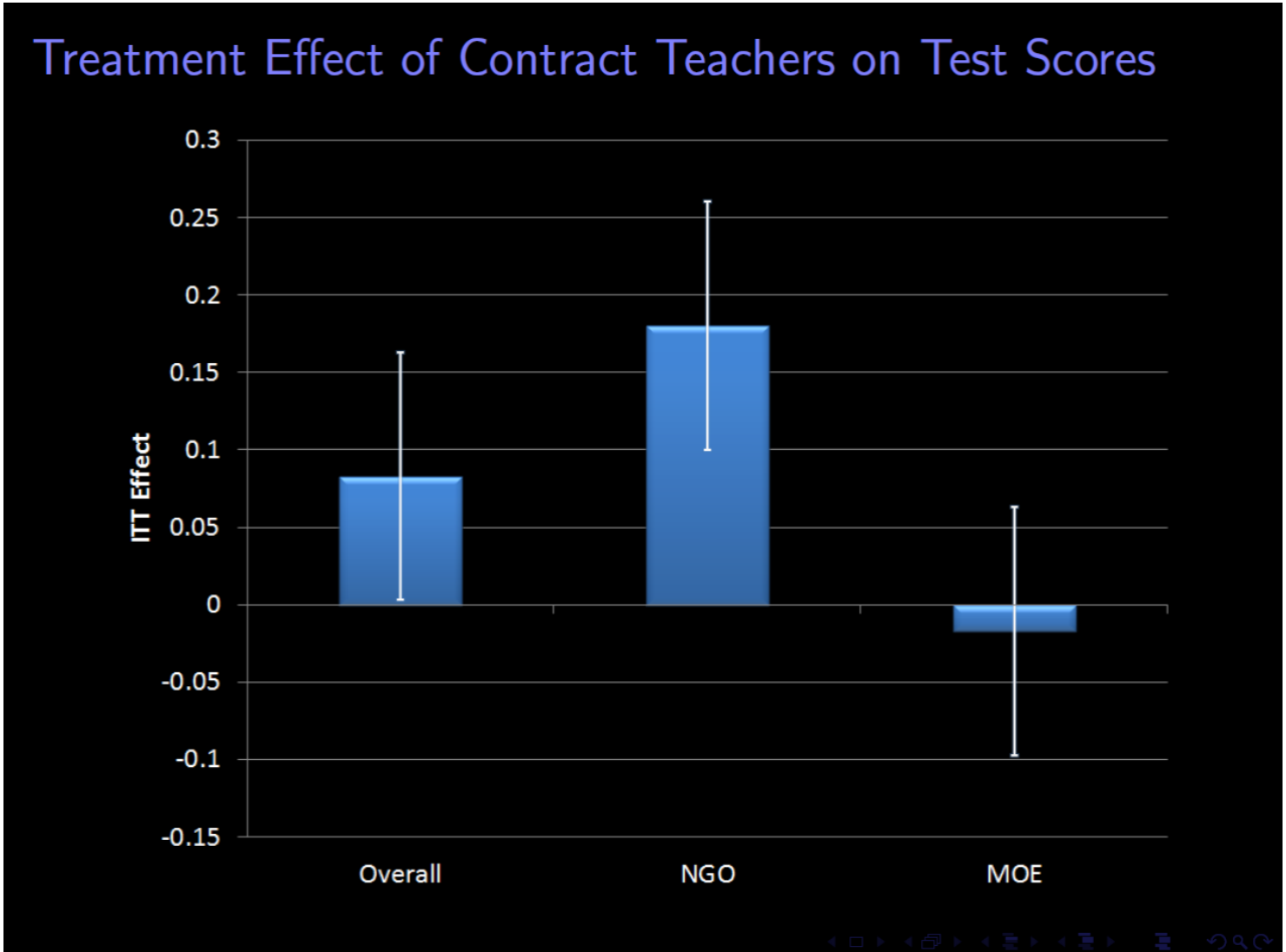
This is what two decades of intellectual regress looks like…two decades of research to make the same point as was made in 1997--*less* well and with *less* sophistication about politics of adoption



If normative were positive these should all be the same (for the same policy maker)—but they differ by *infinity.* So on one level these empirical results reject that normative is positive. But then you cannot make "policy recommendations" based on a rejected positive model.

Exact same program except for one design feature, who implements

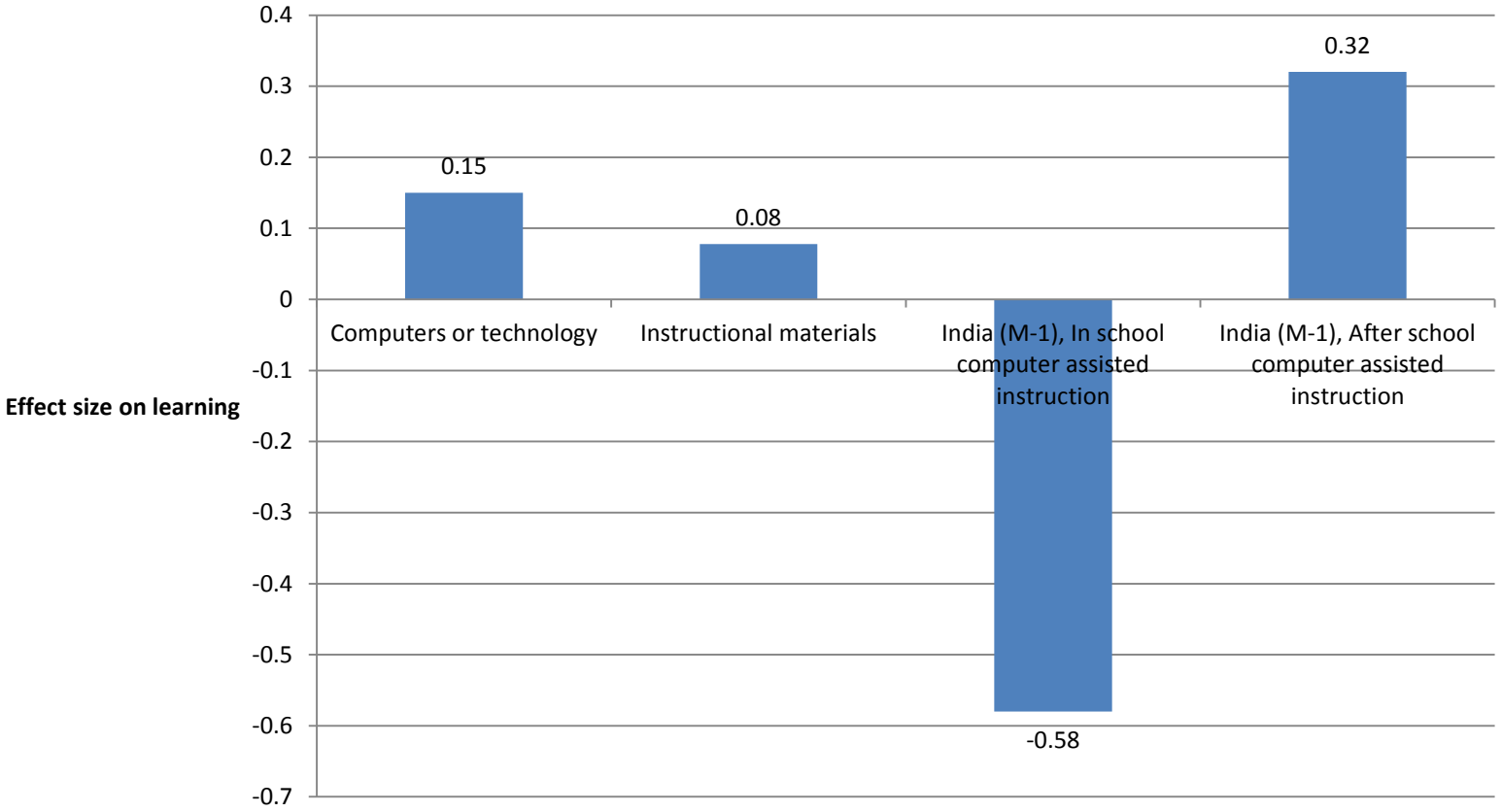
Source:
Bold et al 2013



Do findings from RCTs about the impact of a specific project design lead to knowledge with construct validity?

No. Response surfaces are typically rugged over a high dimensional (and unknown) design space

Existing "systematic reviews" that compare across classes of projects produce gibberish in domains with rugged response surfaces as they lack construct validity



The variation across studies is in fact massive—and mostly appears to be construct validity not external validity

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.

Rapid feedback loops beat all hell out of rigorous when response surface is rugged—particularly with respect to robustness of conclusions

Table 5: Learning results varied across ruggedness of the fitness space

(1) Ruggedness parameter	(2) Ruggedness (absolute difference)	(3) Gain CDS over RCT (ratio to max less average)	(4) Percent excess of RCT over CDS standard deviation
.25	.020	.319	1.04
.5	.042	.445	1.19
1 (base case)	.074	.489	1.64
2	.094	.461	2.36
4	.103	.412	4.25

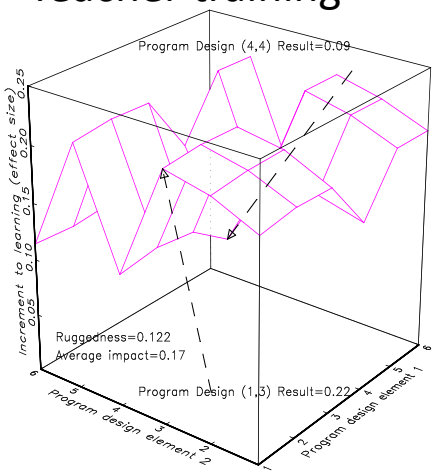
Source: Nadel and Pritchett 2016

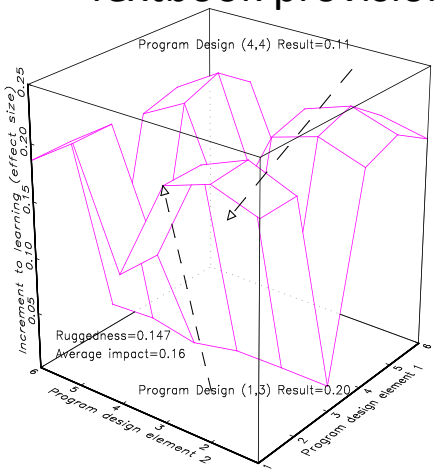
Suppose this is our world, two contexts (A and B), two classes of programs ("teacher training" and "textbook provision") with two design alternatives evaluated (1,3) and (4,4)

"Teacher training"

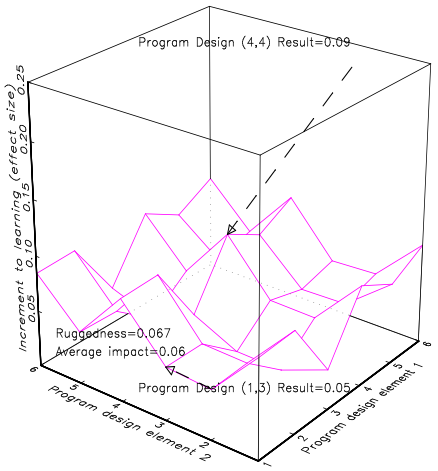
"Textbook provision"

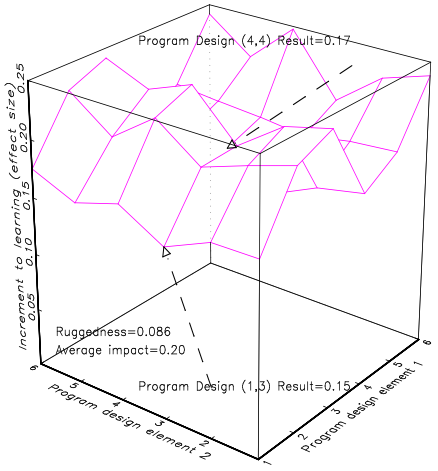
Context A





Context B





Impact sizes of different project designs in hypothetical world			
		Teacher Training	Textbook Provision
Context A	Avg	.17	.16
	(1,3)	.22	.20
	(4,4)	.09	.11
Context B	Avg	.05	.20
	(1,3)	.06	.15
	(4,4)	.09	.17

"Rigorous" evidence and get it *exactly wrong*…in many ways

Impact sizes of different project designs in hypothetical world			
		Teacher Training	Textbook Provision
Context A	Avg	.17	.16
	(1,3)	.22	.20
	(4,4)	.09	.11
Context B	Avg	.05	.20
	(1,3)	.06	.15
	(4,4)	.09	.17

Evidence base A: best project is Teacher Training design (1,3)

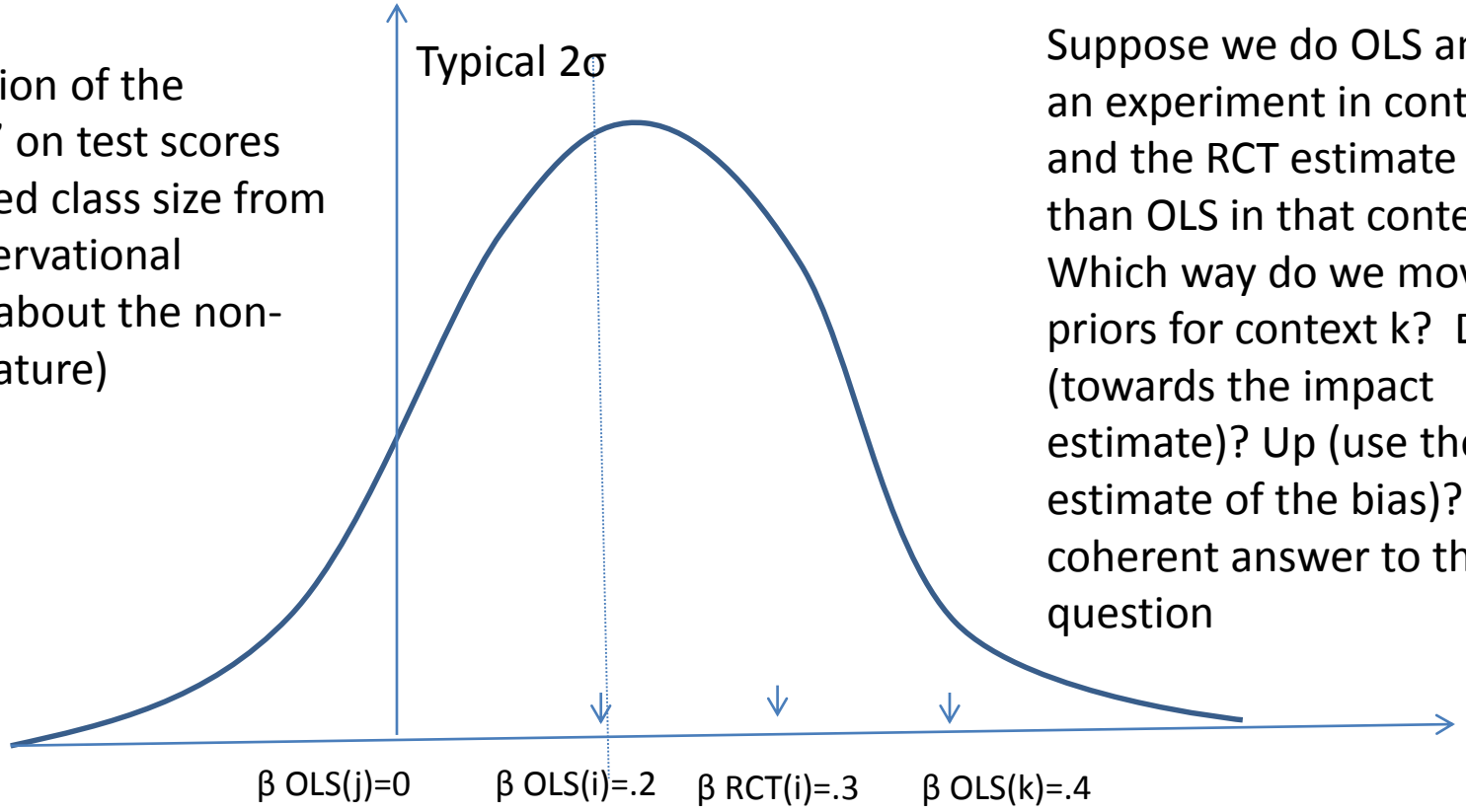
TT(1,3) is the *worst* project in Context B

Do findings from RCTs “resolve” debates through “systematic reviews” that generate findings with external validity?

No. They logically cannot and hence it is fortunate the evidence to date suggests they don’t.

Cannot work: No claim to external validity is coherent because the gap between observational and RCT results *is* the result of behavior

Distribution of the "impact" on test scores of reduced class size from non-observational studies (about the non-RCT literature)

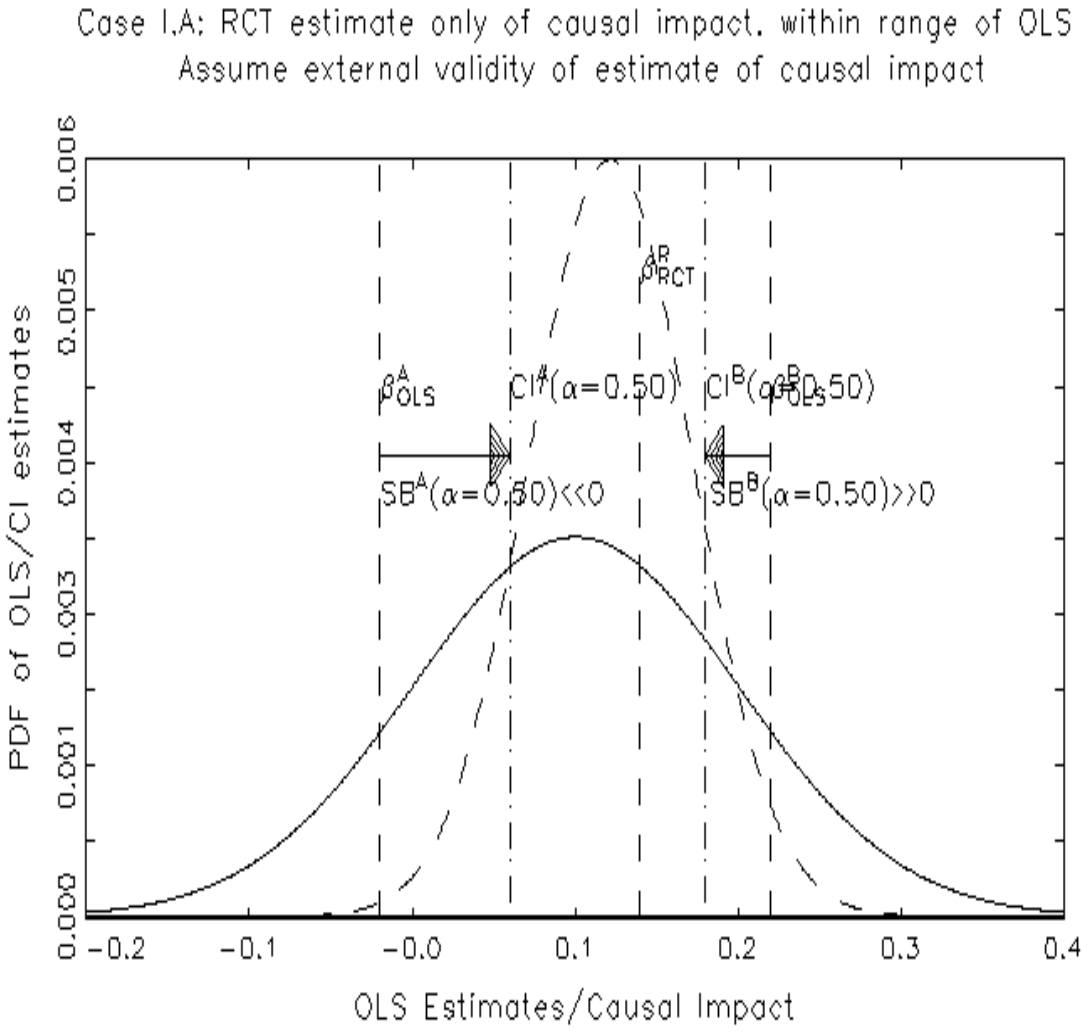


Suppose we do OLS and then an experiment in context I and the RCT estimate is bigger than OLS in that context. Which way do we move our priors for context k? Down (towards the impact estimate)? Up (use the estimate of the bias)? No coherent answer to this question

Zero

"Gold standard" RCT from one specific context (country, region, grade, range of class sizes)

Suppose we assume "external validity" of RCT estimate in context R to adjust estimates in context A and B (weight α=.5)



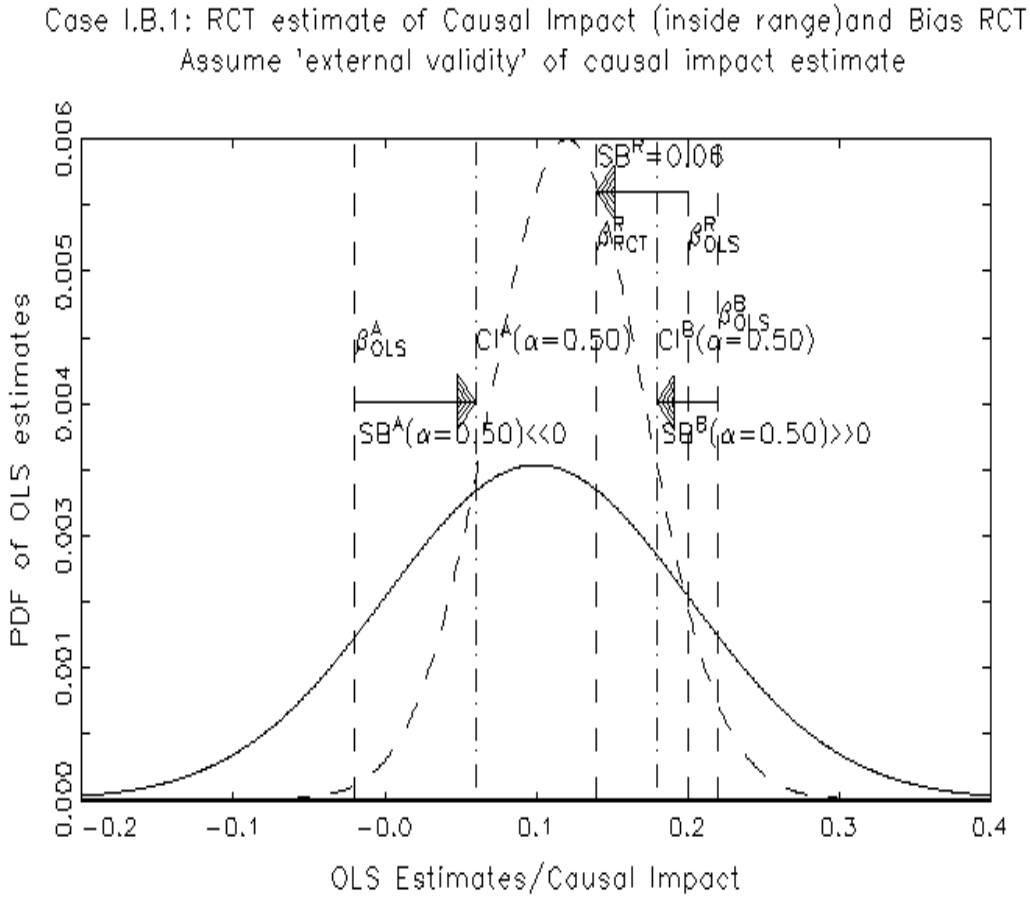
Three (related) big problems with using RCT evidence from one context for another:

I) This implies the causal impact (CI) estimates move closer to each other but the SB (structural bias) estimates move *in opposite directions*

II) This implies the *variance/heterogeneity* of OLS estimates was *larger* than the "true" heterogeneity and the heterogeneity was due to (massive) *heterogeneity in SB*

III) The implied SB of at least one estimate is different from estimated SB (see next slide)

Suppose RCT in context R produces estimate of *both* CI and SB in R—which has "external validity"?

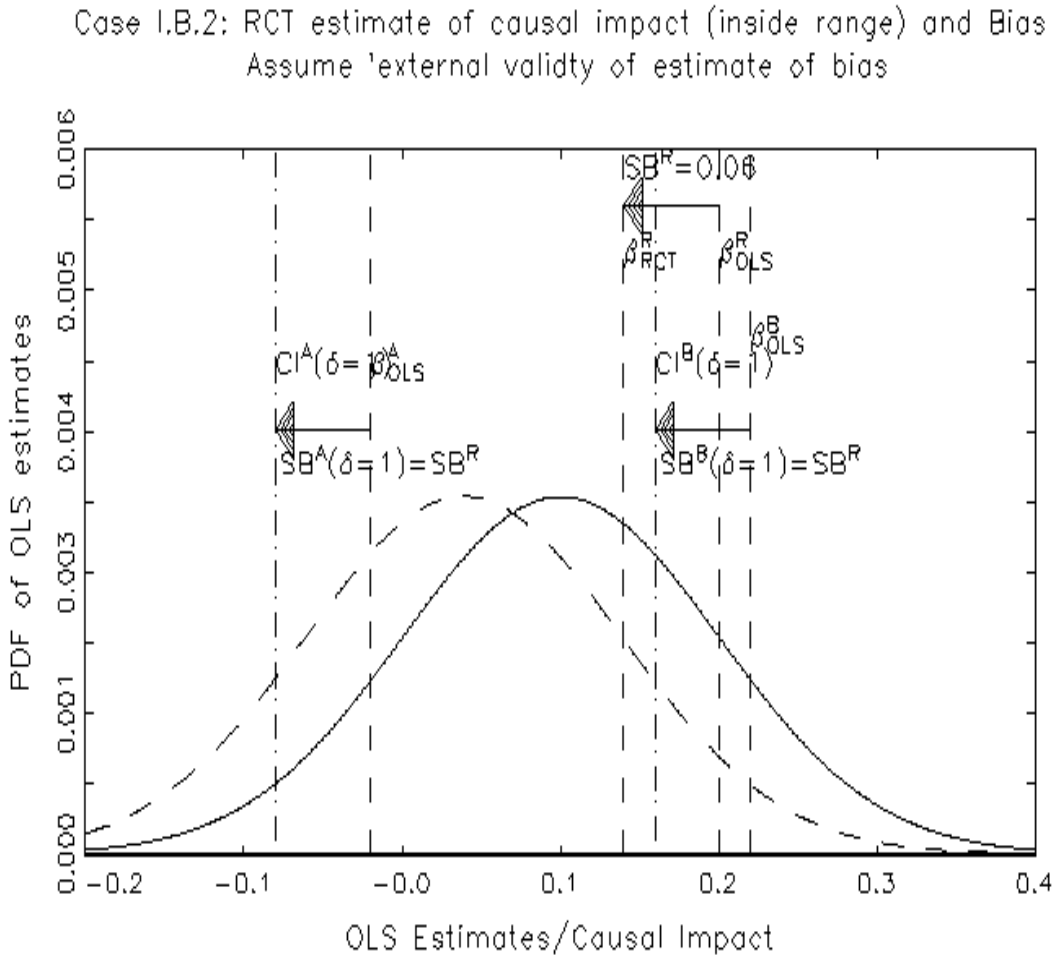


A model of the world that generated the data has to have a component of the model that explains why OLS is biased and that sub-model can be parameterized as can the sub-model in which causal impact is determined.

In this situation (that is empirically the most common when there are multiple studies across contexts) then:

a) Any positive weight on CI in context R for inference about context A and B implies that SB in either A or B is of the opposite sign from SB in R.

If one assumes external validity of the SB estimates the pattern of CI estimates adjustments is very different



There is no justification for arguing that the parameters of the sub-model that determine CI are "more" externally valid than those that determine SB.

If one assumes SB is externally valid:

A) The estimates of CI in context A move *away* from the rigorous estimate in context R (is that counter-intuitive?)
B) The *variance/heterogeneity* is preserved (not reduced) and the central tendency is shifted.

Attacking a straw man?

Every single argument I am saying the randomistas make I have personally heard them make.

So, while the first generation RCT claims might have been a straw man, it was up and dancing (and sucking in money).

Conclusion

So here I am, in the middle way, having had twenty years— Twenty years largely wasted, the years of l'entre deux guerres
Trying to learn to use words, and every attempt Is a wholly new start, and a different kind of failure
Because one has only learnt to get the better of words
For the thing one no longer has to say, or the way in which
One is no longer disposed to say it.
And so each venture Is a new beginning, a raid on the inarticulate
With shabby equipment always deteriorating In the general mess of imprecision of feeling,
Undisciplined squads of emotion.
And what there is to conquer By strength and submission, has already been discovered
Once or twice, or several times, by men whom one cannot hope
To emulate—but there is no competition—
There is only the fight to recover what has been lost
And found and lost again and again: and now, under conditions
That seem unpropitious.
But perhaps neither gain nor loss.
For us, there is only the trying.
The rest is not our business.
TS Eliot
East Coker