RISE
RESEARCH ON IMPROVING
SYSTEMS OF EDUCATION

INSIGHTS

We Knew Fire Was Hot

by Lant Pritchett
Center for Global Development

Key Points

- The best possible methods should be used for identifying causal impacts of system interventions of whatever type, wherever possible.
- The RISE programme starts from the premise that there is little or no promise in additional research efforts that examine the impact of specific, known, inputs into education processes, like class size or textbooks.
- One needs to embed individual results in a framework, like that of Bates and Glennerster (2017), in order to understand how to extend them to new contexts.
- The emphasis should be on endogenous, context specific, systems for improvement, rather than imposition of external "best practice".

Ludger Woessmann, whose paper on testing I recently blogged about, just sent me a note with his research website. The page is now organised by topic, and in a way that is both helpful to academics trying to find papers and for more general audiences looking for overviews and basic messages that emerge from his work. This site's effort to reach both academic and general audiences is admirable, and it is worth many visits.

You can see why RISE is interested in his research:

> "***Schools...***
>
> *My research using the international student achievement tests such as PISA and TIMSS suggests that school systems play a central role why students in some countries perform much better than elsewhere.*
>
> *In particular, institutional structures such as external exams, school autonomy, competition from private schools, and tracking contribute a substantial part to the international differences in student achievement and to the equality of opportunity of the school system.*
>
> *By contrast, the role of spending, class size, and computer endowments is rather limited.*"

Source: Ludger Woessman's webpage.

His statements look a lot like the messages of RISE (Vision Notes One and Two)—and not by coincidence or correlation, but by causation. Woessmann's excellent and careful empirical research has informed what I (and many others) understand about the efficacy and impact of education around the world.

Looking over his body of research I want to emphasise one particular paper of his (joint with Martin West), which I feel has been under-appreciated (it only has 396 Google Scholar citations as opposed to over 3,000 for his Journal of Economic Literature work on economic growth). When the paper was written sixteen years ago in 2002, it already spoke directly to the likely external validity of estimates of causal impact from "education production functions" relating inputs to learning outcomes. This paper should have been the dead "canary in the mine", warning everyone that more and more randomised controlled trials (RCTs) revealing "rigorous" estimates of causal impact in particular circumstances were not going to resolve the problem of the already observed heterogeneity across contexts (countries, regions, time, subjects, grades) of the existing estimates of correlates, associations, and impacts on learning.

If “what works” is truly heterogeneous across contexts—and this Woessmann and West paper shows that even for simple inputs like “class size” and even in mostly developed country contexts, it is—then aggregating estimates of “what works” (through whatever procedure, “systematic review” or otherwise) is not going to be of much use in deciding what to do in specific circumstances.

In 2018, we have excellent pieces showing that RCT evidence lacks external validity, including Eva Vivalt’s outstanding work in general and more specifically in education. Other superb reviews of the literature include a Glewwe and Muralidharan RISE working paper and an Evans and Popova piece.

But my point here is that *we already knew that was going to be the case, both empirically and logically.* We did not need to spend millions of dollars and do hundreds of RCTs to know they were not going to add up to a single clear answer about “what works” in any given context. The lack of external validity (and construct validity, though that concern is bracketed here) of the emerging RCT evidence is not a surprise or even a “finding” that we could have only reached by doing the RCTs—we knew (in the sense of “true, justified, belief”) the eventual outcome, we just didn’t want to know it.

The paper of Woessman and West referred to earlier is *Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS* (the paper was available in 2002; I used it in my Copenhagen Consensus 2004 review). The paper uses the structure of the sampling in the TIMSS (Trends in International Mathematics and Science Study) assessment to recover a plausibly well-identified causal impact on student learning of class size for eighteen different countries.

The trick with class size (and this is a generic reason why one needs to worry about causal identification) is that the correlation doesn’t indicate causation because of the purposive behaviour of individuals. Woessmann and West use the TIMSS data to estimate the regression coefficient between student learning in mathematics and science and the size of the student’s class (this is a partial correlation with their weighted least squares [WLS] results—which I will call ordinary least squares [OLS] just for simplicity as the weighting isn’t conceptually essential—include twelve other correlates of learning besides class size). The expectation is that this (partial) association would be negative: students would learn less in larger classes, all else equal. Across the eighteen countries, that mostly isn’t so. The coefficient has the “right sign” only two times for mathematics and four times for science.

But of course, one can come up with all kinds of reasons why the behaviour of students or teachers or school administrators could lead to better performing students being in larger classes. Here are two examples of this type of causal narrative/structural model:

- Story One: Suppose administrators want teachers to have equal workloads and it is harder to teach poorly performing students. The administrators could use their knowledge of student performance to create a larger classroom with well-performing students and a smaller classroom with poorly performing students that equalised teacher workload. It could be the case that causally each individual student would do better in a smaller classroom, but the data would show a positive association of test scores and class size.
- Story Two: Suppose students (and their parents) want to be with better teachers, and better students are more skilled at getting themselves allocated to better teachers. In this case, again, the true causal impact could be that each child would do better in a smaller classroom (all else equal) but the (partial) association would show that larger classes had higher learning because of the selection effects due to the purposive selection of better students to better teachers and hence larger classrooms.

This is why people look for ways to identify the causal impact of class sizes, using actual experiments (e.g., Tennessee STAR) and “natural” experiments like demographic variation (Hoxby on Vermont) or discontinuities from class size rules (like the “Maimonides” rule in Israel). But a big question is whether the “well-identified” estimates will themselves agree across contexts.

Woessmann and West took advantage of the structure of the 2003 round of TIMSS, which included data on class sizes in the seventh and eighth grades, as well as test scores of students in those grades—usually one class of each per school. This allowed them to use instrumental variables and school fixed effects (to eliminate from the “class size” estimate any effects just correlated with a “better school”) to identify the causal effect of class size. They use the difference in average class sizes in seventh and eighth grades as an instrument for the actual size of a student’s class. One can debate whether this is a perfect instrument (and, as Angus Deaton points out, even randomised treatment is not a perfect instrument), but it is a pretty damn good instrument. It eliminates any of the selection stories involving either the teacher or the student, as neither the quality of the eighth grade teacher nor the ability or learning of the eighth grade student plausibly impacts the size of a given seventh grade class.

So, I would argue that one can plausibly think of what Woessmann and West did as doing thirty-six experiments to identify causal impact, one for math and one for science in each of eighteen countries.[1] What did we learn? In particular, what did we learn about the likely benefits to knowing "what works" from doing an RCT (or even doing many RCTs and doing a "systematic review")? The results are in Table 1.

The empirical results are a complete, unholy, mess of heterogeneity of every possible type. That implies that there would be very little to be learned about the impact in a particular context from RCTs, as the underlying heterogeneity is not at all reduced by causal impact estimates. *And we knew about this heterogeneity of well-identified estimates of inputs in education production functions across contexts in 2002—sixteen years ago.*

Table 1: It is immediately obvious from replicating "well-identified" estimates in methodologically identical ways that "rigorous evidence" of "what works" isn't at all helpful—the heterogeneity across countries (and across subjects within countries) remains massive.

Country	Mathematics			Country	Science		
	Partial association (WLS)	Instrumental variable estimate of causal impact of class size (sorted)	Difference of observational (WLS) and causal estimate		Partial association (WLS)	Instrumental variable estimate of causal impact (sorted)	Difference of observational (WLS) and causal estimate
France	**2.59**	**-2.73**	**5.32**	Czech Rep.	1.43	-1.03	2.47
Iceland	0.16	-2.59	2.75	Spain	0.19	-0.70	0.89
Greece	0.46	-1.53	1.99	Belgium (French speaking)	-0.58	-0.67	0.09
Romania	2.14	-0.30	2.44	Portugal	0.17	-0.31	0.48
Japan	3.81	0.07	3.74	Japan	2.59	-0.26	2.85
Canada	0.76	0.25	0.51	Iceland	-1.01	-0.16	-0.85
Singapore	4.69	0.45	4.23	**France**	**0.55**	**0.14**	**0.41**
Belgium (French speaking)	1.51	0.80	0.71	Slovenia	-0.39	0.29	-0.69
Slovenia	0.52	1.25	-0.73	Singapore	5.03	0.52	4.51
Portugal	0.77	1.54	-0.77	Belgium (Flemish Speaking)	1.47	1.08	0.39
Czech Rep.	2.37	2.67	-0.30	Romania	1.43	3.31	-1.88
Mean	*1.80*	*-0.01*	*1.81*	*Mean*	*0.99*	*0.20*	*0.79*
Std. Dev	*1.48*	*1.69*	*2.09*	*Std. Dev*	*1.71*	*1.19*	*1.84*

[1] One tricky issue is that their instrument just did not work in the "first stage" for a number of country/subjects and hence (as a simple consequence of the formula for the standard errors for instrumental variables [IV] estimates), the "second stage" estimates were imprecise and therefore all over the map with huge standard errors (for instance, the point estimate for math in the USA was 20.26 with a standard error of 69.2 compared to the median coefficient of -.12, the point estimate for Scotland in science was 31.6 with a standard error of 51.8). I am going to limit the results to just those country/subjects where the IV estimates have standard errors less than five times as large as those of the WLS estimates (the average was about two and a half times bigger, which is not unusual for IV).

Think of the OLS (WLS) results in the first column of Table 1 as the existing situation before there is any "rigorous evidence": all we have are partial associations. We know that these partial associations are facts about the world (in the way that any summary statistic of data, e.g., mean, variance, bi-variate correlation, is a fact about the world), but we also know they do not directly reveal causal impacts. There are, as we saw above, many causal stories (structural models) consistent with class size reductions improving learning, but under-which these partial associations could be positive. Strikingly, *all of the OLS coefficients on mathematics are positive* (larger class size is associated with higher student learning) and all but three for science—and no one believes these estimates are causal impacts.

Now, suppose we do a RCT in one country and get an estimate of the causal impact of reducing class size; for example, a RCT for mathematics in France (if we take the results of the instrumental variable [IV] as if it were a RCT—even the strongest advocates of RCTs will admit other modes of causal identification). How would we use that new "rigorous" evidence to change our beliefs about the impact of class size in the other ten countries for which we haven't done an RCT? The first point is that it is not even obvious whether we move our beliefs about the impact of class size to the *magnitude of the point estimate* of class size on mathematics in France (France [M]) or whether we should move all estimates by the *magnitude of the bias* between OLS and the RCT estimate in France[M]. That is, should we guess every country's estimate of class size impact is -2.73 (the point estimate for France[M]) or should we shift every country's OLS estimate by 5.32 (the OLS-RCT gap in France[M])?

Suppose we use the France[M] "rigorous" estimate of impact or bias, would that make our estimates of the impact of class size in the other countries better or worse?

Again, let us suppose that the IV estimates for each country/subject reveal the "true" impact. Then we can judge any given set of estimates by its root mean square error (RMSE) of prediction across all country/subjects.

The RMSE of using OLS in mathematics is 2.7. This is the "naïve" base case of how bad predictions would be if we were completely methodologically and theoretically naïve and mistook OLS partial associations for causal impacts.

If we assume that the "rigorous" estimate of -2.73 from France[M] is an externally valid estimate across these other countries, then the RMSE is 3.2. Using the rigorous evidence from France[M] gives a typical prediction error for the other ten countries about 17 percent *worse* than the methodologically naïve scenario.

If we assume that the estimate of the bias of OLS versus IV from France has external validity and therefore reduce each country's OLS estimate by 5.32, the RMSE is 4.03—now much worse (50 percent) than the prediction error of the methodologically naïve estimate.

And what if we use the "rigorous" result from France[M] to predict class size effects for science [S]?

Note the IV estimate for France in science (France[S]) is *positive* .14, so even after controlling for the selection effects using IV, the estimate is very small, but still positive. Whatever "context" means, it appears it includes not just "country", but "country/subject" as even the "rigorous" estimates for France in mathematics and science differ, by a lot. (And apparently "context" isn't "country" either as the authors report the results separately for the French and Flemish speaking regions of Belgium and their IV estimates of class size impact on science are completely different [-.67 versus 1.08]). The RMSE of the "naïve" OLS prediction for science is 1.92. Using the France[M] of -2.73, the RMSE of prediction is 3.16.

Taking a heterogeneous set of observed partial associations and "correcting" them with this one single "rigorous" piece of evidence makes the RMSE of predicting causal impact *worse* across countries or subjects.

Of course, even the less astute reader will notice I picked France[M] as it was the biggest impact of class size in the "right" direction. Naturally, one would expect the most extreme of a set of estimates not to lead to the most accurate prediction of other countries (just be "winner's curse"). But, on the other hand, this is exactly the behaviour one would expect of advocates of class size reduction, to take single a piece of "rigorous" evidence that favours their position and argue for its general validity. Moreover, if we have only done one experiment, France[M], we wouldn't know what is "typical."

According to Woessmann and West, the class size impact would have to be around 3 or larger to justify class size reductions as cost effective. Therefore, unless one uses something like France[M], the policy implication of there being some learning impact of class size reduction is not "reduce class size," as it likely isn't cost effective relative to other spending.

Three points about using the IV estimate for each country:

First, if we take the IV estimates as representing causal estimates, we see that for mathematics the variability across countries *goes up* from the OLS to the IV estimates (1.48 versus 1.69 in the last row of Table 1). So the heterogeneity of the estimates of causal impact is *larger* than of the naïve partial associations. This is important as it rules out many otherwise plausible uses of new "rigorous" evidence. That is, a common-sense idea might be to move one's "priors" of the impact in each country away from the naïve estimate towards the new "rigorous" estimate. But doing so implies the variance of the distribution of the estimates goes down. Obviously, the exercise above using France[M] for all countries reduced the variance of the predictions to zero (as all were predicted to be -2.73). However, Woessmann and West's results rule out not just this, but any uniform "scrunching" of the distribution towards the "rigorous" evidence, as that would imply a reduction in the variability of the "true" parameters that isn't borne out by the estimates (and this is even after ignoring the very high variance results).

Second, suppose we had done a "rigorous" estimate for any one country (as with France[M] above). The average of the RMSE using each country's IV estimate individually to predict the impact for all other countries is 2.07 for mathematics—compared to 2.7 for the most methodologically naïve approach using each country's OLS estimate.

Third, even if we take the "systematic review" and have done an RCT for each country, the RMSE of predicting impact in any one country is still 1.61. This is because the well-identified estimates have such variance among themselves. Given that a crude indicator of whether class size reduction is cost effective has a value of 3 (with estimates of this scaling), a RMSE precision of 1.61 is not very helpful.

This means there is nothing particularly special about "well-identified" estimates from a given context. The 1999 Angrist and Lavy paper on the Maimonides rule finds a positive impact of reducing class size on learning and has over 2000 citations, even though it was only about a small country of 6.3 million people. The Woessmann and West paper completely encompasses the Angrist and Lavy work by showing that, whatever is true of the impact of class reductions in Israel, it has little or no bearing on predicting any other country—but Woessmann and West have only about 400 citations. Hence, any use of the Angrist and Lavy results outside of Israel (on the premise that these empirical results on class size are especially valuable outside of their context because their estimates are "rigorous") ignores the fact that rigorous evidence isn't.

The reader may note, correctly, that I have made this point about the lack of external validity before. Pritchett and Sandefur (2013) notes that since both the magnitude of causal impacts and the magnitude of bias in the estimate of causal impact are dependent on parameters that are likely to vary across context, there is no logically coherent argument for external validity of estimates of causal impact. Pritchett and Sandefur (2015) showed that RMSE of naïve "own" estimates was actually better than using "rigorous" estimates of causal impact from another context. And I have previously pointed out Vivalt (2016) (and other indications) suggest a lack of external validity (and/or construct validity) in RCT estimates.

But my current argument is much stronger. My argument is that *by at least 2002* any fair person should have been convinced that RCTs could not, and would not, produce estimates with external validity—at least about impacts of inputs into "education production functions". While one might say we have learned that there is a "generalisability puzzle" or that RCT estimates lack external validity, one does worry a bit that learning fire was hot was by touching it. Yes, one knows fire is hot in a deeper and more directly experiential way by touching it, but this is still a confirmation of what was already known than a deep new insight.

What is the implication for research moving forward, particularly for the RISE research program? I think there are four important takeaways:

First, there is no sense in which I am arguing against using the best possible methods for identifying causal impacts of interventions of whatever type, wherever possible. The RISE research programme includes a number of RCTs to assess the impact of various interventions—but more "system" interventions than "input" variation (like assessing the introduction of school improvement plans to improve school governance in Madhya Pradesh, India)—and we avidly anticipate finding out what these methodologically sophisticated estimates of casual impacts reveal.

Second, the RISE programme starts from the premise that there is little or no promise in additional research efforts that examine the impact of specific, known, inputs into education processes, like class size or textbooks.

This literature has been reviewed, meta-reviewed, meta-meta-reviewed and what can be learned, has been learned, and one thing that has been learned (early, then often) is heterogeneity.

Third, as suggested by the RISE paper of Muralidharan and Glewwe (2015) and more recently by Bates and Glennerster (2017), one needs to embed individual results in a framework in order to understand how to extend them to new contexts. Bates and Glennerster (2017) suggest using the following four questions to frame this process:

Step 1: What is the disaggregated theory behind the program?

Step 2: Do the local conditions hold for that theory to apply?

Step 3: How strong is the evidence for the required general behavioral change?

Step 4: What is the evidence that the implementation process can be carried out well?

These four questions can be usefully applied to system level interventions in the RISE 5 by 4 conceptual framework. For example:

- The creation and sharing of additional information about student performance, which change the "information" in the accountability system.
- Different approaches to training teachers, which change the "support" in the accountability relationship
- New methods for teacher performance evaluation, contracting or compensation, which change the "motivation" element in the accountability system
- Approaches for teaching at the right level, which change the "delegation" in the accountability system.

The four questions that Bates and Glennester outline are helpful for assessing each of these.

Fourth, and perhaps most importantly, is the emphasis on endogenous, context specific, systems for improvement, rather than imposition of external "best practice". Abstract, codifiable, "knowledge" from the outside about "what works" is not the key to country success. Instead, success will flow from the creation of education systems that themselves are motivated to achieve learning and in which the system is built to learn how to improve their own performance and expand their own capabilities—including the capability to use rigorous research techniques, embedded in local context, as part of a country driven learning strategy to generate, evaluate, and scale effective practices.

Lant Pritchett is the RISE Research Director and senior fellow at the Center for Global Development. Pritchett has published two books with the Center for Global Development, Let Their People Come (2006) and The Rebirth of Education (2013), and over a hundred articles and papers (with more than 25 co-authors) on a wide range of topics, including state capability, labour mobility, and education, among many others.

Please contact information@riseprogramme.org for additional information, or visit www.riseprogramme.org.

RISE is funded with UK and Australian aid from the UK and Australian governments.

UKaid from the British people

Australian Aid