Center for Global Development

CGD NOTES

"The Evidence" About "What Works" in Education: Graphs to Illustrate External Validity and Construct Validity

6/20/17 (https://www.cgdev.org/publication/evidence-about-what-works-education-graphs-illustrate-external-validity)

Lant Pritchett

Introduction

Currently, the bulk of the new empirical work on estimating the impact on learning of various education projects/ programmes/policies, while based on sound principles of estimating causal impacts, is far too inadequately theorised and specified to be of much immediate and direct use in formulating effective action to accelerate learning. Therefore, just "more of the same" empirical research is unlikely to be of much help or to add up to a coherent action or research agenda as it faces massive challenges of external and construct validity. The RISE research agenda is moving forward by: (a) embedding research into a prior diagnostic of the overall system which allows a more precise characterisation of what "context" might mean, (b) evaluating on-going attempts at education reform at scale (rather than isolated field experiments), (c) specificity about the details of programme/project/policy design, and (d) acknowledgment that policy relevant learning is itself part of the system, not a one-off exercise.

A concrete analogy (literally)

My grandfather was a construction worker and, among other things, poured and finished a lot of concrete (he said he knew Utah well because he had crawled across it, backwards). This led to my spending a summer pouring concrete for a highway overpass. Every truckload of concrete poured was tested by a state inspector. Why such vigilance? Because of the well-known relationship between the cured compressive strength of concrete and how much water is in the concrete when it is poured. Wetter is weaker. Concrete is roughly three times as strong when poured with little water (.25 water to cement) than very wet (.85 water to cement). But levelling dry concrete is like trying to level a pile of dirt (hard work), while levelling wet concrete is like levelling a puddle of water (easy work). Wetter is easier.

But it is more complicated than that. Concrete is also weaker if it has air pockets and is not fully compacted when poured. When the water/ cement ratio is low, it is more difficult to achieve full compaction. When very dry, active vibration is required to compact the wet concrete. So while Figure 1a gives the compressive strength-water/cement relationship at optimal compaction, Figure 1b shows that insufficiently compacted low water/cement concrete has very low strength[1].

Figure 1a: Concrete is stronger when poured with a lower water-cement ratio

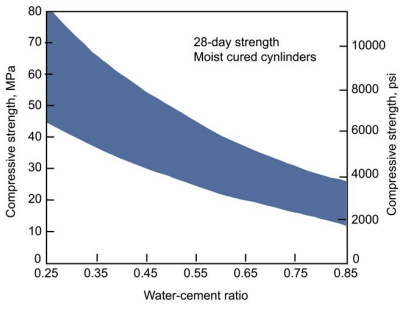
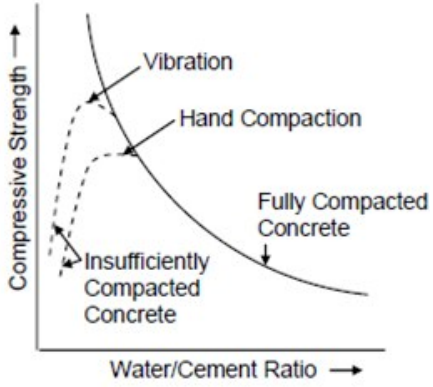


Figure 1b: ...but insufficiently compacted dry concrete is much weaker than fully compacted wetter concrete



The evidence about "the evidence"

Now with that concrete background, let's discuss "the evidence" about "what works" for increasing learning in the developing world, and more specifically, about the evidence of "the evidence." By "the evidence,"[2] I mean the current agenda of producing "rigorous" estimates of the casual impact of specific projects/programmes/policies through the use of identification techniques that minimise bias and then summarising those individual estimates of "what works" through *systematic reviews*.

There are six important pieces of evidence about "the evidence."

First, many common sense, widely accepted, and likely true, facts about education are not proven, or even appear contradicted, by "the evidence." For instance, Glewwe and Muralidharan (2015) report on four well-identified estimates of the causal impact of providing textbooks in cases where textbooks were not available to every student. Each of them show that the causal impact on learning of the typical child of additional textbooks is zero (or that the hypothesis tests fail to reject zero). So "the evidence" would conclude there is no, or weak, evidence for the universal (and almost certainly correct) practice of seeking to provide a textbook for every child.

Second, the only rigorous evidence shows that rigorous evidence isn't. To my knowledge Bold et al. (2013) is still the only rigorous developing country education study on the impact of a government scaling up something that had been "proven to work" with "rigorous evidence." When the government of Kenya's Ministry of Education scaled up a programme that had been rigorously shown to produce substantial learning effects by world class researchers, the programme had exactly zero impact (the estimate was slightly negative). The "rigorous evidence" about what works when implemented by a non-governmental organisation (or researchers as a field experiment) did not prove to be rigorous evidence about what would happen in the context of government implementation [Vivalt (2016) shows this difference in impact across implementers is true more generally, not only in education].

Third, *systematic reviews* of the same body of "the evidence" come to widely different conclusions. Since education interventions are easily amenable to randomisation, there have now been sufficient studies of "what works" to do *systematic reviews* and in fact, there have now been five or six such *systematic reviews*. Evans and Popova (2015) show the *systematic reviews* come to very different conclusions about "what works"-in part because (a) there is so much heterogeneity across the studies, even of the same type or class intervention, that modest variations in the inclusion rules of what studies are to be included, lead to very different conclusions and (b) the classes of interventions were not consistently defined.

Fourth, the variance around the estimates is enormous for the same class of intervention compared to differences across classes. For instance, a *systematic review* by McEwan (2014) suggested that the class of interventions he calls "ICT," have an average effect size (impact to standard deviation) of .15 compared to only .049 for "information" programmes. But the variation across ICT programmes was massive compared to the variation across the classes, or types, of programmes he reviewed. Within the ICT programmes included in the review, there was one with an effect size of positive .32 and one with an effect size of negative .58 - a range of .86 compared to the standard deviation across the classes of interventions of only .05. Moreover, the positive .32 and the negative .58 impact estimates were different treatment arms in the same experiment and hence from exactly the same context.

Fifth, most people agree there are plausible phenomena that produce interactions such that the exact same intervention will have different impact in different situations. For instance, Beatty and Pritchett (2013) articulate a simple model of "overambitious curriculum" in which there is a mismatch between the teaching level and the average student level of ability. In the model, it is easy to show that rigorous evaluation of the same programme would produce the same learning gain at each grade, only if there were no curricular mismatch. With curricular mismatch, differences across contexts of the same programme will produce wildly varying results across countries and regions, and these variances will occur at different grades. For instance, interventions that group students by skill level and not grade in Bihar, India, produced the equivalent of three years of regular grade-based learning in literacy in just eight months. But this almost certainly is due to the extreme heterogeneity that characterises these schools and the same intervention in "curricular matched" schooling systems would be expected to be much lower. So the "rigorous evidence" about "what works" cannot be interpreted or extrapolated independently of the specification of these interacting factors that everyone agrees are present.

Sixth, the studies often lack sufficient contextual detail to allow replication, direct policy application, or the analysis of the impact of programme design. For instance, Evans, Popova and Arancibia (2016) note that pretty much everyone agrees effective teaching is at the heart of good education systems, but at the same time acknowledge that general assessments (or evaluations) of teacher training are often pretty dismal. Hence, they attempt to dig into the question of, "What type of teacher training works?" honing in on details of the in-service training (what type? where? about what?). The first finding is that the published literature on which "the evidence" is based, simply lacks adequate contextual detail to answer these questions. For instance, of the twenty-six article types that were called "rigorous evidence," and hence that a systematic review would include, only forty-three percent of the information needed to understand programme content was available in published works. Likewise, only thirty-six percent of the information needed to look at how the teacher training was delivered was available. If content and mode of delivery are key factors in "what works" in teacher training (and they likely are), then the existing evidence simply lacks evidence to assess these dimensions of programme design.

No one ever imagined that improving learning would be simpler than concrete. Go back to Figure 1a and imagine a randomised experiment that increased the water/cement ratio. What would the "expected" result be? It depends-if the water/cement ratio were low and compaction inadequate, this experiment could produce big improvements in compressive strength from a higher water/cement ratio, whereas if compaction were adequate, the experiment would produce reductions in compressive strength. Without adequate specification of the context, the interacting factors, and all of the details of the intervention; the results of an experiment, even if absolutely rigorous about the causal impact of what happened, have no general value (and, when misapplied, can be worse for formulating policy than simpler, context specific evidence that isn't "rigorous").

Visualizing "the evidence": Simple illustrations of external and construct validity

I will articulate two concepts, *external validity* and *construct validity*, and then I will illustrate those concepts with fairly simple graphs. I argue these help explain why the currently conventional approach to "the evidence" has been, and will continue to be, of limited value without being more deeply embedded in system approaches (to deal with context) and performance oriented learning approaches (to deal with construct validity).

To define the *construct validity* of the causal impact of a project/programme/policy, we need the ideas of a design space and an objective function/response surface/fitness function over that design space.

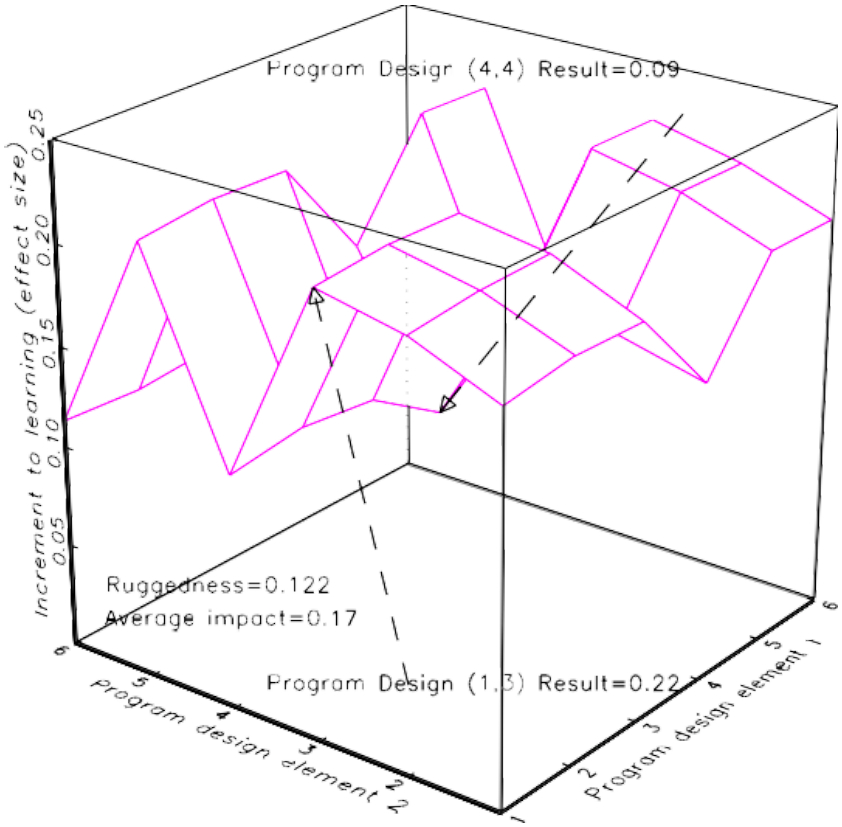
The design space is the combination of (a) each of the elements of programme design and (b) each of the options for those elements. These are what need to be specified to move from the design space for a class of programmes to an implementable specific instance of programme. For instance, a conditional cash transfer (CCT) is the name of a general class of programmes which share the basic feature that people receive cash transfers only if they do certain things, hence the term "conditional." But as Pritchett, Samji and Hammer (2012) illustrate, to move from the generic class of CCTs to implementation of an actual programme, there have to be choices about each of the design elements. Who is eligible to be a beneficiary? How large is the transfer (on average)? Does the amount of the transfer vary across households (e.g., larger transfers for larger households, larger transfers for poorer households)? How often is it paid? To whom is the cash transfer paid? Which agency is responsible for implementation? Suppose the conditions are in relation to school attendance-there are more design elements: How high does attendance need to be? Are there learning conditions? Who certifies the conditions are met? Any specific CCT programme is one instance in the overall design space of the class of CCT programmes.

And CCT programmes are an example because they are so simple.

Evans, Popova, and Arancibia (2016) developed a design space (my description of their work) for the generic *class* of in-service teacher training programmes with a survey instrument to produce indicators of elements and options. For instance, a design element is: "What is the primary content of the training?" This element then has four options: subject content, pedagogic practice, technology use, and counselling. Another design element is: "How is the content delivered?" The options are: lectures, discussion, lesson enactment, materials development, how to conduct diagnostics, lesson planning, and use of scripted lessons. They also have design elements for who implements the programme, the delivery mode, degree and type of follow up, etc. Their classification produces fifty-one indicators they feel are minimally necessary to describe a specific teacher training programme. Their design space has fifty-one dimensions (keep this in mind when looking at the following graphs with two dimensions of the design space).

The second background concept to *construct validity* is the mapping from elements of the design space to specific indicators of outputs, outcomes, or impacts. In different disciplinary domains this is alternatively called a *fitness function* (evolution, biology), *objective function* (computing, mathematics), or a *response surface* (medicine, social science). The *response surface* is most easily thought of as the average gain on an indicator (output, outcome, impact) of a selected population exposed to a specific programme (as an element of the overall design space) compared to those of a similar/ex ante identical population not exposed.

Figure 2: Illustration of a rugged response surface (learning gain in effect sizes) over a design space with two elements and six options for each element (thirty-six possible programmes)



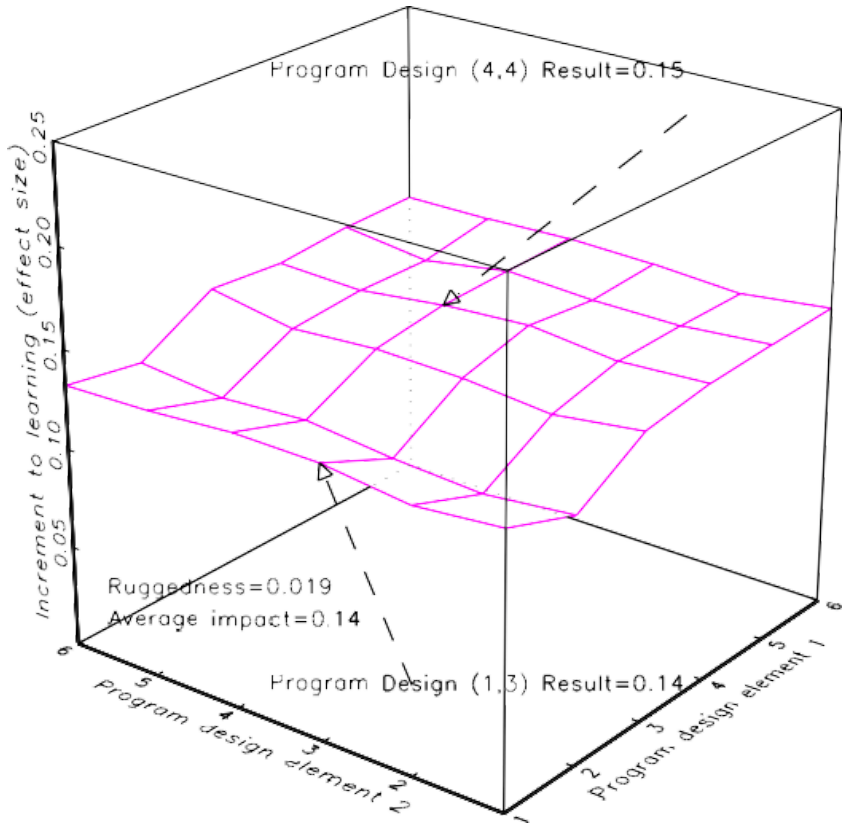
With the concepts of design space and a response surface over the design space, we can produce Figure 2 which is an entirely hypothetical illustration of the learning gain of a student population (response surface) exposed to different instances of a class of education programmes with two design elements, each with six options (design space). Figure 2 is not simple, but is as simple as possible to understand the existing evidence.

Figure 2 illustrates a "rugged" response surface over the design space, by which I mean that different combinations of the design elements produce very different impacts. In this entirely hypothetical situation, a programme design with option one for element one and option three for element two [programme design (1,3)] produces an impact of .22. If we use option four for element one and option four for element two [programme design (4,4)], this only produces an impact of .09. The ruggedness of the space is defined as the average difference between a programme design and all designs that are just one step away (all of the programme's one design step neighbours) from all possible programme designs.[3] In this response surface, the ruggedness is .12 and the average over all thirty-six possible designs is .17, so typically moving one design step would result in a very large absolute difference in outcomes. For instance, one can see that moving from programme design (1,3) to (1,4) would cut the programme impact in half.

Figure 3 illustrates a response surface that is "smooth" over the design space. In this (again, entirely hypothetical) illustration, differences in programme design make relatively little difference to outcomes so that programme design (1,3) and programme design (4,4) produce roughly the same result (.15 versus .14). The ratio of ruggedness (.019) to average impact (.14) is very small.

With these two figures we can illustrate several points that emerge from the evidence about the "evidence."

Figure 3: Illustration of a smooth response surface (learning gain in effect sizes) over a design space with two elements and six options for each element (thirty-six possible programmes)



Lack of construct validity (even with external validity)

Imagine there are two classes of education programmes: *teacher training* and *textbook provision* and that each of those classes has a rugged response surface over its design space. A good design would produce big results and a bad design would produce zero impact, but the average impact of the two types of programmes (over the design space) or their "best design" impact would be roughly equal. Imagine Figure 4a represents the rugged response surface for different teacher training designs and Figure 4b represents the rugged response surface for the design space of textbook provision. Now imagine that researchers do two excellent experiments, one for each programme type, that produce cleanly identified estimates of the causal impact of the specific programme of each type. What can be concluded? Nothing, absolutely nothing (nothing that is, beyond a literal repetition of the results).

Figure 4a: Possible response surface for teacher training (same as Figure 2)

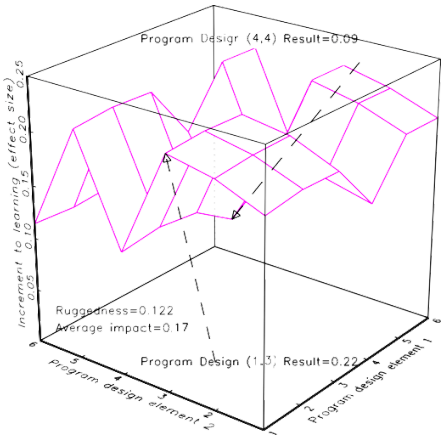
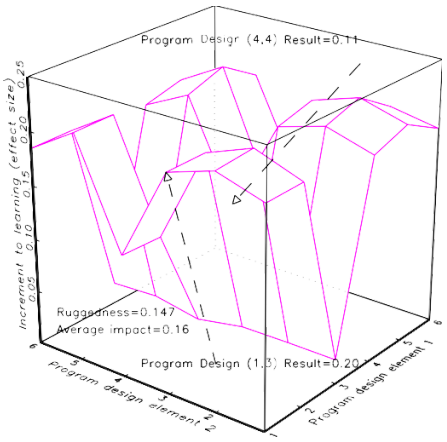


Figure 4b: Possible response surface for textbook revision

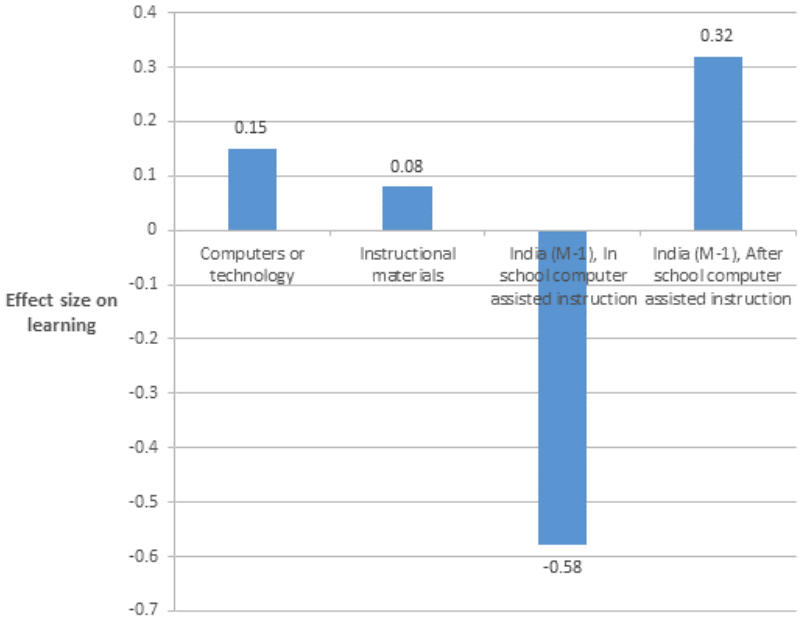


Suppose the researchers (unknowingly, of course) happened to choose a good design for *teacher training* and a bad design for *textbook provision*. A "review of the rigorous evidence" would conclude, "teacher training is better than textbook provision." Of course, it could have also happened, given both response surfaces are rugged, that equally rigorous research happened to investigate a bad design for *teacher training* and a good design for *textbook provision* and a "review of the rigorous evidence" would conclude, "textbook provision is better than teacher training."

The use of experimental evidence is often promoted by an analogy to the assessment of pharmaceuticals, where double blind randomised control trials (RCT) are the "gold standard" often insisted on by regulatory agencies to approve new drugs. But no one in medicine asks the question, "Do drugs work to fight disease?" as it is obvious the question lacks construct validity-What drug (in an exact chemical specification that can be reproduced globally)? In what dosages (in exact and replicable amounts and timings)? Given in response to what observable diagnostic indicators? But people do, unfortunately, write systematic reviews about "the evidence" on "what works" in education without any of this specificity on the instances of each class.

If there is lack of *construct validity*, then we would expect to see the *within class* variance of estimates of causal impact would be large relative to the *across class* estimates of casual impact. In fact, one would expect to see large differences in causal impact across treatment arms in the same experiment. The treatment arms would be testing specific designs in the same context, so differences cannot be the result of *external validity*. For instance, in his review of "the evidence" about impacts of various classes of education interventions, McEwan (2014) compares an intervention class called *computers or technology* to an intervention class called *instructional materials* and shows that averaged over the available rigorous evidence, *computers or technology* had an average impact of .15 and *instructional materials* had an average impact of .08. The naïve policy recommendation might be, "pursue interventions in *computers and technology* rather than *instructional materials*." But, if one looks at the range of impacts within the class *computers or technology*, one can see a massive variation as demonstrated in Figure 5. In fact, one study in the same context reported that one treatment arm had a *negative* impact of -.58 when computer assisted instruction was during school and a massive *positive* impact of .32 when computer assisted instruction was after school. The difference on the same response surface (same context) within the design space was .9 compared to an average difference across classes of interventions of .08. Clearly the message is: "get the intervention design right" rather than "do the right class of intervention."

Figure 5: Differences within treatment arm variations of the same intervention class can be orders of magnitude larger than differences across classes of intervention



Source: McEwan (2014)

Lack of external validity (with construct validity)

An alternative possibility that can be illustrated with these same hypothetical response surface graphs, is that *construct validity* is not a big concern as the response surface is smooth over the design space, but there are contextual factors that affect the impact of the class of intervention. Figure 6 illustrates the case in which *textbook provision* has a big impact in context A and a small impact in context B, but about the same impact across all programme designs.

The distinction between *construct validity* and *external validity* is whether the factor(s) that determine the difference in the response surface is *contextual* or *design*. This distinction is obviously not hard and fast, as what is or is not in the "design space," depends on what, in a given political and policy environment, is feasible in the "authorising environment" (Moore 1984).

Figure 6a: Possible smooth response surface over design space of textbook provision in context A (same as Figure 3)

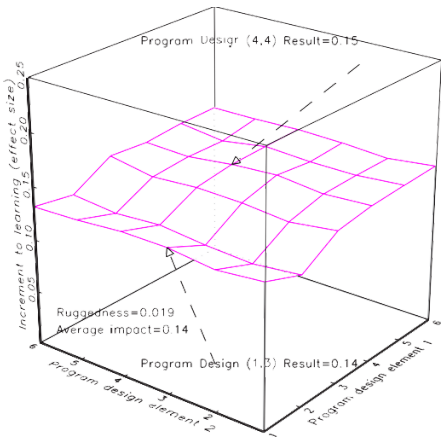
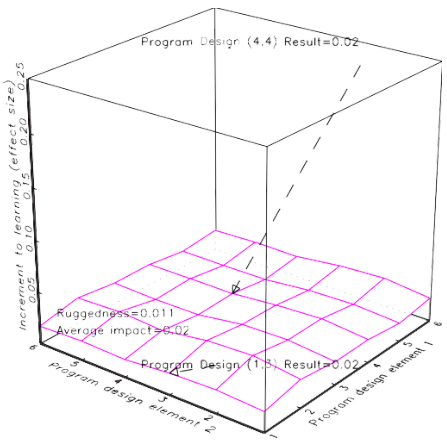


Figure 6b: Possible response surface over design space for textbook provision in context B



For instance, take the Evans, Popova, and Arancibia (2016) classification of the design space of in-service teacher training. Many features are always elements of the design space, like the content of the training (e.g., subject matter versus pedagogical) and the mode of delivering the content (e.g., lecture time versus time practicing with students). But they also include in the classification overarching

features like participation, which has implications for promotion, points towards promotion, or salary implications. Often those designing an in-service teaching training project/programme will (rightly) perceive that whether participation has salary implications is not part of their currently politically authorised design space, but rather has to be taken as a *contextual* feature.

Or consider that textbooks may be far too difficult for the typical student. A programme of providing additional textbooks to a given school or given students may, rightly, perceive that their programme/project design space does not include revising the content of the textbook, even though it can be shown analytically [e.g., Beatty and Pritchett (2012)] that this is a feature that will affect programme impact.

One empirical consequence of a lack of *external validity* will be that "the evidence" will show large variability of results across contexts. This variability across contexts can be large enough that the "rigorous evidence" from context A could be a *worse* predictor of actual impact in context B, than the simplest possible evidence (not a "rigorous" estimate of the true response surface) in context B. Visually, just imagine that in Figure 6a, the biased response surface estimates from a simple procedure like OLS lie uniformly twenty percent below the true response surface, so that if the true value is .14, the biased estimate will be .14*.8≈.11, but the average rigorously estimated impact for context B is .02. The error from the bad method in the right context is .03=.14-.11, whereas the error from the good method, but wrong context (using rigorous evidence from context B for context A), is four times as big .12=.14-.02 [Pritchett and Sandefur (2013) use actual RCT empirical results from microcredit programmes across six contexts to show the mean square error of prediction is actually larger using the rigorous evidence from other contexts, than using the simplest OLS from the actual context].

Lack of both construct validity and external validity

It is, of course, possible that both (a) the response surface is rugged over the design space in each context and (b) due to interactions with contextual factors, the response surface is different in different contexts in both average level and shape.

Then the world that we are researching could look like Figure 7 in which both design and context matter, and matter interactively. That is, Figure 7a and 7b compare two (possible) response surfaces for a class of *teacher training (TT)* programmes. In context A, design TT(1,3) works substantially better than design TT(4,4) (.22 versus .09) whereas in context B, design TT(4,4) works better than design TT(1,3) (.09 versus .05). Moreover, in context A, *teacher training* design TT(1,3) works much better than *textbook provision* design TP(4,4) (.22 versus .11), but in context B, *textbook provision* design TP(4,4) works much better than *teacher training* (.17 versus .05). Suppose there were two studies with "rigorous evidence" from context A that compared treatment arms on *teacher training* TT(1,3) and TT(4,4) and treatment arms on *textbook provision* TP(1,3) and TP(4,4) (or perhaps four separate studies from context A, each of which did one of the options). In this illustrative case, everything about the conclusions drawn from this "rigorous evidence" would be wrong for context B. Whereas *teacher training* design TT(1,3) is the *best* option of the four evaluated in context A, it is the *worst* option in context B.

I realise this was a slog, but I hope the payoff is worth it. I feel that you cannot understand "the evidence" about "what works" to improve learning that emerges from the systematic reviews of "rigorous" studies estimating causal impacts on learning of various education programmes/projects/policies, without understanding the notions of design spaces, response surfaces, construct validity, and external validity (or something very much like them).

However, once you understand these notions, it is easy to understand all of the "puzzles" raised by the existing evidence- common sense not supported (textbooks don't matter?), rigorous results not replicating, contradictions about "what works" emerging from seemingly identical systematic review processes, persistent heterogeneity within classes of interventions, etc. These are exactly what one would expect from a world in which there is a lack of construct validity due to rugged within-class response surfaces and inadequate specification of the programme design details, as well as a lack of external validity due to interactions of contextual features and design to produce different response surfaces in difference contexts.

Moreover, there can be no presumption that ignoring these construct and external validity concerns and acting on the "best available rigorous evidence" will work out. It is easy to construct counter-examples in which "rigorous" evidence from context A would recommend exactly the wrong policies for context B (Table 1).

Illustration of the possibility of the lack of either construct validity (response surfaces are rugged over each design space) or external validity (response surfaces differ across contexts):

Figure 7a: Possible response surface over teaching training design space in context A (same as Figure 1)

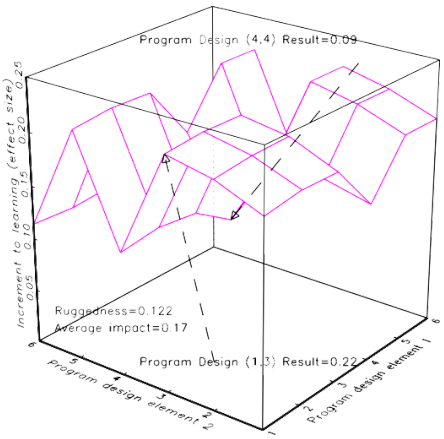


Figure 7b: Possible response surface over textbook provision design space in context A (same as Figure 3)

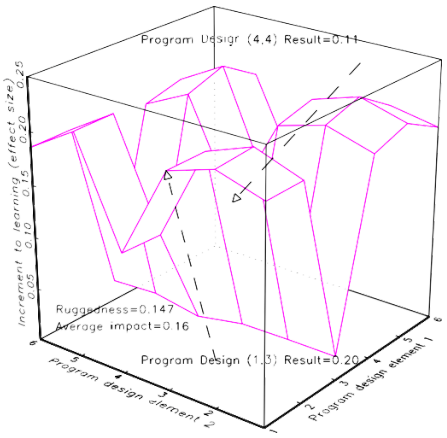


Figure 7c: Possible response surface over teaching training design space in context B (lower average impact, different shape)

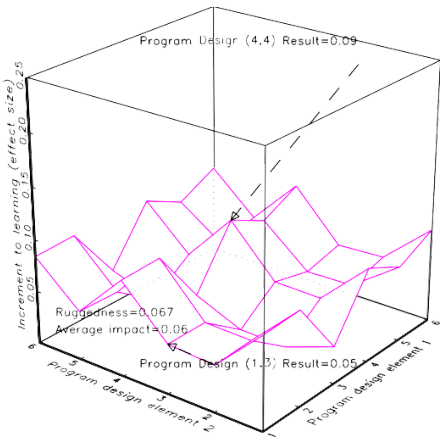


Figure 7d: Possible response surface over textbook provision design space in context B (lower average impact, less rugged, different shape)

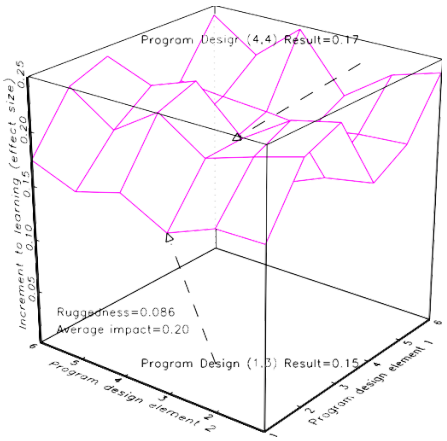


Table 1: With issues of external validity and construct validity issues the "rigorous" evidence about what is best in one contest can make the worst recommendation for another

		Teacher Training	Textbook Provision
Context A	Avg	.17	.16
	(1,3)	.22 {best of four options}	.20
	(4,4)	.09	.11
Context B	Avg	.05	.20
	(1,3)	.06 {worst of four options}	.15
	(4,4)	.09	.17

Conclusion

At the current juncture, with literally hundreds of "rigorous" studies of "what works" to improve learning in developing countries completed, and literally hundreds more on the way, it just cannot be that high value research opportunities are just funding more studies of specific interventions, no matter how clean their identification strategy. The premise of RISE is to expand the body of rigorous research, but in a way that explicitly puts each of the collection of studies into a country context. RISE will evaluate reforms at scale, with a common system diagnostic, with specific, theory grounded hypotheses about what programme/policy/project design might (or might not) work. This approach seeks to encompass the body of knowledge, contribute new knowledge, and move forward from the body of evidence to actionable, context- specific recommendations, including that "learning about learning" has to be embedded in projects/programmes/policies.

This piece was originally published by the RISE programme.

Endnotes

[1] And it is, of course, much more complex than that. The compressive strength of concrete, even made from the exact same Portland Cement, de- pends on a whole host of other factors like the ambient temperature when poured/cured, the size of the aggregate, the mix of aggregate to cement, etc. Given its importance (it is what is under your feet right now, I bet) there are entire handbooks about concrete.

[2] I semi-apologise for all the scare quotes, but these indicate reference-not use of the terms, as I will argue that rigorous evidence isn't.

[3] Since the programme design options are not assumed to be either cardinal or, for that matter, ordered (e.g., think of the teacher training example where one of the dimensions has the content of *subject matter* and *pedagogical method* and cannot be assigned a distance in any meaningful way). Therefore, there is no intrinsic meaning to what is local in the design space and the ruggedness is relative to the ordering.