Chapter 4: More of the same won’t work

The easiest thing for any movement to do is to ask for more. Especially more of the same, as more of the same provides more to all of the existing interests without demanding any disruption or change. More is an especially easy for a successful movement. A beauty of the drive the universal schooling was that more really was an answer. If a school has just been built here then building, equipping and staffing another school pretty much just like it over there is the obvious next step. Universal schooling really does require universally available schooling and, while there might have been other ways to do it, just building out through replication was one way. And, as we saw in chapter 1, it really worked.

It would be nice if goals for universal education could be solved exactly the same way as universal schooling: more, more, and more of the same, and eventually we get there. If this were true then exactly the same coalition of advocacy, altruism, and self-interest can just ride the same system from a grade attainment based schooling goals (like the MDG) to a learning achievement goals, a new Millennium Education Goal—with no need for any innovation, disruption or change from the existing access axis.

“Wouldn’t it be nice” is not a plan. In this chapter I marshal the existing empirical evidence to show three things. First, no developing country currently has an evidence based plan for achieving significant progress in education—in spite of the fact nearly all of them have plans to improve “quality.” Second, no “more of the same” plan based on intensification of existing inputs is going to be the major part of meeting an education or learning goal, “system change” innovations are needed for countries to reach education goals. Third, copying the educational fads that come out of rich countries is not going to work: these countries are just not working on the same problem.

I. All countries plan to improve quality, but no country plans for learning

To expand enrollments education systems needed more schools, more inputs and more people trained as teachers. So the natural extension would be that to improve “quality” one needs “better” schools with “better” defined as more infrastructure (e.g. more space, more rooms, playgrounds, toilets for students), more inputs (e.g. more chalk, more learning materials, more textbooks), more teachers (e.g. smaller teacher to pupil ratios), and more training of teachers. This creates a plan for improving the “quality” of education that is pure logistics.

A common approach to school quality improvement, one which illustrates the quintessentially “high modernist” (Scott 1994) approach, has been to create an Electronic Management Information Systems (EMIS). Using new information technology these aim to create and centralize information about schools. This centralization allows a plan for school improvement is a list of the inputs for quality schools (e.g. buildings, toilets,
classrooms, books, desks, trained teachers in desired proportions to number of students). The EMIS system keeps track of all of those school by school in real (or at least realistic) time. Progress is defined as schools having more inputs and success is reached when every school is a "quality" school. This approach sees only the "EMIS visible"—those aspects classrooms, schools, and teachers that can be coded into bits and bytes.

The internal logic of these "quality" improvements is circular and hence unassailable. School quality is defined in EMIS visible terms, the EMIS system tracks the allocation of resources to inputs, and the goals of school quality are therefore reached when the EMIS system records that schools have the infrastructure and inputs. Success is more or less guaranteed as either budget improvements lead to measured input improvements which, in the absence of any objective measurement of education output, success is achieved when the inputs are in place or budgets are not available, in which case the EMIS system creates better and better documentation of the "quality" deficit.

Suppose the EMIS-visible agenda was successful. How much closer would a country be to learning outcome goals? Think of improving learning improvement as building a tower, brick by brick. How many bricks we will need depends on the size of each brick compared to the height of the tower.

How tall is the tower that needs to be built? We have seen if developing countries want to reach anything like developed country levels of learning achievement—either to reduce the number of their citizens who are below a minimal learning target, or to improve the average quality of education, or increase the number with global capability—they need massive improvements. The easiest way to fix magnitudes is to use a student standard deviation because this is at least somewhat comparable across test instruments while any absolute scoring (e.g. from 1 to 10 or centered on 500 or "percent correct") is a conventional normalization. As we saw in chapter 1 nearly all developing countries (including many middle income countries) require something very near an internationally comparable (PISA or TIMSS) student standard deviation worth of improvement.

How thick are the blocks? Suppose we divide the "EMIS-visible" input expansion possibilities into four building blocks:

---

1 This might seem a caricature, but it is a common description. For instance, in 2009 India’s national parliament passed a law guaranteeing each child the right to an education in a "quality" school, but the only legally actionable elements of a "quality" school are the available, EMIS visible" inputs.

2 This is in obvious homage to James Scott’s landmark “Seeing Like a State” (1998) which describes many “schemes to improve the human condition” based on making the complexity of life “legible” to nation-states which rely on "bureaucratic high modernism" for implementation.

3 Even the “student standard deviation” is not comparable across tests of different construction if the tests are not properly calibrated. For instance, if a test is too easy for the tested population then nearly everyone may get a perfect score and hence a very low standard deviation or conversely if the test is too easy everyone may get a very poor score and, again, a low standard deviation. Part of the science of constructing an instrument is to produce a range of difficulty relative to the tested population to get sufficient variation to capture both low and high performance.
The thickness of these bricks is the simple product of two features.

The effect size, which is the impact of expanding the block on the learning outcome—better buildings, more textbooks, smaller class sizes, more trained teachers—measured (hopefully) in common student standard deviation units. So, think of two students who are finishing their 8 years of basic education, one of whom had a trained teacher all 8 years and one of which had an untrained teacher all eight years. How much higher will the learning outcomes for the student consistently exposed to the trained teachers?

The second feature is the scope for expansion of the input. Often there is a practical upper limit either to the input or to the range over which the input has a large impact. So, nearly everyone believes that moving from no available textbooks to shared textbooks can improve performance and from share textbooks to one per student might improve performance over shared textbooks, but almost no one expects the same magnitude of learning impact from each student having their own textbook to each student having two textbooks. So the expected achievable scope for learning gains from textbook expansion is from none to one per student. Other input measures also have a natural upper limit, having a leaky roof might inhibit learning but once the roof doesn’t leak it seems hard to picture that better and better roofs improve learning outcomes. Limits on the achievable scope might come from practical considerations like cost, so that class sizes might be reduced from 40 to 30 but a reduction from 40 to 12 might be far beyond the practicable cost.

Take a simple example. Suppose the effect size on a measure of learning achievement of a child being exposed to a “trained” versus “untrained” teacher is .2 (remember “effect size” is automatically in student standard deviation units). What is the feasible gain to the average student of training teachers? If say, 70 percent of teachers are already trained then the feasible additional gain from having teachers universally “trained” is exposing an additional 30 percent of students to training so that gain is a .2 effect size gained for each student exposed times the .3 incremental students that will be newly exposed to trained teachers, which means the increased in the average would be .2*.3=.06, hence in this super simplistic example the gain towards the goal of an increase of one SSD would be a brick of thickness .06.
An evidence based plan for achieving a learning goal target based on input expansion would have three simple parts:

- A clear statement of the goal which can be: (a) in any learning domain or skill/competency/functioning—from basic literacy to working in teams to thinking creatively and (b) about any part of the distribution of performance in that goal—e.g. fraction of students above a certain minimal level, performance of the average student, students in the global ranks.

- A judgment of the “effect size” that expansion of the inputs would have based on a fair reading of the available and relevant empirical evidence.

- A feasible plan, in budgetary and logistical terms, for the expansion of the inputs that would reach this goal.

One possible schematic of a learning goal plan is illustrated in schematic terms in Figure 4.1. The goal might be to gain one (internationally comparable) student standard deviation in mathematics performance. The plan would be individual elements in which the thickness of each brick is determined by an empirically estimated “effect size” plus a planned expansion of the input (where each of the four illustrated bricks may be built up of even more specific components detailing the individual learning inputs or specific training programs or qualifications or specific infrastructures).

No developing country in the world has a feasible plan for accomplishing any significant learning goal through an expansion of known EMIS-visible inputs.

This is not to say countries don’t have plans to expand inputs, of course they do. But plans to expand inputs lack one or more of the features that would make for a real plan for education.

First, very few countries have a goal, a clear statement of the magnitude of the improvement in learning objectives that will be achieved by their plans for input expansion. Rather, plans are entirely circular: quality will be improved when inputs are expanded because inputs are quality.

Second, no country has plan that has a link between their plans for input expansion and their learning objectives based on any evidence based assessment of the effect sizes. While learning objectives are often mentioned the magnitude of the learning gains are not quantified—if you are a kilometer from your goal whether your planned steps will take move you one centimeter, one meter, or a hundred meters closer makes a big difference. So while some countries might even pay some attention to evidence of which inputs are “effective” in moving some distance, none pay attention to magnitudes.
I am deliberately making claims in order of increasing complexity and hence also potentially controversy so that the reader can accept one without necessarily accepting the next, more controversial claim. My first claim is just that countries as of yet simply do not have plans to meet ambitious learning goals. Not that they have “bad” or “wrong” or “inadequate” plans, they just don’t have anything that even minimally qualifies as a plan. They have plans for expansion of schooling and they have plans for the expansion of inputs but they don’t have realistic plans for increasing the learning profile of students in schooling to meet learning goals.

II. The learning gap cannot be filled with inputs

I want to argue a much more controversial position. I argue that the reason no country has an evidence based EMIS-visible input expansion plan to meet significant learning goals is that you cannot have an evidence based plan because the evidence says the bricks are too thin to build the tower. The empirical question is:

Will expansion of known EMIS-visible inputs will be roughly sufficient (with perhaps some minor contribution from system changes) to achieve learning goals?

or
Are the building blocks from input expansion are just too thin to add up and major gains will have to come from system or structural changes with either (a) produce more with existing inputs or (b) accelerate the process of discovering and systemically adopting learning achievement improving input.

**Figure 4.2: Will expansion of known inputs be enough?**

Of course the answer is going to be different for every country and in every study, but I argue that, broadly speaking, the overwhelming bulk of the empirical evidence suggests input expansion will constitute only a small part of countries’ progress towards meaningful learning goals. I draw on three literatures:

- Studies that compare student learning across schools with different levels of inputs (whether using natural variation, quasi-experimental or experimental methods)
- Studies that compare learning achievement and expenditures at all levels: schools, districts, states/provinces, or countries.
- Studies that track the evolution of learning achievement performance over long-periods.

**II.A) Education Production Functions: Thin bricks can’t make a tall tower**

There is a simple arithmetic truism: zeros, even lots and lots and lots of zeros, don’t add up to one. Pushing harder on a string won’t help.
There is a massive literature, both published and unpublished, investigating the links between educational inputs and learning achievements\(^4\). Studies have compared learning performance and exposure to inputs across students, across schools, across districts, across states/provinces within countries, across countries. Many of these studies have been on non-experimental data that use statistical procedures to try and eke out the causal effects of inputs, all else equal. More recently there have been a number of rigorous randomized experiments which, along the dimensions they have explored and across the units they have been able to explore (mostly student or school) give internally valid estimates of the impacts of inputs.

The basic idea of all these studies is to compare the learning achievement of students over time, some portion of the learning profile, either by tracking students over time (which is rare) or by comparing student achievement at a point in time. The learning profile for any given student might be steeper if the student is in a learning environment with more rather than less inputs. How big are these effects? Are they big enough such that a plan based on EMIS-visible inputs could potentially steepen the learning profile sufficiently that countries could pursue a “more of the same” path to reach learning goals? How much higher could a student’s *cumulative* mastery be from exposure to higher levels of schooling inputs?

\(^4\) I want to stress again that while most of the literature is relating educational inputs to learning achievements that are easy and low cost to measure (like reading and mathematics) using these studies does not imply a “back to basics” approach that asserts these are the only or even most important elements of schooling. The process of education has a large number of important goals, both in socialization and in competencies and skills sets which are not well measured on the typical standardized exam.
Any estimate of the learning impact of schools or school inputs must take into account the many other factors that affect any student’s outcome such as the student’s family background, motivation, innate abilities, etc. One can only to estimate the incremental impact of schools “all else equal.” Nearly all empirical studies find that the “all else” about students matters, a lot. The single most reliable empirical finding in the “education production function” literature is that student performance differs enormously, even among children with nearly identical schooling experiences, and that these differences are robustly associated with both student and parental background characteristics. It should surprise no one that children of better off, more educated, and more achievement oriented parents tend to perform better in school than other children. This is important to keep in mind as, if these student background characteristics matter (and they do) and students with similar background characteristics tend to cluster into schools (and they do, either because of patterns of residence or school choice) then there will be large differences in the observed performance of schools that have nothing intrinsically or causally to do with the school’s performance.

Let me start with three examples which are typical of the massive existing “education production function” literature.
A first example is a study from Jamaica which combined economists and sociologists of education and so did a careful job of measuring not just “inputs” but also pedagogical practices, school organization, and community involvement (Glewwe, Grosh, Jacoby, and Lockheed 1995). The included fourteen input measures (about the school physical facility, availability of instructional materials, student to teacher ratios, teacher qualifications and training, and school-level pedagogical inputs) are all components of most EMIS-visible plans for school improvements.

First, in the estimated associations for mathematics achievement, seven were positively related, seven were negatively related, and none was statistically significant at the conventional 5 percent level. This is exactly what one would expect from pure randomness. For reading achievement, eight were positive, six were negative, and one was statistically significant (at the 5 percent level)\(^5\).

Second, the “effect sizes” even for many of the variables that are “statistically” significant are small, although in this study some were as large as .2. For instance, the study found that the percent of students in the school with desks was associated with higher reading scores but the effect size was only schools with an “adequate” number of desks versus “not adequate.”

Third, the “scope” for intervention—the extent to which an intervention could be pursued—even for the items that were positively associated was often limited. So, for instance, the study estimated the impact of textbooks arriving two months or more late, with less learning in both math and reading. However here the “scope” was to get the textbooks on time and hence was a “once off” improvement, not something that could be continuously improved to reach higher and higher achievement. Similarly, the numbers of students who have desks has a maximum of 100 percent—and 85 percent of students already did have desks. So the maximum gain from moving to the best possible value of this is the gain of the effect size times the maximum scope is about one tenth of one standard deviation for reading (less for mathematics).

What if we used this study and say, suppose we achieved the input fantasy—all of the inputs were at their best possible values. What if all teachers had diplomas, all teachers had training in the last three years, all students had desks, class size was 25, all schools had reliable electrical services—everything. How much of a one standard deviation gain (on this particular test) in math and reading (averaged since these are school or classroom, not subject specific) would this incredibly difficult and costly achievement in the expansion of inputs produce? The puzzling (but not uncommon) result is that the predicted gain would be modestly negative—as the magnitude of the inputs with negative estimates is slightly empirically larger than those with positive estimates. This anomalous negative result is principally the result of one odd result, so let’s throw that out as a statistical fluke, even with this biased estimate the answer is about .10. So a massive program of input expansion, a reform agenda that would be

---

\(^5\) This was also the case for the “pedagogical processes” as of the ten measures only one was significant at the 5 percent level, while for reading three were, so this is not just true of “inputs.”
fiscally very costly and take a decade to achieve would be predicted to have (generously) about a tenth of student standard deviation impact.

Another empirical study measured learning achievement of grade 4 students in Sri Lanka in three subjects of Mathematics, English, and the child’s first language (Aturupane, Glewwe, Wisniewski 2006). As “inputs” the study measured of not only of student performance but also had very detailed information about the children’s own background (e.g. parent’s education, resources in the home, participation in private tutoring) and very detailed information about the school (more below). The study then used standard statistical techniques to estimate the association between the three learning achievement measures (Math, English, Language) and child, household, and school level variables.

First, this study has a rich array of variables intended to measure school quality, including infrastructure variables, teacher training, characteristics of teachers, principals, etc. All told there were 35 input variables that nearly everyone would have agreed were “inputs” into a “quality” education. The study dismissed most of these in one telling sentence: "variables with no explanatory power in any of the tests are dropped." Of the 35 input variables 28 were never statistically significant in any specification for any learning measure (not Math, not English, not native language). Only one-fifth (seven of 35) of the “EMIS-visible” inputs had any demonstrable association with any learning outcome. You can add up all the zeros you want and still get nowhere.

Second, of the seven variables that were included in the “final specification” (because they were statistically significant in at least one specification for at least one learning outcome) the magnitude of the “effect sizes” was empirically tiny (and inconsistent). For instance, it is commonly asserted that more experienced teachers are superior and that retention of teachers a policy priority and this justifies paying more to teachers with more seniority. Their simple (ordinary least squares linear regression) estimates were that an additional year of teacher experience was associated with less than .01 effect size improvement.

Third, even where there were strong associations the available scope was modest. The largest impact of any measured school characteristics was a measure of whether all students had adequate desks (so either zero or 1 for each school). Bringing the others up to that standard would (if the effect were causal) add 8.1 points to student scores, so the effect size is substantial. But, there was a limited scope for this variable to improve performance as 58 percent of schools already had adequate desks. The gain from moving to 100 percent of schools is 8.1*.42=3.4—which is good, but we are looking for a gain of 100 points so this will at best get a tiny part of the way.

Suppose we again calculate the EMIS-visible fantasy plan—that each of the seven input variables are expanded to their maximally feasible scope—what is the gain? (And keep in mind this is already biased by the exclusion of the variables which were not

6 This study also had measures of pedagogical processes based on actual observation of teachers in the classroom.
statistically significant as some of those might have been positive, some negative). The total is .177, which is undeniably something, but also undeniably not all of the way. Of course, that estimate relies on the “point estimates” which were each themselves subject to great uncertainty. What are the plausible upside and down risks of an input expansion? Well, the result is that each of the impacts was on the low side (one standard error below the point estimate) then the total possible scope for input expansion would be only .026—almost nothing. This same calculation on the upside is that even if the impact of every single variable was larger than the point estimates then even the maximum possible EMIS-visible expansion at uniformly optimistic assumptions about impact produces only about a third of the desired impact.

As final examples, I take studies of two states of India, Rajasthan and Orissa which examines the connection between standard inputs like teacher salary, qualifications, training, class size, and whether schools have multi-grade teachers. Again the same experiment of having the maximally feasible expansion of these inputs—eliminating multi-grade schools, having all teachers as graduates, all class sizes smaller by 10 students for each inputs times its effect size. The results are total gains of .12 to .13 SSD—again, substantial, but nowhere near 1.

**Figure 4.4: Empirical estimates of the total impact of even ambitious, across the board, improvements in inputs on learning outcomes are small**

These are just four specific studies in four specific regions, but I could do this all day (and have on many days, e.g. Pritchett 2004) as these studies are typical of an enormous literature of “education production function” studies. Unfortunately this application to the feasibility of the scope for overall improvements is not the way this literature is usually reported or summarized. As with much academic literature there is an unfortunate obsession, some would say fetish, with “statistical” over “practical”
significance. Moreover, there is rarely any attention to the “scope” for expansion of a particular input so that the total feasible achievement gain is rarely explored.

II.B) Why the published literature proves the opposite of what you think

I can hardly hope to have convinced the advocates of the access axis of the dubious wisdom of “more of the same” by the few examples above, nor even perhaps yet the fair minded reader. The reaction is “but what about the studies in countries X, Y and Z that conclusively showed that inputs A (class size), B (teacher qualifications) or C “matters”? They were published in academic journals and hence subject the rigorous refereeing. Why shouldn’t we believe those?” I deliberately used examples from “old school” regressions and studies not published in academic journals at all because the way the academic literature is published and read make it subject to enormous bias.

I use a simple simulated scenario to show that it is possible to have many published studies showing inputs, say class size or teachers’ qualifications, are associated with better student performance even if it is never true. Simulations are useful for illustration because in a simple artificial world of a simulation there are no mysteries: it is easy to see exactly what is happening and why.

In my “world” I simulate “learning outcomes” for 1000 “students” as random numbers drawn from a standard Normal distribution so in this “world” nothing causes student outcomes. Four “inputs” are generated, also as completely random numbers. Using this data on I estimate an “education production function” by associating with a multivariate regression student “learning outcomes” with the “inputs.” In the simulation I create many such studies, each with exactly the same draw of random numbers as “outcomes” and “inputs.” Sometimes, just by chance, an input with be “statistically significant” as the standard statistics will reject the hypothesis that the association is zero. In fact, a conventional standard of statistical significance of a ten percent two sided test will produce a false rejection of the hypothesis that the association is zero in about once for every twenty studies7. So 100 “education production function” studies will typically produce 5 that show an input matters in being “statistically significant” even when we know in this simulation world for sure that this is false.

The problem comes when these features of statistics are combined the way that social science happens in practice (and a fair amount of non-experimental science from data, like studies in medicine or health) from study design to publication. Most non-experimental studies start off by generating data about a large number of variables that are potentially associated with learning—features about the school, the classroom, the teacher, the student.

The first bias happens when researchers then analyze the data. Suppose the examination of the data does not reveal any statistically significant associations on only associations with “unexpected” sign. Most researchers, who are busy people with other

7 Again, if you are confused about why this is so, see the statistical appendix.
things to do, do not even bother to write the empirical results up into a study. They just move on. This creates a “write-up” bias which no review of the literature, however comprehensive, will be able to find these negative results as the papers just are not out there.

Second, suppose some input variables are statistically associated with learning outcomes, this leads to the second bias: what does the title, abstract, and write-up make the paper “about”? Obviously it is much more interesting to write about the findings the researcher does get. In the study in Sri Lanka for instance, there is no table showing the empirical results for the 28 of the 35 variables that were never statistically significant. In spite of the fact this data was generated and analyzed the results are never reported (other than to say they were never significant in any regression). After data mining researchers prefer to show the gold and hide the tailings—hiding all those just did not pan out.

What is often thought of as “publication bias” is actually the third in the chain of biases (but which affects the previous behavior). For very good technical reasons, scholarly journals are reluctant to publish articles about those variables that were found not to be associated with learning. This means that studies without “statistically significant” results will have a hard time making it from “study” to “publication.”

In the simulated “world” I add an agent I will call a “researcher” who combines the biases outlined above. Suppose now that if none of the four “inputs” turns out to be significant there has been a “study” but no published paper is generated. But, if one (or more) of the inputs turns out to be statistically significant then the researcher publishes a paper “about” the significant input.

Now, what is the result of this computer simulation? The result is a “published” literature with apparently compelling proof that inputs matter. Table 4.1 shows the results for a simulation in which there were 100 “studies” (which is just one run, but repetitions of the simulation for another 100 studies produce broadly similar results) each with 1000 “students.” Column 6 shows the number of “published” studies “about” each input that report a positive and statistically significant association. For Input A there happen to be 9, but there are also 6 for Input D and 4 for Input C. What more conclusive proof could anyone want that Input A is an important part of a plan for quality education than 9 articles in refereed academic journals showing its association with learning, which is almost half the entire published literature. Moreover in all the published literature 71 percent of papers show an association that is positive (even if not statistically significant)?

But since we control this simulated “world” exactly we know for sure that every one of those published studies is wrong.

---

If “low statistical power” and “attenuation bias” mean nothing to you and you are not intrigued read on, if they do mean something to you then you know what I mean, and if they don’t mean anything to you and you are intrigued see the Technical Appendix section “Why failing to reject will lead to rejection.”
Table 4.1: Randomness plus statistics plus publication bias produces apparently compelling evidence of the completely false...in this simulated experiment inputs do not matter at all (by construction of the simulation) and yet there is a substantial empirical literature about each input showing it does matter

<table>
<thead>
<tr>
<th>Inputs into learning quality to be studied</th>
<th>Total number of simulated studies</th>
<th>Fraction positive (theoretical expected=50%)</th>
<th>Fraction positive and statistically significant (theoretical expected=5%)</th>
<th>Number of &quot;published&quot; studies &quot;about&quot; the input in which the given input (row) is positive and statistically significant</th>
<th>Proportion of all &quot;published&quot; studies in which the input (row) is positive</th>
<th>Proportion positive of the input (row) in &quot;published&quot; studies &quot;about&quot; other inputs (theoretical prediction=.05)</th>
<th>Estimated &quot;effect size&quot; of each input from reviewing the &quot;published&quot; literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input A</td>
<td>100</td>
<td>59.0%</td>
<td>9.0%</td>
<td>9</td>
<td>71.4%</td>
<td>50.0%</td>
<td>0.061</td>
</tr>
<tr>
<td>Input B</td>
<td>100</td>
<td>50.0%</td>
<td>2.0%</td>
<td>2</td>
<td>71.4%</td>
<td>68.4%</td>
<td>0.058</td>
</tr>
<tr>
<td>Input C</td>
<td>100</td>
<td>51.0%</td>
<td>4.0%</td>
<td>4</td>
<td>57.1%</td>
<td>47.1%</td>
<td>0.064</td>
</tr>
<tr>
<td>Input D</td>
<td>100</td>
<td>47.0%</td>
<td>6.0%</td>
<td>6</td>
<td>61.9%</td>
<td>46.7%</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Estimated impact of expansion of all inputs from the "published" literature (sum of the effect sizes of each) 0.24
Average impact of expansion of inputs from all country studies (no bias) 0.00

Maximum estimated impact from any of the 100 country studies 0.14

The fourth bias is “reader” or “reviewer” bias. Someone who “reviews the literature” by reading papers about a given input, say class size or teacher qualifications, will produce an entirely different result then if they looked at all studies. So, in this simulation all papers about input A show positive and statistically significant results (a feature built into the simulation). But what if we read the papers “about” inputs B, C and D? Then we see that input A is positive about 50 percent of the time and negative about 50 percent of the time—exactly what should be true of the random numbers that were generated in the simulation.

This “reviewer” bias also affects the magnitudes of gain to be expected from expanding inputs. Suppose in our simulation we think of each study as a “country” with its own study. A country could either form a plan based on its own study of the four inputs or the country could commission a review of the “published” literature. The average of the “published” estimates will be about .06 due to the publication bias. The predicted impact on learning of a one unit increase in each input would be .24. But we know this is pure artifact. If instead each country used their own study they would, on average, get it right—the average impact in the simulated world would be zero. In fact, reading the literature “input by input” would result in a larger mistake on the predicted impact than any country reading its own study would make (the “biggest” country result is around .15).
Simulations are not evidence about the real world, but do show exactly what happens in a simple world under our complete control. With just a single rule about moving from “study” to “published” literature this simulated world in which inputs do not matter at all, by construction, has a large published literature showing the importance of inputs, from which one could form seemingly “evidence based” plans for input expansion.

The only way around the problems is to meta-research which summarizes the results of all available studies—which is not a complete solution as one cannot find the studies which were never written up nor the results on variables whose results are not reported due to data-mining, but at least it is something. Professor Eric Hanushek of Stanford University has been doing this now for decades both in the USA (Hanushek 1986) and reviewing the literature from developing countries (1995). His summaries of the existing studies finds exactly what one would expect if inputs are only loosely related to outputs—that one is almost exactly as likely to find that inputs have negative associations as that they have positive associations and that the number of “positive and statistically significant” results is not far off from what one expect from pure randomness.

II.C) Why not take “the best” estimates of the impact of inputs?

“Why not just use the ‘best’ studies and ignore the rest?” Studies of the type discussed above that rely on non-experimental data can establish associations but have a very difficult time establishing causality because the world is full of people, and people are very, very complicated and behave in purposeful and sophisticated ways. When we observe schools with good learning outcomes also have students with books it is difficult to disentangle whether books cause learning or that students who are capable and motivated to learn are more likely to go out and get books. Especially when parents and students can explicitly or implicitly choose their school and classroom the associations between school characteristics and learning outcomes cannot be interpreted causally because it may be that the most ambitious and motivated parents and students seek out teachers with the best qualifications so that there will be a strong association between teacher qualifications and student performance, even if the actual causal impact of teacher qualifications is small.

This has led, particularly among development economists, to a re-emphasis on the use of “natural” or “controlled” experiments to investigate the impact of various potential interventions on learning outcomes. These newer, more rigorous, studies have even further called into question the conventional wisdom that might support known input expansion as a strategy for school improvement. This use of rigorous testing of innovation is in fact key to creating improved learning—and the next many chapters will draw on many insights from a variety of recent experiments (and other rigorous methods of estimating causal impacts).

However, this approach is not the solution to an EMIS-visible input plan for meeting learning goals.
First, while these methods are terrific for establishing estimates of causal impact with “internal validity”—that is, less contaminated with the many sources of potential bias in non-experimental data—they have no, and claim no, general “external validity.” For instance, the STAR (Student/Teacher Achievement Ratio) experiment is a massive achievement. There is no question that if one wants to know what the causal impact on learning of reductions in class size from small (e.g. 25) to very small (e.g. 15) sizes in the early grades in an educational system like Tennessee (e.g. levels of other available inputs, quality and motivation of teachers, administrative capability, legal background and framework, citizen and parent systems of engagement and accountability) there is no estimate with a better claim to internal validity than the STAR experiment.

So perhaps, as one of my colleagues once asserted, “one good study is worth a thousand bad regressions” and hence we should just throw out all existing regressions and just use the world’s best experiment. Certainly, if we were measuring the gravitational constant or proton mass, we should. But we are not. Suppose I wanted to know what the impact of an (a) across the board reduction in class size in all grade, (b) from an average of large (40 plus) to an average of medium (30) in Kenya or a state of India, or Indonesia or Pakistan. Would I be better off with an estimate from non-experimental data or from the Tennessee STAR experiment? Who knows? Certainly a general claim that experimental estimates from a different context (e.g. Tennessee or Israel) are “better” estimates of the causal impact of a change in inputs than even the simplest possible non-experimental estimates from the same context (e.g. Pakistan or Kenya) has no scientific basis, in fact is just pretty goofy.

On certain basics like class size or teacher to pupil ratios there are already enough experimental or well-identified estimates from different contexts to know, from the various experiments themselves, that the estimated impact is not constant across contexts. To exactly no one’s surprise, the estimates of class size reduction from experiments in Kenya or Pakistan or India are not the same as those in Tennessee or Israel. Experiments themselves have already established limits on “external validity” of experiments, a finding that surprised exactly no one.

Second, while rigorous and randomized evaluation of innovations will be key to achieving high learning outcome systems in developing country contexts, they are not going to be the basis for input based plans because they build estimates of the range of policy relevant factors so slowly. The reason randomization works as a technique is reduces the biases from the many other determinants of learning, not by estimating them, but by balancing them between treatment and control (if this isn’t clear see the discussion of randomization in the Technical Appendix).

III. How much will learning goals cost? All you have—or nothing at all

I worked for many years at a large development agency. At one point a colleague of mine was asked to be part of a team to provide an estimate of the incremental cost of reaching the Millennium Development Goals. He was asked how long it would take him to estimate achieving one of the goals. His response was, “I can tell you right now. The
evidence suggests there is zero connection between the money spent and the outcome and mathematically anything divided by zero is infinity so the answer is the spending to achieve the goal is infinity.” This was regarded as a not particularly helpful response, but he did not more work any more on this clearly pointless question. To his surprise, there was a next meeting at which he was asked how his work was coming. He said, I gave you the only answer the evidence supports the first time, but if you want a different answer I can tell you not only how much one MDG will cost but how much the incremental spending to achieve all the MDGs will be. My answer is that since this “costing” exercise is really an exercise in advocacy to expand foreign aid and since foreign aid is currently about 50 billion, then asking for less than doubling of aid is too little and asking for more than doubling is unrealistic, so my second estimate of the incremental aid to reach all of the MDGs is 50 billion. He then got his wish and was no longer asked for his views. Many weeks later, after much evidence was reviewed and data crunched, the committee issued its report estimating that the incremental aid needed to meet the MDGs was 48 billion.

What are the basics of the cost of improving learning? This is going to take a bit of arithmetic and a few definitions, but there is no “theory” involved at this stage, this is just the inevitable arithmetic of accounting. To make the arithmetic easy we’ll assume a linear relationship between learning outcomes and the four types of EMIS-visible inputs: infrastructure (IF), learning inputs (IP), teacher training and qualifications (TTQ), and Teacher to Pupil ratios (TPR) so that the expected learning outcome (L) gain from a plan of expansion of these inputs is:

\[ L_{Plan} - L_{Base Case} = \beta_{IF} * (IF_{Plan} - IF_{Base Case}) + \beta_{IP} * (IP_{Plan} - IP_{Base Case}) + \beta_{TTQ} * (TTQ_{Plan} - TTQ_{Base Case}) + \beta_{TPR} * (TPR_{Plan} - TPR_{Base Case}) \]

where the \( \beta \)'s for each input are scaled in effect sizes.

The cost of this plan is just the sum of the cost of expanding each inputs which is just the magnitude of the change in each input times the unit cost, averaged over the range of expansion. This expression is simplified but could be built up directly from whatever level of detail of specific infrastructure or inputs or training is needed):

\[ Plan Cost = (IF_{Plan} - IF_{Base Case}) * UC_{IF} + (IP_{Plan} - IP_{Base Case}) * UC_{IP} + (TTQ_{Plan} - TTQ_{Base Case}) * UC_{TTQ} + (TPR_{Plan} - TPR_{Base Case}) * UC_{TPR} \]

With two each ratios we can make things clearer. The cost effectiveness of a given input is just the ratio of its effect size to its unit cost—how much it costs to achieve learning through the expansion of that input:

\[ Cost Effectiveness_{IF} = \frac{\beta_{IF}}{ UC_{IF}} \]

The second is just the share of the total cost that is devoted to any given input:
The overall cost effectiveness—learning gain per expenditure of any given plan is therefore just the sum of the cost effectiveness and cost share of each input (or, alternatively put, the cost share weighted average of the cost effectiveness of each input):

\[
\text{Cost Share}_{IF} = \frac{(IF^{Plan} - IF^{Base Case}) \times UC_{IF}}{\text{Plan Cost}}
\]

These few simple equations allow us to link the literature on “education production functions” which link inputs and learning outcomes of various inputs with the voluminous literature that links spending per student and learning outcomes.

Just from this simple equation for the cost effectiveness we should already know two things to expect from this literature: first, anything can happen and second, mostly nothing does.

*Anything can happen* in associating learning outcomes and expenditures across jurisdictions (schools, districts, states/provinces, countries) or in the same jurisdictions over time because they are just recovering some mix of the underlying cost effectiveness, the effect sizes of various inputs (which themselves vary) and prices and of cost shares, which are choices made through some type of budget process.
Table 4.2: How to understand associations of learning outcomes and spending:

anything could happen…

<table>
<thead>
<tr>
<th>What could happen</th>
<th>Empirical result that would be found</th>
<th>Relation to underling inputs and outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some schools face higher prices for inputs, but all schools use the same inputs</td>
<td>No association of spending and outcomes</td>
<td>Higher spending schools are less “cost effective” on average because unit costs are higher</td>
</tr>
<tr>
<td>Higher spending schools spend incrementally on highly cost effective inputs</td>
<td>Higher spending is associated with higher learning achievement</td>
<td>The composition of spending (cost shares) is better in higher performing schools</td>
</tr>
<tr>
<td>Higher spending schools spend more on inputs with zero or small cost effectiveness (at the same unit costs)</td>
<td>Higher spending will be only weakly associated with higher spending</td>
<td>The composition of spending by cost effectiveness is worse in higher spending schools</td>
</tr>
<tr>
<td>Cost effectiveness is high across all inputs</td>
<td>Higher spending strongly associated with learning outcomes</td>
<td>The composition of spending doesn’t matter (much) because all inputs have high cost effectiveness</td>
</tr>
</tbody>
</table>

…but mostly nothing does

| Cost effectiveness of most inputs (at existing levels of utilization and efficiency) is low | Higher spending only weakly associated with learning outcomes            | The composition doesn’t matter (much) because either (a) nearly all inputs have low cost-effectiveness or (b) most resources are spent on expanding inputs with low cost-effectiveness |

There has been a massive literature comparing levels of learning achievement and spending per pupil at many different levels—across countries, across states/provinces within countries, across smaller jurisdictions, and even across schools as well as a literature comparing performance over time of the same units. What is found again and again is that resources, per se, have a very small impact on measured learning outcomes. In this literature some find no statistically significant impact while other studies do find an impact. But as with the impact of inputs the question is not “statistical significance” but real significance, the size of the estimated relationship matters.

While there are many sophisticated ways of attempting to eke out from cross-national data what the impact of expansions of resources in any given country might be, the basic data tell the big picture story: countries have exactly the same measured learning results with very different levels of spending and countries have very different
**learning outcomes** with exactly the same level of spending. Figure 4.5 shows data from 2009 round of the PISA which shows the average PISA results across all three domains for each country and the cumulative educational spending on a child from 6 to 15 (in comparable PPP units). It doesn’t take any fancy statistical procedures to see that the United States spends PPP$ 105,000 per child and Poland spends PPP$ 39,000 and they get nearly the same results. Or conversely, Spain spends PPP$ 74,000 and Finland spends slightly less, PPP$ 71,000 and yet Finland outperforms Spain by 50 points (half a student standard deviation).

Suppose we estimate the association between spending and these measured learning outcomes using OECD data (excluding Mexico since we want to do the thought-experiment for Mexico) and asked: If, for some reason Mexico’s learning outcomes were to increase with spending in exactly the same pattern as across the OECD countries, how much spending would it take to reach an OECD based learning goal? Not surprisingly the answer is that you cannot get there that way. The simple bivariate relationship is that each $10,000 of cumulative student spending is associated with 2.2 additional points, while the Mexican gap with the OECD is 75 points (425 versus 500). So any reasonable expansion—say a 50 percent increase from Mexico’s base of PPP$ 21,000—would lead to a gain of 2 points. If Mexico doubled spending, from 21,000 to 42,000—which would entail massive fiscal costs—the gain would be 4.6 points. Even if Mexico were to reach US levels of spending (which is beyond the realm of the feasible over any time horizon) the ‘predicted’ impact would be only 14 of the 75 point gap.
Figure 4.5: If Mexico learning outcomes were to increase with spending at the same rate as the estimated cross national association even an impossibly large five-fold expansion in spending would leave Mexico well short of a learning goal.

Source: Author's calculations from data from PISA 2009.

If one extends beyond the OECD and compares countries around the world there is exactly the same pattern—countries with same performance at very different levels of spending and countries with very different performance at the same levels of spending—which means the overall learning outcomes/resources relationship is weak. Figure 4.6 takes the cognitive skills estimates that Hanushek and Woessmann (2009) produced using all of the available internationally comparable assessments and compares them to data on education spending as a fraction of GDP from the UN. Brazil’s cognitive skills estimate is 364 and Korea’s is 540—176 points higher but both spend roughly the same fraction of GDP.
Suppose the cross-national pattern revealed the typical path a country would take it expanded spending and Brazil expanded from its current levels of spending and reached the highest level of spending in the world (Denmark). At the usual estimate this would only increase Brazil’s performance by about 21 points—this isn’t nothing, but would still leave Brazil 115 points below the OECD norm. Moreover, what statistically “weak” association means is that the data cannot really tell what the association is, suppose the estimate of the association as one standard error larger, so we were “optimistic” about the impact of spending expansion. Even if Brazil were the world’s highest spending country and the impact of spending were “high” this still only gets Brazil to 405.

Figure 4.6: Using data on cognitive skills and spending as a percent of GDP show that resource effort is not strongly associated with learning outcomes

Source: Author’s calculations with data from Hanushek and Woessmann (2009) and UNESCO UIS data.

Associations between spending and learning outcomes do not reveal some deep underlying hard, technical, facts about the world, but rather the opposite. They are more consistent with a view in which the “cost effectiveness” of spending—including the individual “impact effects” of various inputs very widely around the world. So studies of “inputs” like class size or “resources” on inputs like teacher salaries do not reveal the impact of class size—they show that there is no such thing as “the” impact of class size.
In many physical sciences there are hard physical facts, like the mass of proton or neutron or electron. We know that if one atom has exactly one more proton than another atom its mass is higher by exactly that amount—whether this is in Kenya, India or Tennessee. Moreover, if we know an atom’s weight exactly we can also know exactly that it is.

But everything important about education involves human beings—as students, as teachers, as parents, as head masters—and human beings are not reducible to physical facts, they have hopes, fears, identities, likes, tastes, motivations. Human beings choose. Therefore “the” impact of adding a teacher to a classroom on the learning of students is not a fixed quantity, like proton mass, but rather is itself determined by the behavior of people.

What appears to constitute the major difference in the performance of educational systems in producing outcomes is the effectiveness with which people in those systems—students, teachers, administrators, parents—use resources. In low effectiveness systems no amount of additional resources that is not accompanied by a substantial increase in the effectiveness with which people work can achieve the education countries strive for.

IV. The West does not know best

Follow the leader makes some sense. In nearly all the international data there are four groups of countries: a set of East Asian countries that have the highest scores, most of the OECD tightly clustered around the OECD mean of 500, the Eastern European countries and Former Soviet Union also mainly near the OECD mean, and the “rest”—the “developing” countries around 400 or below. Therefore a natural instinct might be to just do what the OECD did to get to their current levels. Perhaps even education experts from the high performing countries can come and just teach us how they did it based on their own experience.

But, the striking fact is that no living Western education expert has led, been part of, or even lived through, a truly major national improvement in measured student achievement. Except for the USA the evidence is indirect, but strongly suggests that all of the currently high scoring educational systems in the OECD countries were already high scoring 40 years ago. This means even a very experienced education expert, say someone 60 years old who finished their education training at 25 years old in 1975 has never seen their country make a major improvement in average performance on these type of learning outcomes. Their lived expertise therefore can include many issues—such as dealing with racial and gender inequalities, expanding reach to learning disabled students, coping with fractious social issues—but does include creating a high performance system.

This long-run stagnation in scores also means that the massive expansion in “more of the same” in OECD education systems—much smaller class sizes, much higher real

---

9 In this case “Western” excludes Finland, to which we will return.
spending per child, much higher levels of teacher educational qualifications—followed rather than preceded the achievement of high learning performance.

How do I defend this claim about the performance of the West? Of course nearly all schooling systems are much too clever than to allow their performance to be tracked on a consistent basis over time. That would provide precisely the information needed to judge where the “improvements” being provided were really useful, and for what objectives. On this score, “it pays to be ignorant” (Pritchett, 2000)—by not tracking achievement over time this frees the system to the internally circular legitimization that more is better just because more is better.

The USA however did track performance on a consistent basis over time using the National Assessment of Education Progress (NAEP). These assessments tracked performance in two basic competences that all schooling systems promote: reading and mathematics. The really astounding results of those assessments is that over the 33 year period from 1971 to 2004 the average reading score of 17 year olds in American schools did not change at all. In 1971 the scaled score was 285 and in 2004 the scaled score as 285. No change. None. Moreover, while there have been many socio-economic changes in the USA that might have shifted population composition that affect the national average, if we limit to just “white” students with a parent who graduated from college the exact same trend holds (over the shorter period for which the data is available)—in 1980 their score was 305 and in 2004 it was 303.

What would have progress looked like that was capable of improving learning by the amounts developing countries are looking for—a student standard deviation in a generation, say. As we have seen the actual numerical scaling of learning assessments is just convention, and on this assessment a student standard deviation in 2004 was 43, so a student standard deviation gain among the privileged students over this period would have taken scores from 305 to 348 from 1980 to 2004 and this pace of progress is illustrated in the figure as a clearly entirely hypothetical pace of progress. While any education data is subject to quibbles, we can be confident nothing like a student standard deviation in improvement happened.

---

10 In the Mathematics assessment in the same population (17 year old, white, parent college graduate) scores increased from 319 to 322 from 1978 to 2004.
Figure 4.7: Progress in reading scores among 17 year olds in the USA from the National Assessment of Educational Progress (NAEP)—on average and for “elite” children compared to a student standard deviation improvement trend

Source: Author’s calculations, data from http://nces.ed.gov/nationsreportcard/littdata/

While this lack of progress has occasioned massive debate in the USA, it is actually not at all a US phenomena, in fact the US rate of improvement is roughly that of the OECD as a whole. Two German researchers, Eric Gundlach and Ludger Woessmann realized that, although the USA was unique in tracking its own performance over time, many countries participated in a variety of international comparisons that compared their national average scores to those of the USA on reading and mathematics. Therefore one could link the assessments to estimate the speed of progress in countries that were not measuring their own speed. A simple analogy is that imagine you are watching a marathon. As long as two runners remain the same distance apart if you know the speed of one you automatically know the speed of the other—whether that one is ahead or behind. In exactly the same way if the USA measures its progress over time using one instrument and Germany measures its position relative to the USA at various points in time, even using a different instrument, we can estimate the pace of progress in Germany by comparing its position relative to the USA. If it gained, it was improving faster, if it fell behind it was improving slower. Woessmann has updated this calculations recently with Eric Hanushek, using this simple concept, but carried through with the complexity to takes to make international assessments of different designs, carried out in different subject domains, different student populations comparable.

The results show three important facts.

First, all of the 15 OECD countries were at or above 500 by 1975. However the OECD established their existing levels of learning achievement they did so before the 1970s. This means no living education expert from the West has had the experience of
implementing programs/practices/activities that led to the type of massive improvements in learning achievement that countries today need to be looking for. All existing education experts inherited systems with high learning performance. Therefore their professional experience is in operating or improving functional, high-performing, systems, not how to build them.

Second, as Gundlach, Woessmann and Gmelin (2001) show for the period 1970 to 1994 there have been massive increases in real education expenditures in every OECD country—in most countries expenditures per pupil doubled or tripled over this period. Moreover, in the standard input measures—infrastructure, inputs, teacher training and qualifications, and class sizes—there have been enormous increases over this period. Whatever the gains from these were (and I am not debating whether there were not other valid educational gains from these changes) these post 1975 improvements neither account for the current OECD educational lead nor did they appear to lead to substantial learning gains.

Third, countries who wish to increase their learning outcomes cannot rely on imitating the pace at which OECD is now progressing. If one calculates the average gain in learning outcomes, normalized around the PISA 2000 standard so that an OECD student standard deviation is 100, the average gain in the 15 countries measured was 10 points over 25 years, only .4 points a year. Suppose we just do the mechanical exercise of asking, if developing country X were to achieve the pace of progress of the OECD, how long would it take them to reach the current OECD level of learning achievement? Table 4.2 answers that question and the answer is: a very, very, long time. Take Indonesia, a middle income country with a growing economy, with a PISA 2009 average score of 385. At the pace of .39 points a year it would take 115/.39=298, almost 300 years to reach the current OECD levels.

There have been some advanced countries make more progress than others, for instance, Finland has been widely acknowledged for rapid improvements. These estimates are that between 1975 and 2000 Finland improved at about 1.2 points a year, almost three times faster than the OECD average. This is terrific, and later in the book we’ll discuss more about how they achieved that. But, even at this pace Indonesia would take 100 years to reach average OECD levels.
Table 4.3: The rich countries current pace of progress is just too slow for countries to catch up....

<table>
<thead>
<tr>
<th>Country</th>
<th>PISA 2009 Average Score (R,S,M)</th>
<th>Years it would take the country to reach the level of 500 if the country improved at the pace of....</th>
<th>At the average pace of 15 OECD countries between 1975 and 2000 average (500 to 510, .4 points per year)</th>
<th>At the pace of the three fastest OECD performers (Finland, Canada, Netherlands), 1.13 points a year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kyrgyzstan</td>
<td>325</td>
<td>454</td>
<td>154</td>
<td></td>
</tr>
<tr>
<td>Peru</td>
<td>368</td>
<td>342</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td>Panama</td>
<td>369</td>
<td>340</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td>Qatar</td>
<td>373</td>
<td>329</td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>Albania</td>
<td>384</td>
<td>300</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>Indonesia</td>
<td>385</td>
<td>298</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>389</td>
<td>289</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Tunisia</td>
<td>392</td>
<td>280</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td>396</td>
<td>270</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>399</td>
<td>263</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td>399</td>
<td>263</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td>401</td>
<td>257</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Jordan</td>
<td>402</td>
<td>253</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Montenegro</td>
<td>404</td>
<td>249</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Trinidad and Tobago</td>
<td>414</td>
<td>224</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td>420</td>
<td>208</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td>422</td>
<td>203</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>Romania</td>
<td>427</td>
<td>190</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Uruguay</td>
<td>427</td>
<td>190</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Bulgaria</td>
<td>432</td>
<td>176</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Chile</td>
<td>439</td>
<td>157</td>
<td>54</td>
<td></td>
</tr>
</tbody>
</table>

Source: Author’s calculations with PISA 2009 data and estimates from Hanushek and Woessmann 2009 (table B.3)

The reason I emphasize this about progress in the West is that often debates about educational policies in poor countries become about what are the educational stylish fads in rich countries. But these are, for the most part, completely irrelevant for the solution of the learning challenge poor countries face because they are just not solving the same problem—really improving learning.
Conclusion

Countries should have a plan for improving, among other aspects of their educational systems, the learning outcomes of their students. Given how far behind the global standards these learning outcomes currently are, these plans will have to target massive gains. While these gains are demonstrable possible, the first simple question is whether a “business as usual” approach that just expands the known inputs at their existing level of efficacy will be sufficient.

The answer is no. Pushing on a string, even a politically desirable and organizationally easy string, is pushing on a string.

This is not to say expansion of inputs will not be an important part of the solution, but they will have to be either new inputs or inputs utilized in a more effective way. These are “system” changes—either in the motivation of agents, the organization of activities within the sector—including the activity of innovation itself.

Why am I devoting so much time and intensity on an essentially negative message? Because the first step to success is admitting failure. Edison created a commercially viable electric light by testing thousands and thousands of different materials and their shape. But this innovation was only possible because of three things: one, he was looking for a new source of light, not content with existing sources, two he knew what he wanted, and three, he was willing to admit failure and move on. The next chapter discusses why any of those three things are difficult in existing spider systems.
References [Incomplete]

Technical Appendix: Three issues with statistics

I) False Positives

What is the source of “false positives”? Statistics work by telling us what is unlikely to be true about the general population given the sample we have in hand. So one can look at the association between availability of chalk and student learning in a sample of students and ask, how likely is it that these two are associated in the population at large based on my sample. The way this works is that there is a “null hypothesis” (that the association is zero) that I “reject” statistically at the conventional 5 percent level if the odds are less than 1 in 20 that I would have observed the association actually present in the data if the population association were truly zero.

The problem is that one should expect “false positives” 1 in 20 times just by random luck of the draw. That means that if one tests 20 variables in an “educational production function” then, on average, the data will say that one of the 20 variables is “statistically significant” even when it is not really true about the population.

This problem is even worse if people use lower critical values, say for a “10 percent significance” then false positives will be 1 in 10.

II) Why failing to reject will lead to journal rejection

There are two very good reasons that journals find papers that “fail to reject” a hypothesis difficult to publish.

First, many tests have low statistical power, that is, have a large range of hypotheses they fail to reject. Take the simplest possible case of a linear bivariate relationship between y and x. One runs a regression, gets a point estimate for $\beta$ and then constructs a confidence interval for $\beta$ based on its standard errors and some specification of type I error (the likelihood of “rejecting a hypothesis when it is true”) which is also known as a “significance level.” Crudely put, one “rejects” a hypothesis if the value is outside of the confidence region and “fails to reject” if it is inside of the confidence interval.

Now suppose in a specific application we know for certain that the plausible (or even possible) values of $\beta$ lie between zero and 1. We collect some data, run a regression and our point estimate of $\beta$ is .5 and the standard error of estimation is also .5. If we test a null hypothesis $H_0: \beta=0$ then we would “fail to reject” this hypothesis. If one were naïve about statistics one might conclude this provided evidence that the linear association, $\beta$, was “near” zero. But what if we tested the polar opposite that the linear association was as big as it could be, $H_0: \beta=1$? Given the point estimate and standard errors we would also “fail to reject” that hypothesis. So we conclude that $\beta$ is big because our data cannot reject it? No.
The real conclusion in this instance is that the test had low statistical power (in the sense of the “power” of a microscope to give a clear picture or resolution). That is, what the data were really saying, via the standard errors, is that the existing data just could not say much about the value of $\beta$. In fact, in this case the data were uninformative—the data did not reduce the range of possible conjectures.

So the first reason there is “publication bias” is that it is very hard to distinguish a “low powered” (and hence uninteresting) failure to reject from a “high powered” failure to reject. Finding that one can not reject the standard null that there is no association is oftentimes no evidence at all.

Second, in the simplest possible linear bivariate relationship between $y$ and $x$ the consequence of mis-measurement of $x$ is that $\beta$ is smaller in absolute value, that is biased towards zero (whether the “true” $\beta$ is positive or negative). Suppose I “reject” $H_0$: $\beta=0$ what might happen to make it so that I “fail to reject” that same hypothesis. Well, all you have to do is measure $x$ more badly—that is, introduce some errors into your data, have your empirical $x$ be a less good representation of the true conceptual $x$, have your data collection be weak. In other words, one can always produce a “failure to reject” by being worse research. So trying to publish a “failure to reject” creates an uphill battle that your results are not just a result of the “attenuation bias” from measurement error in the “$x$” variable.

### III) One aspect of randomization as a method

One difficulty of nearly all statistical studies on non-experimental data is that the underlying phenomena of interest, say student learning outcomes, has many determinants or correlates. So suppose we have a simple linear association between the variable of interest, $y$, one independent variable of interest, $x$, that for simplicity only has two values: “red” and “blue”, and many other variables “z”— $z_1 z_2 z_3 z_4$ plus some noise.

$$y = \beta \cdot x + \gamma_1 \cdot z_1 + \gamma_2 \cdot z_2 + \cdots + \varepsilon$$

If we just compare whether $y$ is higher for “red” students than “blue” students this might be because a student’s “color” causes them to have higher scores but it might also be that what really causes better performance is $z_1$ and hence the association between better performance and $x$ is just an artifact of the association of $x$ and $z_1$ and the causal connection of $z_1$ and $y$. This is called “omitted variables bias.”

One “solution” is to track down the omitted variables in include them. This approach has the advantage that to get the right estimate of $x$ one also needs to get the right estimate for all of the other variables.

However, in non-experimental data “omitted variable bias” is pervasive. Since we often have empirical models with very low explanatory power, this implies there is enormous variability that is not captured in the “included” variables and hence while the remaining variation might just be random, it might be an “omitted variable” and that
variable might be associated with x and hence even our best estimates of x might contain substantial bias. With low explanatory power there is no way to be confident in the estimates of any of the variables, the x or the zs.

Experiments handle this by picking people for the “treatment” of being “red” or “blue” at random. What this does is ensure, in a statistical sense, that the “red” individuals and the “blue” individuals have the same distribution of characteristics of all of the z variables. That way, even if the z variables matter, they do not bias the estimate of x. This is great for estimating x.

The downside of this approach to estimating accurately the association, or even causal impact, of x is that it forgoes the estimate of the rest of the variables. You do not need a complete or even workable model of the real variable of interest y to be able to estimate the impact of x on y.

Suppose we believe there are 35 school related variables that potentially impact on student learning. Non-experimental estimates give weak estimates of each of the 35 while an experiment could give us a much better estimate of one of those variables (in a given context). This leaves us good information related to the question “should we do more or less of x” but without the information for a plan.