Is that funny? Is that evidence?

Lant Pritchett
Harvard Kennedy School and CGD

Four Jokes

Joke 1

- Q: What is brown and sticky?
- A: A stick

Joke 2

- The Bible says that wherever two or three are gathered in my name the Holy Spirit will be there.
- Experience says where ever there are four Episcopalians gathered together there will be a fifth.
- Source: Garrison Keilor Jokebook

Joke 3 (A knock knock joke)

- [Audience participation required]
- Me: Knock Knock
- Audience: Who's there?
- Me: Interrupting cow
- Audience: Interrupting…Me: Moo!

Joke 4

Me: Why did the chicken cross the road?

Audience: I don’t know why?

Me: To get to the idiot’s house.

Me: Knock, Knock.

Audience: Who’s there?

Me: The chicken.

Evidence and Projects

- Two aspects of the use of “evidence” of “what works” in the context of development projects.
  - Does the evidence have *external validity?*
  - Does the evidence have *construct validity?*
- I argue that much of the fad for the use of “evidence” is a simplistic and false analogy from physical science (or logistics) to human behavior
- Most “evidence” about “implementation intensive” and especially “wicked hard” problems in the published (and gray) literature has *neither* construct nor external validity.

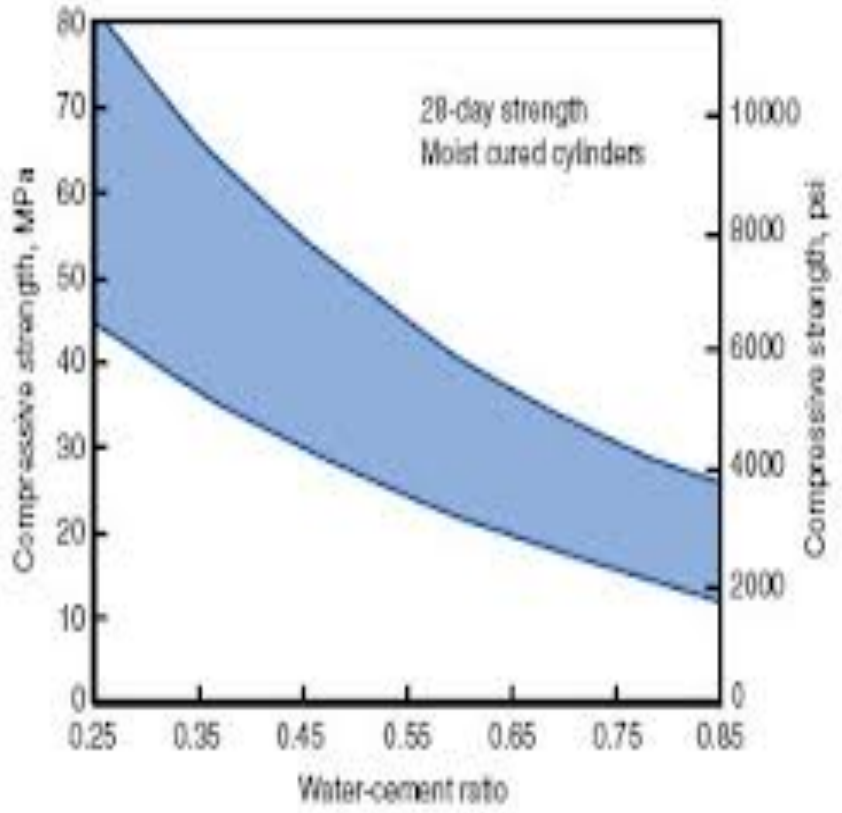
Visualization of the concepts of “construct validity” and “external validity”

- 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

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. labour 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	

Interactive effects and produce rugged response surfaces

Concrete is stronger when poured drier…



…only if it is adequately compacted

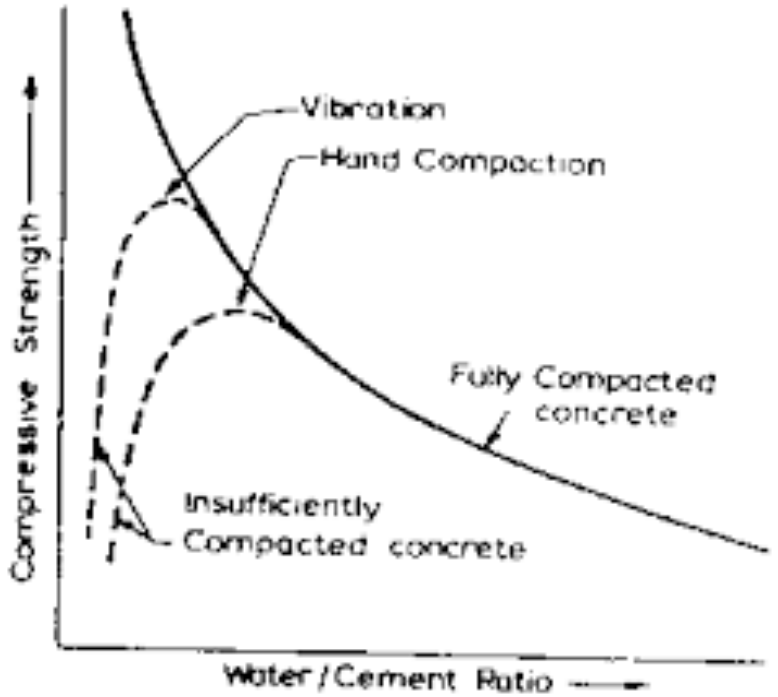
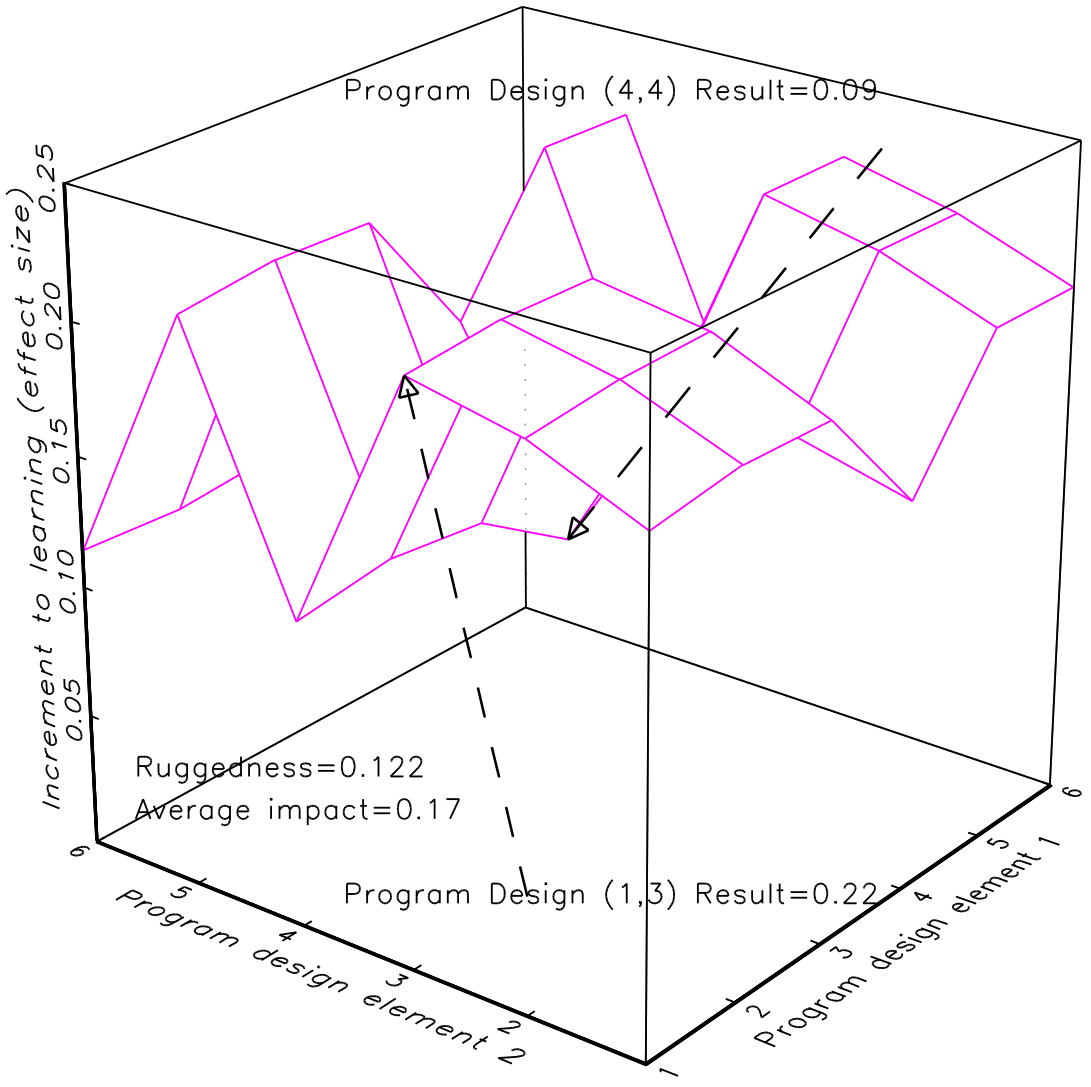


Illustration of a rugged fitness function

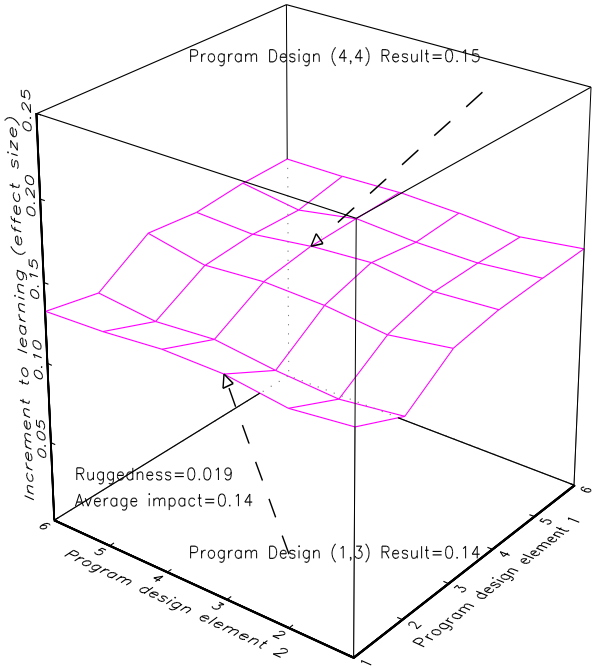


Visually distinguishing “External Validity” from “Construct Validity”

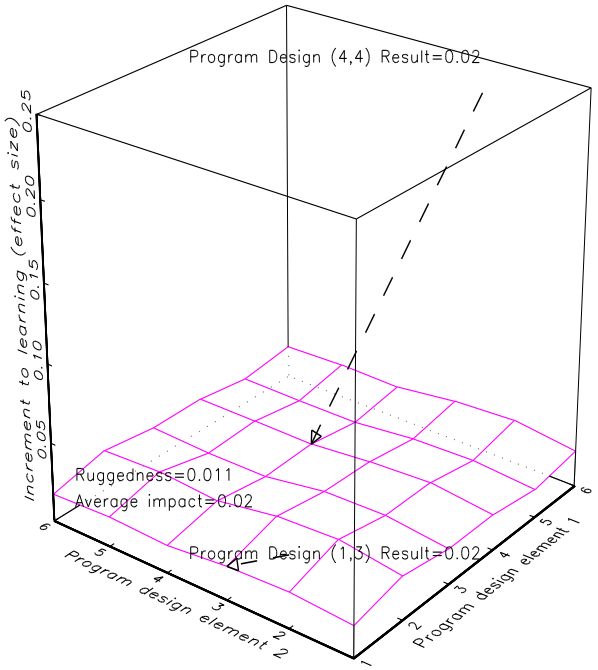
- Four *entirely hypothetical* graphs of possible effect sizes of a *class* of program (e.g. “provision of textbooks”, “ICT in classrooms”, “reduction of class size”, “performance pay”)
- I define “construct validity” to be whether there is large outcome variance across *instances* of a *class* depending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

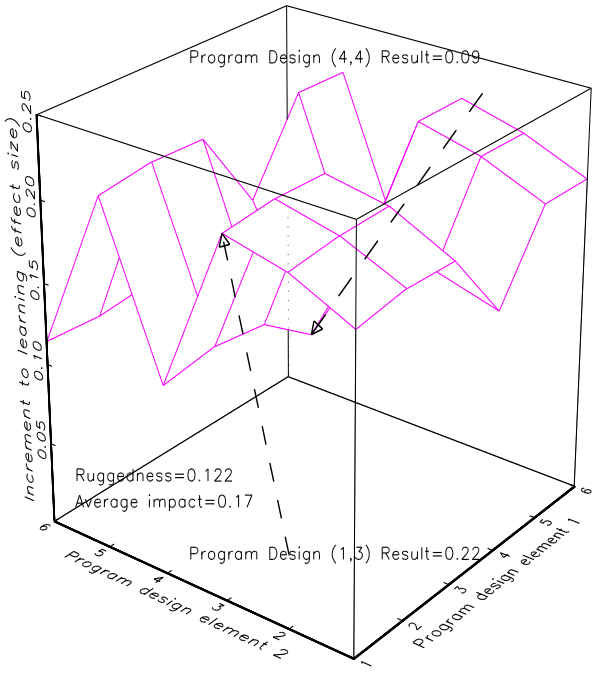


Response surface in context B—design doesn’t matter much, nothing works

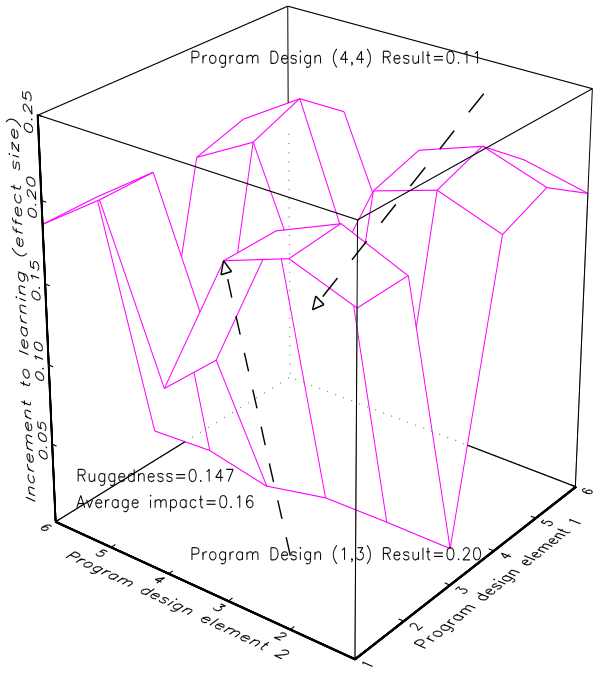


Construct validity: Rugged fitness functions imply different designs produce different results

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

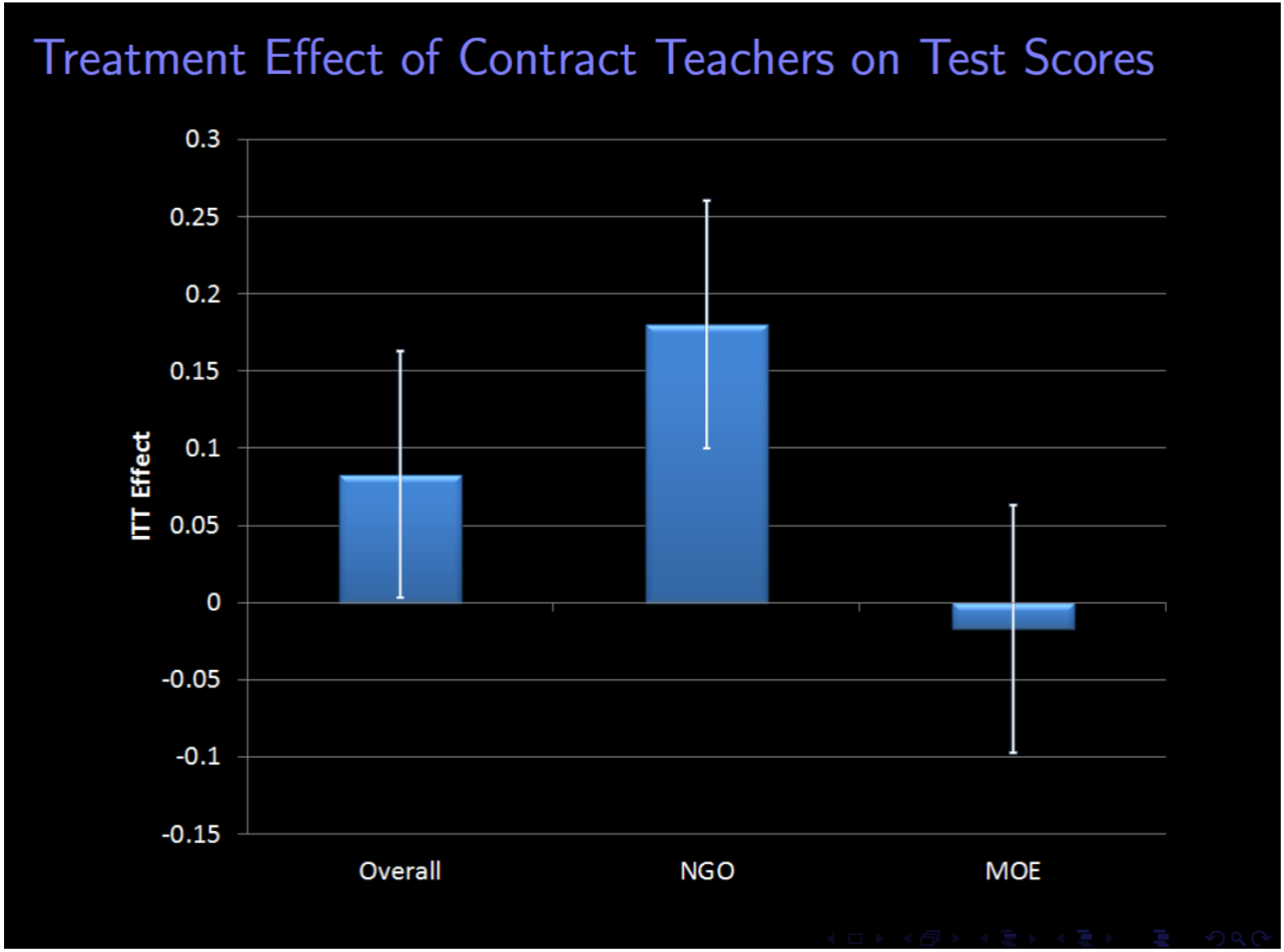


A different class of program ("teacher training")

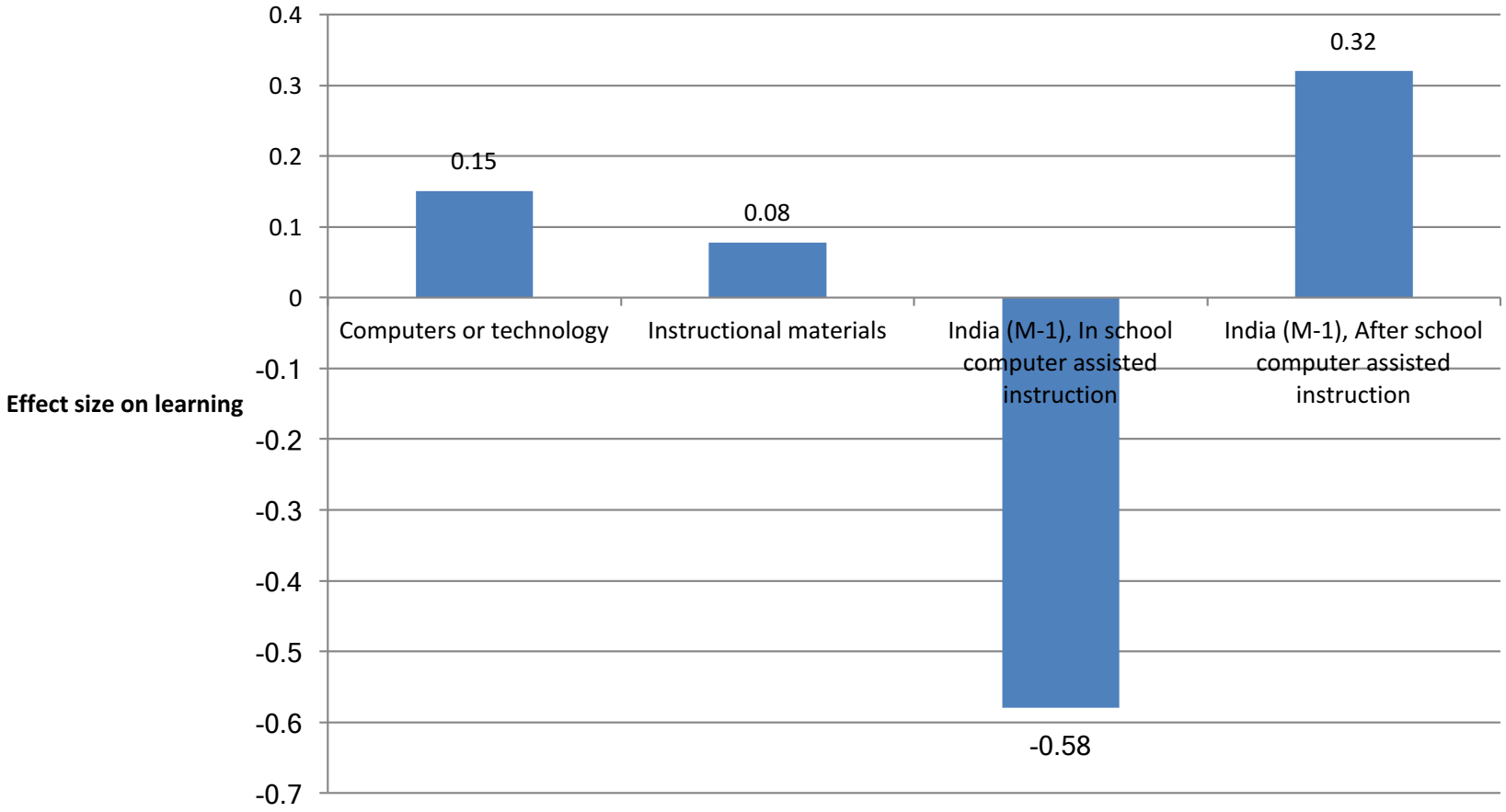


Exact same program except for one design feature, who implements

Source:
Bold et al 2013



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

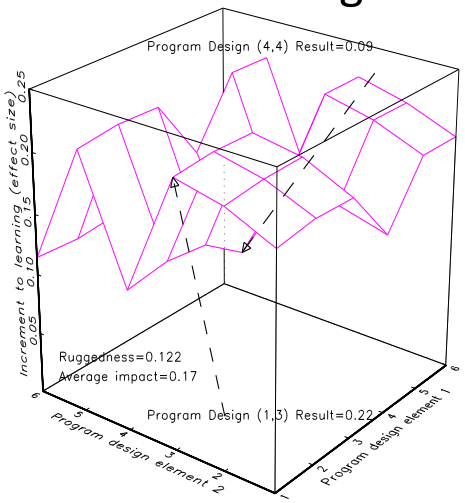


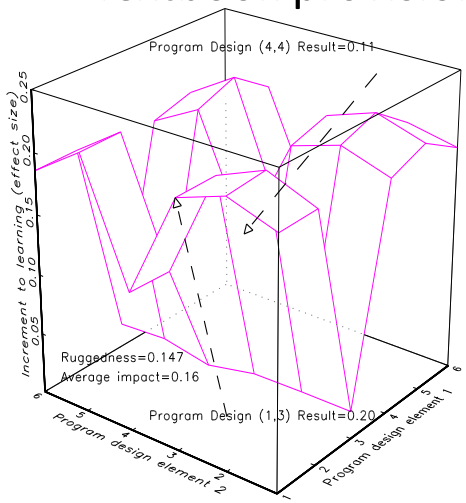
Suppose this is our world, two contexts (A and B), two classes of programs ("teacher training" and "textbook provision") with two design alternatives evaluated (1,3) and (4,4)

"Teacher training"

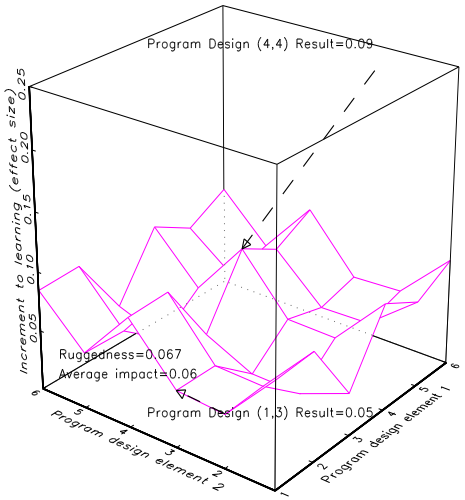
"Textbook provision"

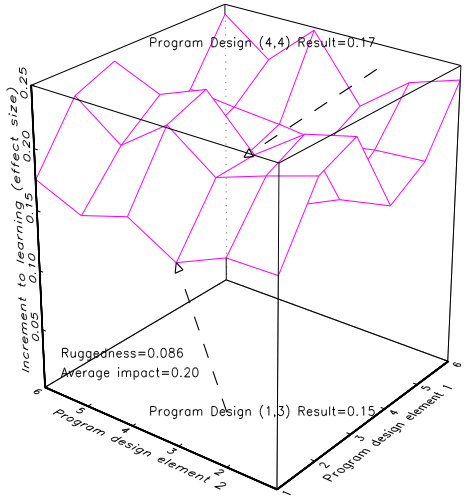
Context A





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

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.

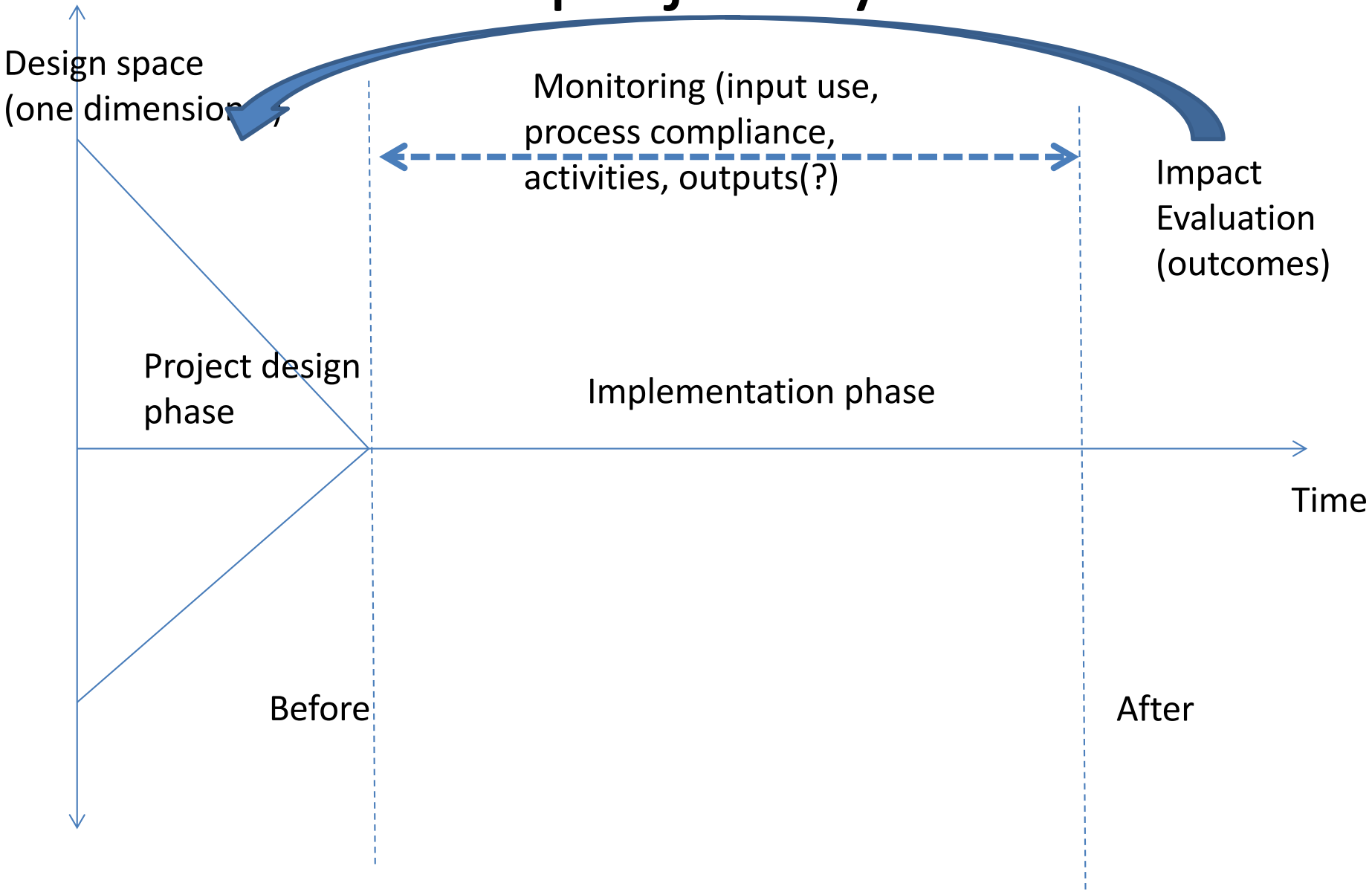
Rapid feedback loops beat all hell out of rigorous when response surface is rugged—particularly with respect to robustness of conclusions

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

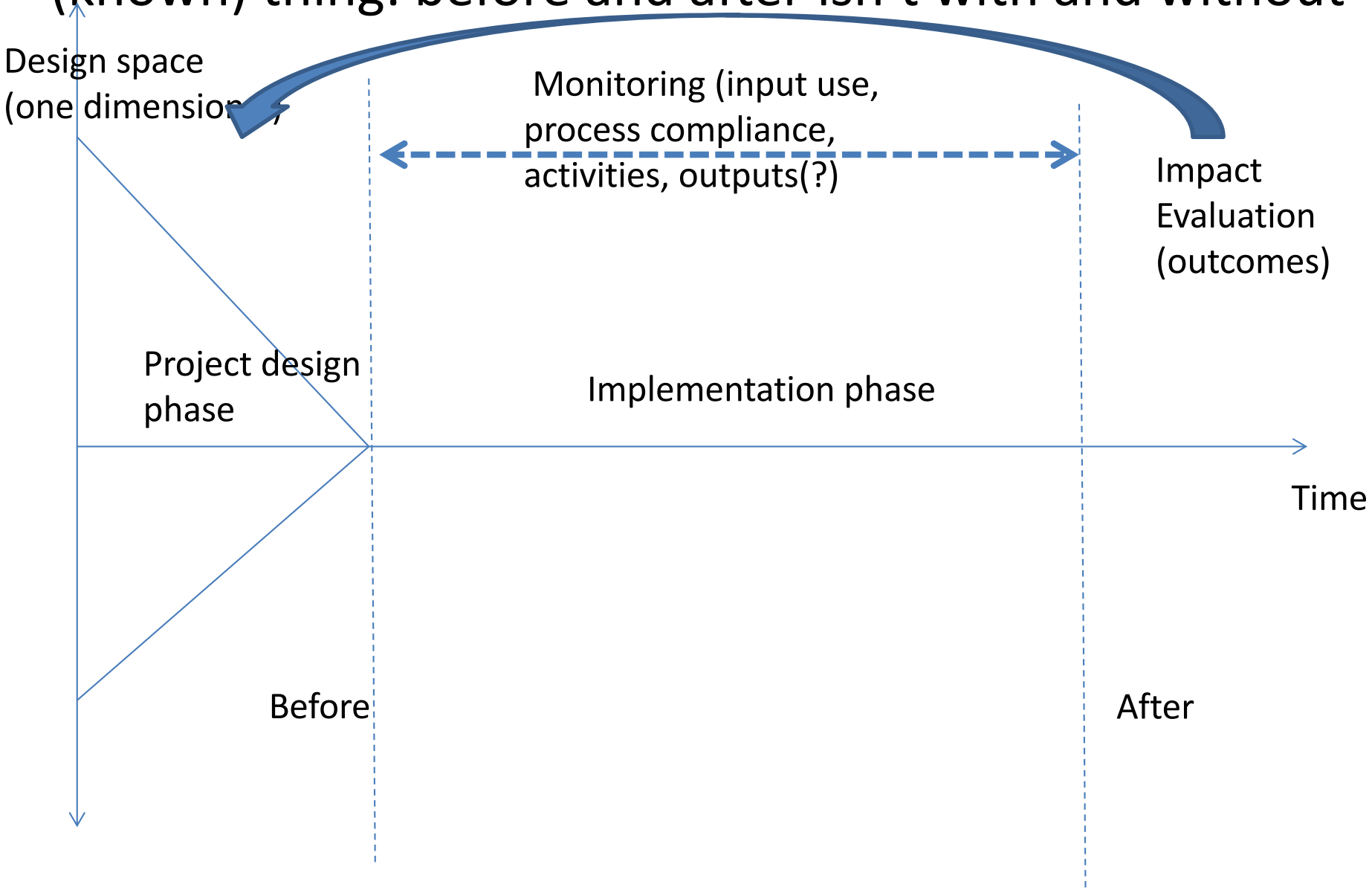
(1) Ruggedness parameter	(2) Ruggedness (absolute difference)	(3) Gain CDS over RCT (ratio to max less average)	(4) Percent excess of RCT over CDS standard deviation
.25	.020	.319	1.04
.5	.042	.445	1.19
1 (base case)	.074	.489	1.64
2	.094	.461	2.36
4	.103	.412	4.25

Source: Nadel and Pritchett 2016

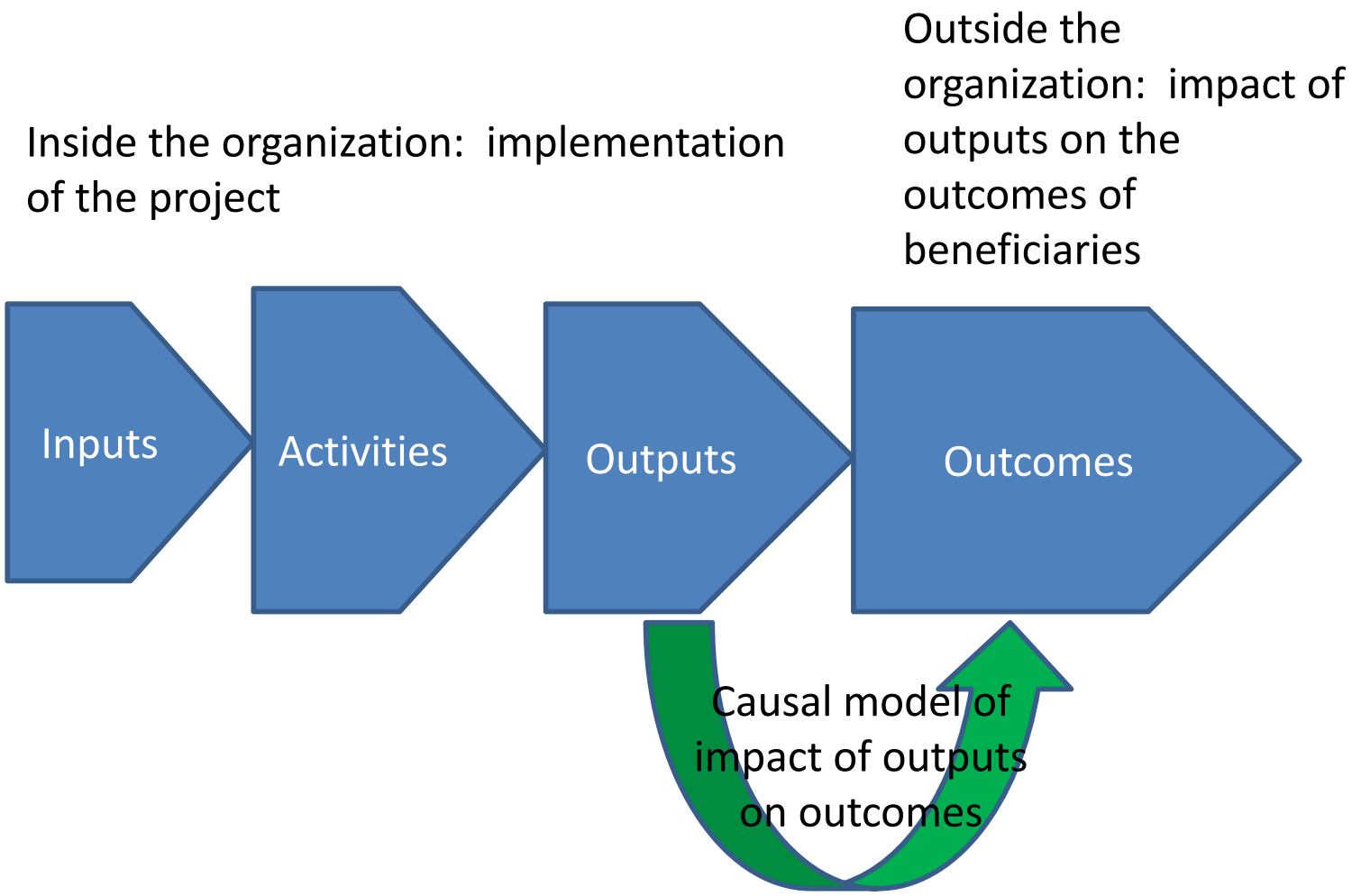
The typical (pre-adaptive) approach to the project cycle



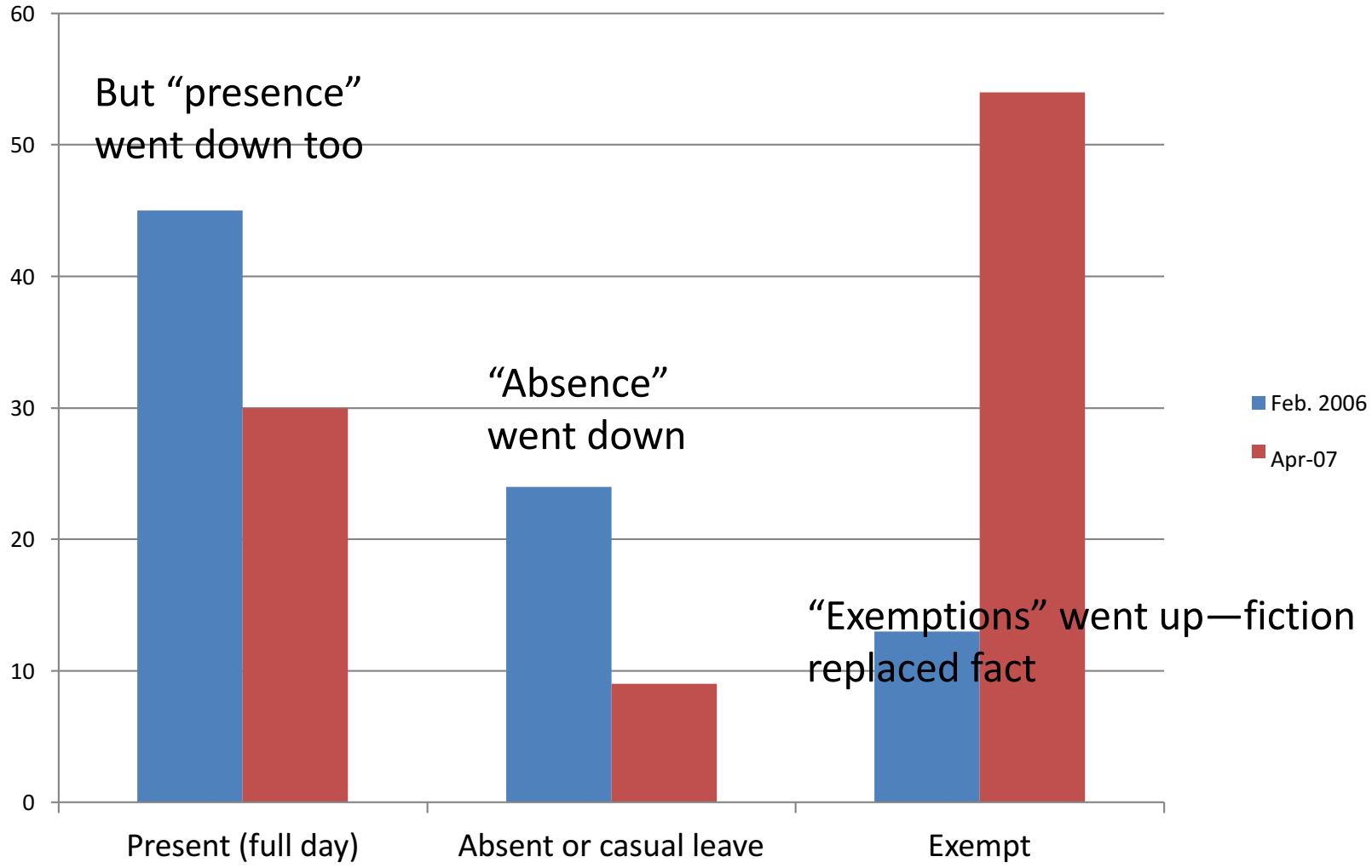
To this model of learning in the project cycle the RCT movement brought was mono-maniacal about one (known) thing: before and after isn't with and without



But many (most?) true "impact evaluations" have been failures *in implementation* not *causation*



During the course of the field experiment to motivate nurses to attend their clinics in Rajasthan…

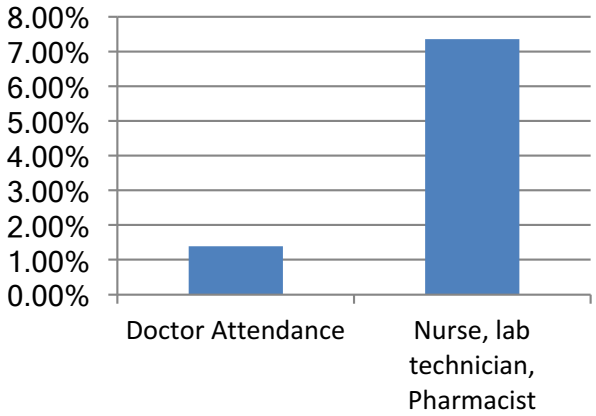


Trying to change the facts about attendance merely created an administrative fiction

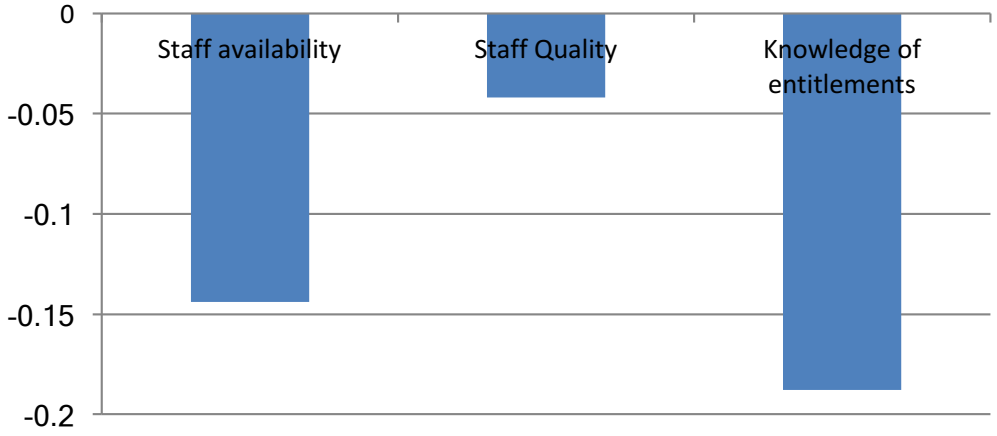
Source: Banerjee et al 2008, **Putting Band-aids on a Corpse**, adapted from Figure 3

“Deal with the Devil” (Recording attendance at PHCs in Karnataka)

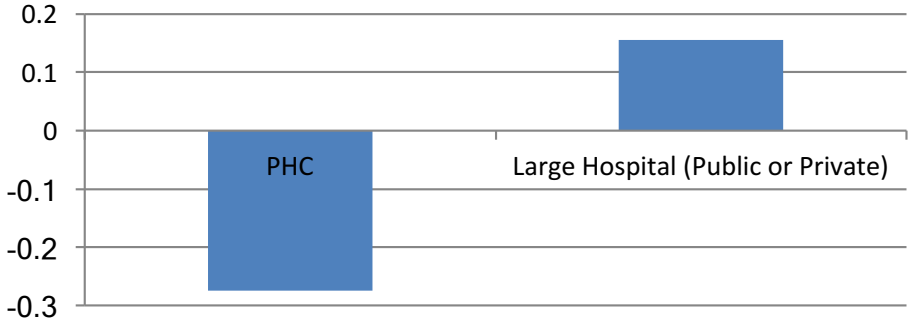
Treatment Effect on Attendance



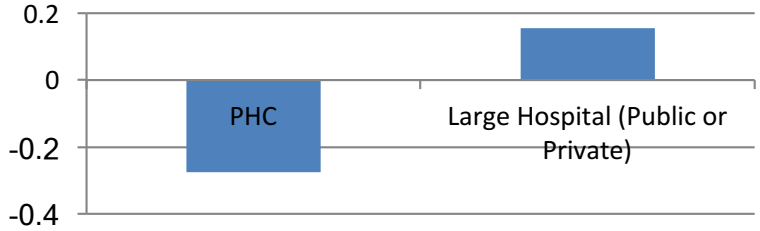
Treatment impact on patient perceptions



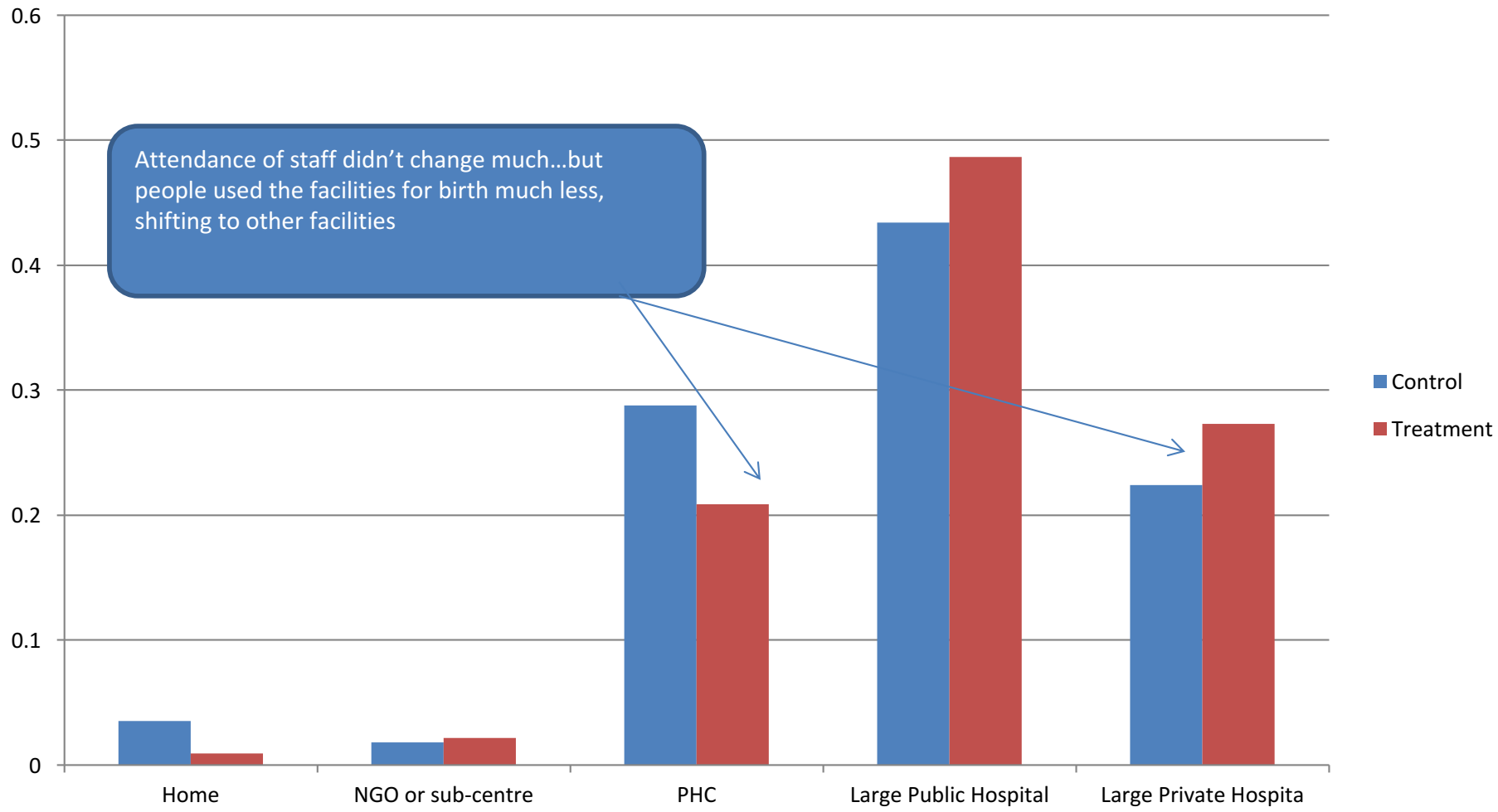
Percentage change in place of delivery



Percentage change in place of delivery

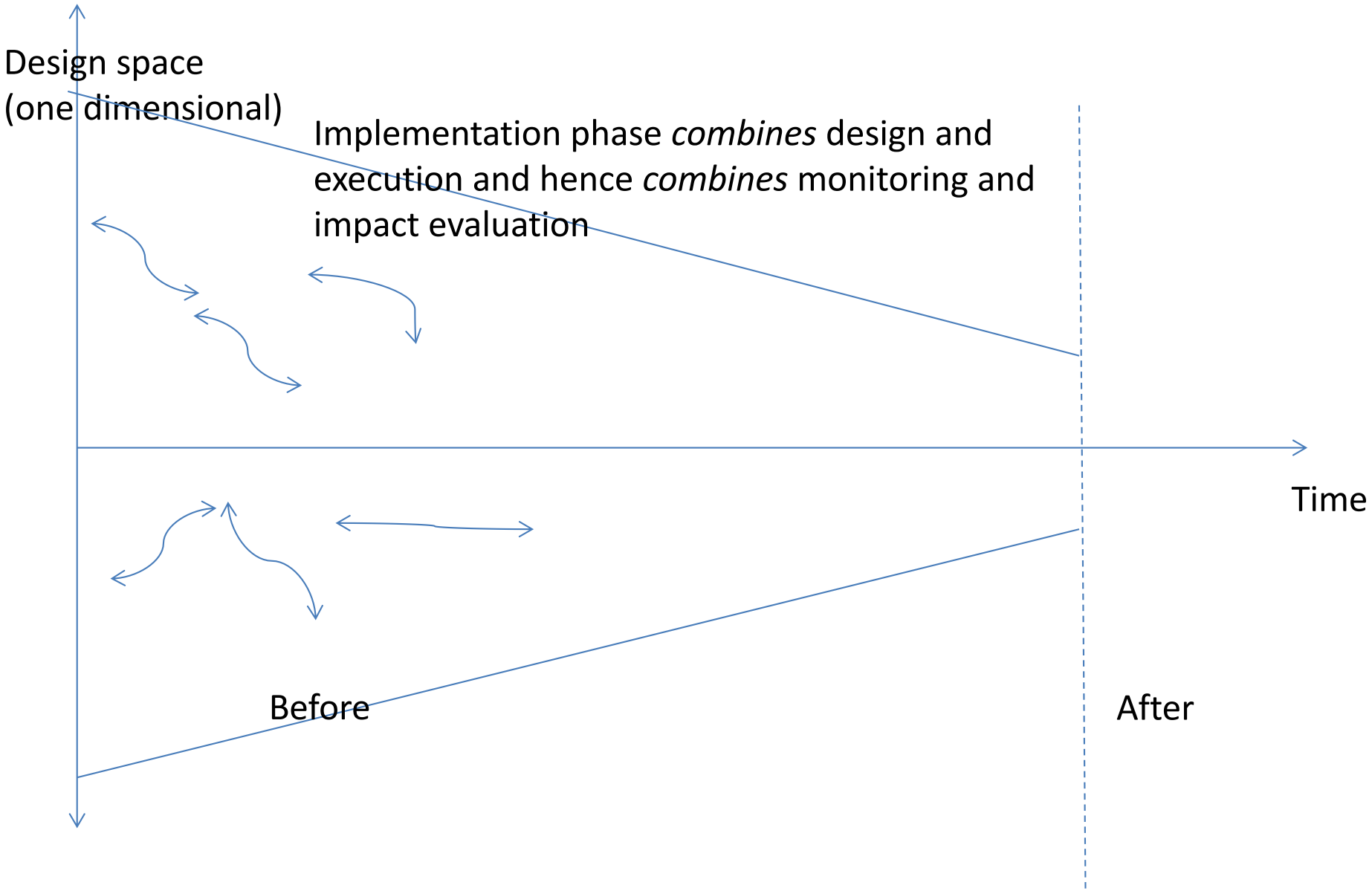


Biometric program in Karnataka India to increase attendance of all staff (Doctors, Nurses, Pharmacists, etc) at PHCs



Source: Dhaliwal and Hanna 2014

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



Big differences between the “rigorous evidence” approach and PDIA

- Who learns? In old model “designers” learn and implementers do. In PDIA implementers learn. So the evidence that is relevant is the evidence for the implementers.
- What is learned? In old model the “implementation” is assumed away and what is learned is just about causal connections of outputs to outcomes. In PDIA “what works” *both* what can be made to work (implementation) and what works to change outcomes.

What is the relation to jokes?

- Jokes are a domain of language that lack both external validity (what is funny is different across contexts, background knowledge matters for wordplay) and construct validity (minor variation in wording matters).
- Is your policy/project/program like a joke? Does its success depend on specifics of context, design and match of context to design?