A Practical Approach to In-Country Systems Research[1]

Luis Crouch
Joe DeStefano
RTI International
June 2015

Introduction

Although this paper is about education systems research, it does not pretend to be a standard academic paper. It is, instead, an academically supported suggestion or "light advocacy" paper. Its focus is also a practical one. We are assuming that even as forms of international aid such as budget support for education ministries are questioned, non-project assistance and policy reform will continue to be relevant to the donor community. We also assume that traditional projects aimed at producing system-wide lessons for reformism will continue to be useful. As such, practical, evidence-based knowledge on how to support such programs will continue to be needed.

One of the coauthors is a member of the RISE program's Intellectual Leadership Team (ILT), and has been involved in the design of the program since its early days. As of May 2015 it seemed as if certain aspects of the program still could use sharper definition or a sharper statement of options—in particular when it comes to operationalizing things at the country level. This paper hopes to contribute to how RISE approaches the challenge of research into systems change. In particular, we draw on years of experience and research dealing with the complexities of education reform to consider how to link changes in system-level capacity to appreciable improvements in learning outcomes.

Four key questions drive how we see the role of RISE.

(1) First, education systems in the developing world need to experience very large improvements in learning outcomes—something on the order of 1 to 2 standard deviations (SDs). Such improvements are usually seen only as the result of a well-supported, direct pedagogical intervention. Absent the means to fund national-scale direct interventions, are there system-level changes that could produce **similar, or even greater, effect sizes** on a broad scale?

(2) Second, comprehensive systems reform is complicated, time consuming, fraught with political obstacles, and too often not connected to tangible, large improvements in student-level outcomes. Is there a **limited set of changes** that while indeed systemic—that is, they are more than a small pedagogical intervention or experiment—are limited enough to be tracked, and are more directly linked to (and therefore more likely to lead to) improved learning than broad systemic change?

(3) Third, system change depends primarily on the will and capacity of local leaders and stakeholders, but experience has shown that it can be accelerated through strategically chosen reform support

[1] This paper was prepared for the Research on Improving Systems of Education (RISE) Program Launch Conference, Washington, DC, June 18–19, 2015.

activities. By using information, communication, and networking tools to shape institutional change and to combat resistance and bureaucratic inertia, implementers of targeted activities can actually exert significant leverage at the system level. Is there a set of **high-leverage policy support interventions** that can increase the likelihood that system changes more directly linked to improved learning outcomes can be successfully implemented?

(4) Fourth, researching how systems changes do or do not contribute to improved outcomes poses several significant methodological challenges. RISE country studies will have to overcome low replicability, a high degree of causal interference, lack of a counterfactual narrative, and challenges to establishing external validity. Could a well-articulated **causal chain and theory of action** be used to document how reform support activities lead to systems changes, which in turn result in improved learning outcomes? And could a research design that relies on a "judicial" approach, similar to a trial by a panel of peers, be employed to verify the underlying causal relationships?

This paper is therefore divided into four sections, each intended to address the questions raised in this introduction. In responding to these questions, we hope that we are constructing an argument for how relatively small investments can accomplish two things: leverage significant improvements in learning outcomes at a system-wide scale, and document and evaluate how specific changes in system capacity and operation lead to those improvements. Experience has taught us that well-targeted and well-timed technical and policy support inputs can help education systems implement major reforms. By focusing those types of inputs on a by-design limited set of core system functions, we contend that one can achieve successful system-level reforms that, most importantly, are directly linked to improving student-level outcomes.

1. Achieving large effect sizes: Motivating a particular approach to country-based research

The existence of RISE is to be applauded. Various RISE Vision documents (Center for Global Development 2015a, 2015b, 2015c) explain the importance of the theme—we do not need to restate their premises. However, we would like to motivate one particular approach that could be tried: A focus on certain subsystems that we hypothesize could make the most difference. In reading the literature on experiments that improve learning performance, we were struck by a dichotomy. Experiments aimed at structural factors (e.g., public/private split, decentralization, school autonomy, results-based teacher pay, school-based management), accountability and incentives (e.g., local voice and choice), and inputs (more infrastructure, cash) seemed to have effect sizes on learning that, optimistically, ranged up to about 0.2 SDs. On the other hand, the experiments that focused extremely tightly and proficiently on certain pedagogical practices (improved teaching methods that meet the children where they are, vastly improved textbooks based on rigorous research, use of the children's mother tongue, etc., and all these in combination) had effect sizes that ranged up to about 0.45 or even 0.50.

We have not had time for a rigorous and systematic review of all the literature. In any case, there are several of those, and there are even reviews of the reviews; and beyond that, the conclusiveness of systematic reviews has been criticized (e.g., Evans and Popova 2015). The numbers 0.2 and 0.45 were our impressionistic recall from reading the literature casually, as it came along. However, to move beyond impressions, we took a slightly systematic (systematic lite?) look at *some* of the literature, with a specific (semi-rhetorical) agenda in mind. The usual provisos apply—such as the fact that categorizing interventions into broad areas is difficult, effect sizes have great heterogeneity, and cost-effectiveness is often not even considered.

As one step, we decided to search just the World Bank papers in the categories of Policy Research Working Papers and Journal Articles from 2010 to 2015, looking for papers on both structural experiments and more pedagogical ones, recognizing that some of the papers represented reasonably rigorous nonexperimental studies. The justification for this simple selection criterion was that the Bank has been doing a lot of good work in this area, organizes its findings well, and can be considered to be an actor with its finger firmly on the pulse of debates and interventions.

The findings were instructive. In an experiment with school-based management and accountability, Blimpo, Evans, and Lahire (2015) found no effect on test scores. Andrabi, Das, and Khwaja (2015) found relatively low impact (around 0.1 SD) from providing school performance information to the market. In research on private actors in South Asia, Dahal and Nguyen (2014) found only that private schools did no worse than public schools. One researcher who did detect a significant impact from a structural change was Yamauchi (2014), who found an effect size of around 0.3 SDs for school-based management in the Philippines. Researching the impact of community input into management, Pradhan et al. (2011) found modest effects: around 0.2. Chen (2011), researching both top-down and bottom-up accountability, also in Indonesia, presented results that were a little hard to interpret (i.e., no direct results in terms of effect sizes, and no SDs of the independent variables), but the results implied an effect size of around 0.2. Das et al. (2011) found that increases in inputs (not exactly a structural change) had an impact on learning outcomes only if unanticipated, but even then the impact was a modest 0.1 SD. Serra, Barr, and Packard (2011) tested for the effect of both upward and downward accountability in Albania and, for the variables that were significant at the .1 level or better, the average effect size was 0.16. Muralidharan and Sundararaman (2013), doing research on contract teachers in India, found an effect size of around 0.155. Goyal and Pandey (2013), on the same issue, found contradictory effects netting out to zero, as far as we could tell. Finally, in a review of many studies, Bruns, Filmer, and Patrinos (2011) reported (by our count) some 19 studies that showed effect sizes in various accountability-related reforms (mostly pay for performance, or school-based management). Putting the results reported by Bruns, Filmer, and Patrinos together with the others named here, the median effect size is 0.17, with an interquartile range of 0.13 to 0.22. So, the optimistic or reasonably high expectation is around 0.22, which confirms our impressionistic recall of the literature.

Given the tendency of rigorous research at the World Bank to be carried out by economists, and given economists' natural predilection to study incentives, accountability, and other structural interventions, the World Bank search tended to yield fewer papers on what might be called pedagogical interventions. But there were a few. In research on the impact of interventions that contained a large dose of pedagogical aspects, Jung and Hasan (2014) found effect sizes as high as 0.3 on poor children's developmental outcomes in an Indonesian early childhood program, if they had never been enrolled before. Effects on some other outcomes were not higher, but also were not lower, than the typical structural impacts noted above. Wang (2011) found that reducing age variance in the classroom—an intervention we would classify as pedagogical, but weakly so—had an effect size of 0.10.

Given the relative paucity of World Bank research on the more seriously pedagogical issues, it was tempting to look at all pedagogical interventions outside the Bank's publications database. However, that would have cast the net too widely indeed. Instead, we narrowed the non-Bank selection by looking at interventions on reading in the early grades—an area we know well. Specifically, some non-World Bank research in the area of early grade reading that has involved a fairly "tight" pedagogical intervention—meaning well-defined, specific, high-fidelity, and mostly with limited objectives—has shown relatively high effect sizes. For example:

- Literacy Boost, a program that is now approximately five years old, has been implemented in a reasonably replicated fashion in 24 countries by Save the Children USA. It seems capable of producing a set of effect sizes with an interquartile range of 0.20 to 0.56 and with a median of 0.38. (These calculations were done by the authors of this paper based on Dowd 2014 and Dowd et al. 2013.) The estimated effect sizes varied from country to country, language to language, and specific skills measured, such as nonword decoding or oral reading fluency.

- Pratham, an Indian nongovernmental organization (NGO) with perhaps the most rigorously evaluated early grade reading programs, has shown effect sizes in various programs ranging from around 0.15 to 0.70, although the literature is a little hard to decipher for purposes of comparison (Banerjee et al. 2011; He, Linden, and McLeod 2009; Pritchett and Beatty 2012).

- Experiments by RTI International in Kenya showed effect sizes with an interquartile range of 0.26 to 0.41 and with a median of 0.33 (depending on skill tested and language) after just one year (Piper, Simmons Zuilkowski, and Mugenda 2014). A similar effort in Liberia produced a range of 0.61 to 0.82 (Piper and Korda 2010) after two years; and in Egypt, RTI researchers saw improvements of between 100% and 200%—depending on the skill in question (Nielsen 2013)—after two years.

- The NGO Room to Read has operated programs in Bangladesh, Cambodia, India, Laos, Nepal, South Africa, Sri Lanka, Vietnam, and Zambia that follow a somewhat replicable pattern. As of 2014, they were able to show that fluency levels in the schools of intervention were approximately 100% higher, on average, than in control schools. The median effect size, across 18 country/language/grade combinations, was 0.91, with an interquartile range of 0.72 to 1.25 (Matthew Jukes, Room to Read, personal communication).

- A single-country experiment on early literacy instruction, combined with malaria prevention, in Kenya showed effect sizes in the range of 0.13 to 0.33 (Jukes et al. 2015).

It is important that programs that are less militantly focused than those summarized above can produce statistically significant impacts, but the effect sizes tend to be smaller. See, for example, an evaluation by Costa and Carnoy (2015) of the Brazilian Literacy Program at the Right Age (Pacto pela Alfabetização na Idade Certa, PAIC), which found effect sizes around 0.15, although higher if certain interactions were accounted for. Lucas et al. (2014) found effect sizes we could call intermediate or relatively weak (0.2 and 0.07 respectively) in Uganda and Kenya. Not all of the programs noted above that appeared to have good impact reported effect sizes, unfortunately. For the ones that did, the median result was around

0.33, with an interquartile range of 0.15 to 0.61. In that sense, the optimistic or better results (the top of the interquartile range) from the more pedagogical interventions were considerably better than those from the accountability interventions, with the medians a bit higher, and the bottom of the interquartile range being about the same as in the accountability interventions.

What does all this mean? Does it mean we recommend focusing only on the systems *directly* needed to sustain pedagogical quality, or, even more narrowly, pedagogical quality in the early grades? Not at all, for several reasons. First, some of the accountability or structural experiments were "soft" or somewhat decoupled from a tight managerial or accountability process, so one might not reasonably expect them to have a big impact. They simply were not about "hard" forms of accountability with real consequences for certain actions. And, they strike us as having been more loosely implemented than the better pedagogical experiments. (In Bruns, Filmer, and Patrinos 2011, this point is made with regard to several of the experiments.) Second, the pedagogical experiments were by no means bereft of accountability. Most of the pedagogical experiments noted above relied on fairly intensive coaching and observational visits by coaches to teachers, on site. Knowing that one will be supervised, and that one will be visited by a coach, exerts a sort of professional accountability on teachers even if there are no "hard" external accountability consequences for not adopting the behaviors on which they are being trained. Third (and to some degree consequent on the second point), it may well be that to sustain the practices—and certainly to expand them to other subjects and other grades, and to scale them up—will require some accountability pressures. Lastly, in some sense, pedagogical interventions "have it easy." That is, if one sufficiently narrows down the goal (reading in the first few grades) and the measurement is very specific, and one gets the teachers to teach to those goals, it is reasonable to expect large effect sizes. The problem, as noted, is creating systems where that kind of change becomes, well, part of the system.

In all this, it is important to note that as those interested in impact (development agencies, academics, governments) subject implementers to rigorous forms of evaluation, including randomized controlled trials (RCTs), implementers react and self-protect by narrowing the range of results desired (evaluated) and the scope of the intervention. Thus, daunted by the massiveness and difficulty of trying to "improve learning outcomes," implementers can react to the responsibility for reform by trying to discover and "prove" a fairly good way to teach reading. Even narrower, reading in the first few grades. Or, even narrower, teaching decoding, which is just one component of reading. This creates a sort of external validity problem: Does what one learned about decoding apply to slightly more sophisticated understandings of reading? Of mathematics? In later grades? What is the sweet spot for a reform goal? What makes it broad enough so one can generalize from it, yet narrow enough that one can measure it and begin to see impact in reasonable time so as to recalibrate and refine the approach, and also begin to share appropriate lessons with a world (hopefully) thirsty for improvement? Trying to hit this sweet spot is what motivates an approach that is simple enough to be called "bare bones" but that contains the key elements needed to drive learning improvements in a truly systemic manner.

Perhaps one ought to start with, and reason outward from, pedagogical subsystems that do seem to work, at least based on our close study of the limited amount of pedagogical interventions literature that we selected; and then ask what specific accountability structures might best leverage the pedagogical practices.

The next section examines what those might be.

2. A bare-bones system

It is safe to say that almost all education systems across the developing world have implemented several reform efforts over the past three decades. One cannot dispute that during that time, access to schooling has expanded, as the RISE Vision Documents explain. But at the same time, most measures of outcomes indicate that actual learning is all too rare an occurrence for many of those children who have gained entry to school, as also documented in the RISE Vision Documents. How is it possible to spend huge sums of resources while undertaking numerous rounds of reforms and still have so little to show for it in terms of educational outcomes? Having co-designed and helped implement complex education reform programs in numerous countries, the authors are comfortable asserting that the complexity of those efforts, combined with an excessively blueprint approach, may in fact be part of the problem.

The combination of policies, procedures, and implementation acumen needed to improve instruction across an entire system of schools is daunting, to say the least. For example, with the Systems Approach for Better Education Results (SABER), the World Bank is attempting to construct a framework to "produce comparative data and knowledge on education policies and institutions with the aim of helping countries systematically strengthen their education systems" (World Bank, 2015). The creation of SABER has led to the identification of 13 topics and over 500 indicators on which to judge how well an education system addresses issues related to everything from early childhood development, to assessment, to higher education, to school health and feeding. The work that has gone into researching each issue and defining rubrics by which to judge whether a country demonstrates "latent," "emerging," "established," or "advanced" capacity in each of the over 500 areas is impressive, but also overwhelming. In some far-off future, the education systems of places like Malawi, Nepal, or Nigeria may wish to have advanced capacity in each of these domains, but in the near term, such systems are failing miserably at providing the most basic educational service—namely, ensuring that children can learn to read and do arithmetic after the first few grades of primary school. Can one realistically expect them to take on all the institutional challenges inherent in even one of the SABER topics?

For this reason, we contend that the challenge of getting an education system to produce reliably better learning outcomes is more about operational capacity than it is about policy and the institutional environment. The distinction we are trying to make here is between *getting the policies right* and *managing implementation.* Developing countries are adopting policy statements and elaborating education sector plans that often say all the things they need to say (much as countries' constitutions do)—and in the past several years, we have seen increasing evidence of sector strategies that explicitly claim improved learning outcomes as an objective. This represents progress, due in part to growing recognition that although there has been success pursuing Education For All (EFA) goals, enrollment is not translating into learning. However, that is still a long way from being able to implement those policies and manage the day-to-day operations that will, even at the most basic level, reveal to a ministry the extent to which its intentions are translating into tangible changes in classroom practice. Further, it says nothing of actually putting into place the scaffolding that is likely needed to increase the probability of that happening.

In thinking about what education systems can do to improve learning outcomes, we are therefore taking a decidedly simpler approach—one that recognizes what we contend are a few important truths.

First, the ingredients for good teaching and learning of basic skills in primary school are well known. Schools—and within schools, teachers—need to dedicate sufficient instructional time. During that time, they must use sound instructional techniques and well-designed materials to afford students opportunities to learn and practice basic skills according to a well-understood sequence of how children learn, for example, to read or do math. The simple view of learning is that learning = learning time × rate of learning. The former is "just" a matter of time (and accountability for using it well), and the second is related to method. This is what informs the interesting work of Barbara Bruns at the World Bank (Bruns and Luque 2014).

Second, many developing country education systems are not ensuring that students acquire basic skills such as reading and securing an understanding of math, precisely because they are not able to guarantee the basic requirements named above. Studies of allocated versus effective instructional time show losses of as much as two thirds of class time due to late start/early ending of the school year and each school day, capricious school closing, teacher and student absenteeism, and poor management of classroom time. Observations of classrooms across numerous countries have also shown that during the times when teachers and students are together for a lesson, classroom time is not spent on instructional activities known to increase learning (Schuh Moore, Smiley, DeStefano, and Adelman 2012). Materials are often not present in sufficient quantity, and those that are available are not used effectively to support instruction, partly because their quality is so low that teachers are not able to use them and partly because teachers are not trained explicitly in how to maximize the value of materials during teaching. Additionally, instruction falls far short of what is needed because curriculum is too often overly ambitious, inappropriately paced, and poorly sequenced. For one, the curriculum content may start at a point well above where students are when they enter school. A recent review conducted by RTI International (2014) for the U.S. Agency for International Development (USAID) of the content of Arabic books in Morocco found that the lesson for the first day of first grade contained a story written at the level of an independent reader. In addition to getting the starting point wrong, there are also mismatches between students' pace of learning and the pace assumed in the curriculum, resulting in most students being left behind at an early point in their schooling. Pritchett and Beatty (2012) documented this phenomenon rigorously, showing how, for a given distribution of students' skills, an overly ambitious curriculum actually produces less learning.

Third, education systems fail at ensuring that schools can assemble the necessary ingredients for good teaching and learning because they barely even try to. Most schools operate in isolation from the system—perhaps receiving some resources such as curriculum and materials, but at the same time perhaps not even guaranteed that staff salaries are paid regularly. Teachers may participate in professional development activities, but those are usually divorced from classroom practice, delivered through a highly ineffective cascade, and devoid of the kind of ongoing support that actually allows teachers to apply what they learned in a workshop. Even if a well-intentioned reform leads to research-based improvements in the structure of, say, the reading curriculum for the initial grades of primary school, it is no wonder that such a reform often fails to alter student outcomes, if teachers receive only a smattering of poorly delivered training, may have materials delivered late if at all, and have no one to turn to for advice on how to align their daily practice to the new program of study.[2]

We therefore see the challenge of "systems change" not in terms of trying to catalogue all the policies, procedures, and institutional capacity needed (say, like SABER), but rather in terms of identifying the bare-bones things a system needs to do to ensure that teaching and learning in all schools improve. This requires stripping down the notion of an education system to its first principles.

The birth of the modern public education system is usually traced to Horace Mann and the emergence of the notion of the "common school" in the eastern United States in the mid-19th century.[3] Simply put, Mann saw the need for an administrative system that could overcome the great variations in the quality of schooling which local communities were able to organize for their children. Schools would not be just local affairs, but would become part of a system. What would be the purpose of such a system? At its origin, the notion was to establish some standardization in the provision of schooling—that is, establish the floor and monitor to see that schools met those minimum criteria.

The century that followed Mann's call for standardized schooling saw the rise of the industrial model of education, relying on "scientific management" to oversee and ensure standardized inputs—e.g., statewide curricula, certification requirements for teachers, increasing regulations governing the organization of schools and the school day. And when the industrialized model of an education system was transferred to developing countries, leadership in those countries became overly focused on copying the bureaucratic trappings of an education system rather than the substance of teaching and learning. This is what Pritchett (2013) referred to as "isomorphic mimicry"—the tendency to create institutions in weak, developing countries that look like those in states with functional public sectors. By pretending to implement the reforms tried in successful countries, but without putting in place the underlying functionality that actually makes the reforms work, countries wind up with education ministries, offices, agencies, and institutions, but still lack the operational ability to actually deliver high-quality services. This notion of isomorphic mimicry actual helps explain why any current institutional assessment in a typical developing country's education sector leads one to conclude that the lack of institutional capacity is a severe constraint, despite decades of projects aimed at "building capacity."

Perhaps the simplest and most basic example of isomorphic mimicry involves the institutional commitment to achieving universal access. The ability to ensure 100% attendance in school is something that developed-country school systems built as they evolved—starting with ensuring that kids came to school, by using highly local truancy systems to monitor who was skipping out and getting them to attend. Developing countries mimicked the idea of 100% attendance by setting goals for full enrollment, but without having in place the local management structures, mechanisms, or incentives to promote regular attendance at the school level. Thus, systems have enrollment rates that meet the goal of

[2] For a general discussion of problems with implementing education reforms at scale see Elmore (1996). For a detailed account of the example of the implementation of the National Literacy Acceleration Program (NALAP) in Ghana, see RTI International (2011).

[3] Among numerous possible references, see an interesting perspective on the emergence of the publicly funded system of common schools in the United States in Cubberley (1919).

universal access, but without the means to even know whether enrolled students attend school regularly.

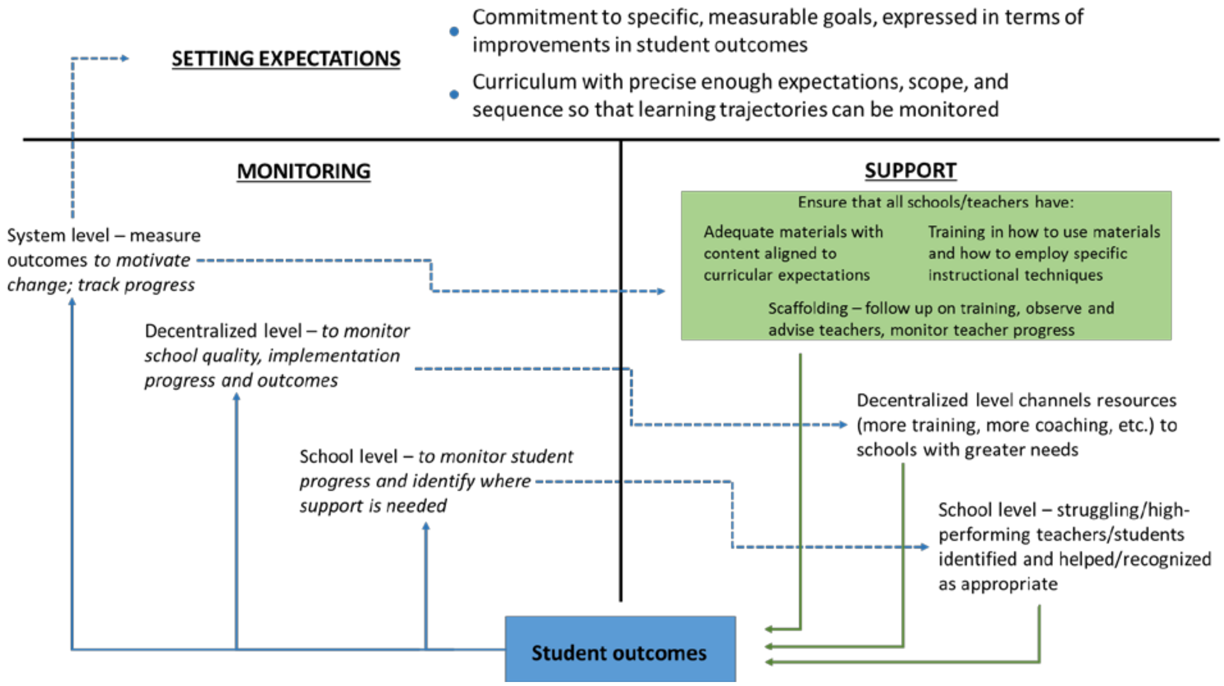
We contend that the challenge of successfully realizing large-scale improvements in learning outcomes is to replace what have become overly bureaucratic, administrative structures intent on isomorphic mimicry with an explicit management imperative to achieve results and equity. This will involve moving from "thin" descriptors of system relationships to "thicker" ones (using the terminology in RISE Vision Document No. 3), but, given the weak capacity of most systems to think through "thick" descriptions, much less implement them, one can propose focusing on the ones that seem to matter the most. In the early 2000s, the Annenberg Institute for School Reform set out to determine how school districts could add more value to what were largely state-driven reforms. Annenberg asked a very basic question: "Why have a school district in the first place?" In a "greenfield" exercise, leading U.S. scholars, educators, and researchers were asked how they would design a school district if they were starting from scratch. Interestingly, it did not take long for that group to define a very bare-bones sense of what a school district should do—namely, establish a culture focused on results rather than administrative compliance, correct for the fact that some schools/communities will function better than others, and protect children by intervening when schools are failing (Ucelli and Foley, 2004). Likewise, we are electing to revert to the basic notion of how a system adds value to schools, namely by performing three bare-bones functions:

- Setting expectations for the outcomes of education
- Monitoring and holding schools accountable for meeting those expectations
- Intervening to support the students and schools that are struggling, and holding the system accountable for delivering that support

What makes an education system a "system" and not just a collection of schools is its ability to operationalize this basic framework. This does not mean that we devalue all the myriad functions an education system may be called on to perform, such as ensuring pre-service training and certification of teachers, deploying staff, and guaranteeing equitable financing of schools. We are contending that the three functions we indicate above are what most determine if the system can influence whether schools can get students to achieve better learning outcomes. Each function is described below and the relationships among them are depicted in ***Exhibit 1.***

Before we go on to describe these three components, we have to note that there is an interesting issue here related to the degree to which having this kind of simple, three-ingredient approach agrees or disagrees with the "MeE" approach advocated by Pritchett, Samji, and Hammer (2013). We will defer that discussion until later (the section on implementation), but we want to point out that we are aware of the apparent contradiction, and to say that we think the contradiction is only apparent.

Exhibit 1. Diagram of three core functions of an education system


SETTING EXPECTATIONS
Commitment to specific, measurable goals, expressed in terms of improvements in student outcomes
Curriculum with precise enough expectations, scope, and sequence so that learning trajectories can be monitored
MONITORING
SUPPORT
Ensure that all schools/teachers have:
Adequate materials with content aligned to curricular expectations
Training in how to use materials and how to employ specific instructional techniques
Scaffolding – follow up on training, observe and advise teachers, monitor teacher progress
System level – measure outcomes *to motivate change; track progress*
Decentralized level – *to monitor school quality, implementation progress and outcomes*
School level – *to monitor student progress and identify where support is needed*
Decentralized level channels resources (more training, more coaching, etc.) to schools with greater needs
School level – struggling/high-performing teachers/students identified and helped/recognized as appropriate
Student outcomes

Setting expectations

Having clear, measurable targets for system improvement is needed to drive decision-making. For example, having measurable goals for Education For All allowed education systems to monitor whether they were making progress or not. From 1990 to now, decisions throughout education systems have been driven by commitments to achieving EFA goals. Enrollment rate targets were translated into the numbers of schools or classrooms to be constructed, numbers of teachers to be hired, and other materials to be procured and distributed. Gender equity objectives led to the development of numerous strategies for increasing girls' enrollment. Plans, strategies, and budgets were all driven by the desire to reach measurable goals. Only with similar, clearly articulated measures of improved learning outcomes, and the systems for translating these down to the school level, will attention begin to focus on what is needed to increase student achievement. Additionally, the articulation of well-defined learning outcomes for each grade and level of education can (and should) drive how curricula are structured and sequenced. For example, the standards movement succeeded in defining not just what students should learn, but what they should be able to do—in other words, expressing curriculum outcomes in terms of desirable levels of proficiency in specific skill areas. This helps teachers, students, and families better understand what students should be able to do as they move through school and therefore better monitor whether students are on track or not. Any aspiration to create accountability for learning outcomes requires as a starting point this shared understanding of what those outcomes should be.

Monitoring and support

If teachers are going to help their students perform up to expectations, then the system needs to ensure that teachers and their schools receive support (the right side of Exhibit 1). In this manner, we can think of system improvement as similar to project implementation at scale. Can the necessary inputs and supports be delivered to all schools? This litmus test provides the distinct advantage of defining system functionality, and therefore its measurement, in precise terms. It also lends itself to a clear set of management imperatives—e.g., make sure (and monitor) that good materials are getting to schools and that teachers are sufficiently trained and supported to exploit those materials). It puts in place the basis for more transparent accountability.

The system also needs to be monitoring overall performance with respect to the clearly stated outcome objectives (the left side of the diagram). System-level monitoring, for example, can be accomplished through periodic, sample-based assessment of student performance in specific skill areas. The past decade of international development work in education has, in fact, seen great progress in helping education systems perform exactly this kind of monitoring—be it through international comparative assessments such as the Trends in International Mathematics and Science Study (TIMSS) or the Program for International Student Assessment (PISA), through early grade reading or math assessments (EGRA and EGMA) performed in numerous countries, or through household surveys such as those completed by Pratham in India (for example, see ASER Centre 2015) or Twaweza's Uwezo initiative in East Africa (see Mugo et al. 2011 for an example). Such assessments have been instrumental in motivating greater attention to learning outcomes as indicators of education system performance and are increasingly being used to measure whether systems are improving over time or not.

At the next level down, the system also needs to monitor whether schools, teachers, and students are receiving the basic package of inputs. This monitoring of basic implementation needs to be coupled with tracking of outcomes so that the education system can verify if the school improvement interventions are having the desired results. Too often, education reforms are launched and at some later point results are evaluated; when little impact is noted, it is not clear whether failure is due to poor implementation or ineffective reforms. We propose getting the system to explicitly state its "theory of change," map the sequence of actions and underlying assumptions linked to those actions to put that theory into operation, and devise the means to not only monitor implementation and impact, but also revisit and test initial assumptions along the way.[4] Accountability for results therefore needs to be tied to accountability for provision of basic supports. Schools, teachers, and students cannot be expected to meet and be held accountable for improved outcomes if they do not receive the materials, training, or ongoing support recognized as necessary to producing those outcomes.

System capacity to monitor school-level implementation and results is decidedly different from what typical education management information systems (EMIS) are structured to do. The EMIS in most countries are designed to perform an annual census of schools—collecting information on students, teachers, and infrastructure. These data are used almost exclusively for planning and broad policy purposes, not for day-to-day management decision making.[5] We depict decentralized monitoring as

[4] What we refer to later in this paper as "crawling the design" and "crawling the implementation space."

[5] In a study of the capacity of data systems in Mozambique, Ghana, and the Philippines (DeStefano 2011), researchers from RTI International found that each country had multiple data sources and systems within the education system, but almost all focused on accumulating the data needed for basic planning rather than providing information relevant to the management of learning. And the additional data systems (such as examinations,

directly linked to the provision of additional support, wherein the information collected is designed to indicate where such support is most needed. Methodologies such as lot quality assurance sampling, or LQAS (Valadez 1992), are being experimented with as a means to change the information culture within education systems. Their purpose is to shift the focus from EMIS counting of inputs to measures of basic operational quality and performance at the school level, and then to use those measures to drive decisions regarding the support schools need to improve.[6]

Developing leading rather than only lagging indicators of improved outcomes is another aspect of how monitoring at the decentralized level needs to change (Foley et al., n.d.). While we do want systems to focus on learning outcomes, test scores are by definition lagging indicators because the tests can be given only after instruction has been provided. And the time lag needed to process test results means the data may not be available until after a school year has ended. Leading indicators, on the other hand, would tell the system if the conditions likely to contribute to student success were in place—at the start and at periodic points throughout the school year. Opportunity-to-learn indicators (Gillies and Quijada 2008) that show whether schools are open when they should be, whether teachers are present, and whether class time is used for effective instruction are examples of leading indicators. Other measures of school management effectiveness can be incorporated into an LQAS-based monitoring framework to signal whether schools are positioning themselves to produce better outcomes.

Knowing which schools are performing and which are not (in terms of both leading and lagging indicators) allows the system to then provide differentiated responses based on that performance. Schools that are succeeding maybe just need to document their success and be evaluated to see how the lessons they are learning could be shared with other schools. Those that are struggling can be targeted with additional help: more visits, more coaching, additional training, or supplementary materials. This approach would represent a cultural shift for most education systems, leaving behind the simple administration of an evenly divided resource pie and replacing it with mechanisms for targeting resources based on need.

Similarly, at the school level, the school community needs to monitor individual student progress so that teachers and students who are struggling can receive the extra support and pressure they need to succeed. The idea here is that teachers need to relate to their communities and their professional supporters in a “thicker” fashion. This requires the use of monitoring and assessment tools at the school level that can show schools which teachers are performing well (e.g., showing up, using improved methods and materials) and which students are reaching desired levels of proficiency. And when data show that some students are not performing well, schools, teachers, and their communities need to make another cultural shift. They need to go from assuming that some students will succeed and others will not because of the basic endowments different children have, to recognizing that all children can and should succeed—certainly in basic education. While research on brain development has moved decidedly from the old notion of intelligence being a fixed commodity to recognizing that intellectually ability is plastic and can grow, most schools operating in the developing world have not made this shift.

household surveys, or analyses of the labor market) that could provide useful information were not being exploited for education system management purposes.

[6] RTI has piloted the use of LQAS-based approaches to school monitoring in Ghana (RTI International 2013), Zambia (Bostock and Rakusin, 2014), and Tanzania (in April–May 2015; report pending).

Children will learn at different rates (especially young children), and schools that serve disadvantaged students need to understand that many children will need more instructional time to develop the levels of expected skill. For example, the development of language—and therefore reading skills—in children from resource-poor home environments requires much more instructional time (Brown and Saks 1986). Organizing supplemental learning activities (remedial classes during the school day, extra time after school, summer school) should be part of what schools and their communities do in response to information that indicates some students (or schools) are performing below expectations. The role of the system is to encourage (if not require) and, more importantly, support school communities in doing this—for example, providing funding, training, and content for effective remedial programming.

3. Putting in place the bare-bones functions: Political-economic aspects

While we are advocating a bare-bones approach to systems change, we are not so naïve as to pretend that the challenge of getting day-to-day instruction to change in thousands of schools is easy. In fact, it is in recognition of how hard it is to change the behavior of tens of thousands of teachers that we are trying to limit the scope of the system functions which we argue for taking on. And since we also recognize that changes in the status quo ante will provoke opposition from various stakeholders, we want to limit the number of fronts on which political-economic battles would need to be fought.

Our understanding of the political economy of reform begins with the recognition that the existing arrangements are not an accident.[7] Various actors in the system are benefiting in a variety of ways from how things are presently being done, even if it is only from the comfort of doing things by inertia. We also recognize that while there has been some success brokering reforms aimed at increasing access, many would argue that expanding the provision of schooling was actually the easy part of education reform, since most of what was needed was more, more, and more (more schools, more books, more teachers). Getting and spending more money does not necessarily challenge the existing arrangements, and may in fact further reward the interests that are already benefiting from things like construction contracts, book purchases, and handing out of jobs (DeStefano and Crouch 2006).[8]

In contrast, what are needed now are reforms that change how actors throughout the education system define their roles, interact, and carry out their functions on a daily basis. Since there is no such thing as a clean slate, concerned actors have to stop doing what they presently do and begin doing what is needed to more effectively fulfill the core functions we have described above. If changing human behavior were easy, we would all already eat healthier food; exercise more; and be better spouses, parents, and colleagues. The usual challenge of changing human behavior is compounded by the vested interests that

[7] See, for example, the Education Reform Support series published by USAID’s Advancing Basic Education and Literacy (ABEL) Project in 1997, beginning with the overview and bibliography (Crouch and Healey 1997).

[8] Of course, important political battles were fought to bring about some of the reallocations necessary to expand access. We are not forgetting or downplaying how difficult some countries found it, for example, to reduce scholarships for university students in order to allocate more money to basic education. In fact, one of us vividly recalls being hunkered down in a minister’s office while university students rampaged through town to protest their reduced subsidies. Some lessons learned in the successes of the reforms implemented in the 1990s to free up resources to fund expansion of access are in fact what we draw on to strategize how to tackle the political challenges that inevitably will ensue in trying to put in place the bare-bones functions we write about here.

may be associated with the existing ways of doing things. Lax enforcement of teacher attendance means teachers may spend the time when they should be in school engaging in other income-generating activities, especially if they are not receiving their school salary regularly. Lack of accountability for district supervisors visiting schools, and ample excuses available to them for not doing so, means they do not have to spend time away from home. Putting in place the bare-bones functions would have to confront exactly these kinds of obstacles.

In fact, we see three categories of obstacles to education systems being able to create clear outcome expectations, monitor and hold accountable schools and teachers for meeting those expectations, and hold the system accountable for delivering the supports struggling schools need to succeed. First, there are technical challenges. Second, there are mental models that need to be changed and the inertia of the existing ways of thinking and doing business will need to be overcome. And third, the active opposition of entrenched interests will need to be combated.

Designing the assessments and the tools and procedures for the kinds of monitoring we see as essential to system improvement will require specific technical know-how that most ministries in developing countries do not possess. However, there is ample experience adapting and applying simple assessments like Pratham's ASER studies, the Uwezo *Are Our Children Learning?* assessments, or the EGRA or EGMA in numerous countries. Assessors can be relatively easily trained to reliably carry out sample-based, national assessments. Applying LQAS techniques (as mentioned above) in education is still in early stages of experimentation, but standardized approaches are being developed for sampling, defining indicators, and training school directors and district administrators in Ghana to collect the necessary data, including implementing periodic group assessments (RTI International 2013). In Zambia, an SMS-based messaging gateway is being employed to transmit data (Bostock and Rakusin 2014) from schools to the district level, where school report cards are then automatically produced, showing how a school compares to itself over time, as well as how it compares to district, regional, and national averages on key indicators.

Systems for electronic monitoring of classroom practice are also growing in application: Projects in Kenya (see Piper et al. 2015) and Malawi are supporting school monitoring personnel using tablet-based software to record observations of instructional practice and to monitor their own provision of support to teachers. Improving the capacity to develop and implement these kinds of data systems as features of an ongoing national, district, and school monitoring system will require some investment, but the outright costs are not large.

Once mechanisms are in place to allow such monitoring data to be collected and compiled on a regular basis, the system then needs to respond to what those data show. This requires decision-making authority, control of resources, and managerial capacity to direct resources based on need. It also requires that those who support schools—be they on site or at a subdistrict or district level—know what to do to help a school be more successful.

For example, in the late 1990s, Chicago schools put in place a system for monitoring and holding schools accountable for student performance. This included the lowest-performing schools being placed on probation, during which they were mandated to receive external support, ostensibly to help them improve. However, the support response provided was found to be too weak, underspecified, inconsistent, and of too low intensity to bring about the kinds of changes in teaching and learning that low-performing schools needed (Finnigan and Oday 2003). Education systems therefore need the capacity to deliver high-quality interventions that target specific aspects of instruction and that bring to bear the necessary infusion of support (e.g., additional training or mentoring of teachers, higher-intensity follow-up, and incentives to encourage staff to increase their efforts). The pilot reading program in Liberia that was mentioned above experimented with using accountability with support and accountability without support, and found that in schools where no intervention was provided, solely reviewing with educators the data on their students' performance could not drive improvement (Piper and Korda, 2010).

Similarly, McKinsey & Company (Mourshed, Chijioke, and Barber 2010) set out to identify the traits shared by education systems that over time have managed to dramatically improve learning outcomes. The authors reached a similar conclusion: School systems that improved outcomes while starting from initially low levels of performance did so by focusing on using data to guide school support, targeting low-performing schools with additional support so that all schools could be raised to a minimum quality level (bringing up the bottom), and providing motivation and scaffolding for low-skill teachers. In short, designing corrective interventions and building capacity within the education system to carry them out is a technical hurdle that needs to be overcome. Many quality-improvement efforts fail because this capacity is lacking; schools and teachers may be exhorted to improve, but with no evidence-based, defined set of things to do to get better, they still languish. Having every single school undergo a discovery process to figure out what to do just takes too long (and often leads local actors to focus on the wrong things).[9]

Changes in how data are collected and used and in how the system manages and targets resources are not just technical challenges. They also require key actors to think differently about the nature of what the education system is trying to do. We talk about this as a cultural shift or a changed mental model—one in which the perception of the role of the system vis-à-vis schools is redefined. All education systems have administrative subdivisions that are responsible for overseeing schools in their jurisdictions. They usually have responsibility for: collecting data (through the school census), monitoring school operation (through periodic, usually quite infrequent and shallow, inspections), managing personnel (processing assignments and transfers), participating in the administration of examinations, organizing teacher professional development (through workshops or cluster/school-based teacher gatherings), and sometimes overseeing school improvement planning (which may include managing school improvement grants).

Staff of these decentralized offices are not necessarily trained to intervene in ways that help schools become more successful, and in fact may see their job as being the monitor of quality, not the one responsible for making sure quality improves. Having these administrators and inspectors become

[9] In a May 2015 review of early grade reading interventions in Cambodia, RTI researchers for the USAID Education Data for Decision-Making (EdData II) project found that the Child Friendly Schools Framework is being used to guide school improvement planning and school support, but that the framework includes 177 different indicators against which schools are supposed to evaluate their performance, of which only 17 relate to teaching and learning (report pending).

accountable for whether they help schools succeed or not will require a dramatic redefinition of their roles and concomitant reshaping of the mental models which people bring to those roles. Individuals' perceptions of the roles they play, of their own ability to be successful in that role, and even their definition of what success means will need to change. If teachers go into teaching because it was a course of study open to them, with the promise of a job that requires them to show up intermittently for a few hours per day for only 10 months out of the year, then simple training is not necessarily going to get them to become dedicated and effective.

In addition to the technical and conceptual obstacles, there are political-economic interests that are likely to push back against proposed changes. To realize that not everyone is in favor of establishing clear, outcome-based benchmarks for evaluating performance, one need look only at the battles being fought against the Core Curriculum in the United States, or note the fact that the National Assessment of Educational Progress (NAEP)—the one measure of how well districts or states in the United States are doing with respect to an objective measure of student performance—remains voluntary and has restrictions on reporting disaggregated results.

In developing countries, similar kinds of resistance to putting an emphasis on learning outcomes as the measure of system performance will surface (or already have). Education system leadership may express interest in measuring outcomes, but then recoil from results that show how poorly their system is performing, opting to squelch results rather than share them publicly. Furthermore, rare (if not non-existent) is the bureaucratic entity that voluntarily signs up for increased accountability. At the school level, teachers and principals may object to being answerable to their communities and to their supervisors for their students' performance. Teacher union leaders are likely to oppose attempts to increase teacher accountability to communities or bureaucracy, as opposed to accountability to the leaders themselves.[10] And the notion of reverse accountability—in which schools and their communities hold the education system accountable for delivering high-quality supportive services—not only is likely to be resisted by the administrators responsible for providing those services, but also may be a concept completely outside their ken. ***Exhibit 2*** summarizes some of the obvious ways in which resistance to the bare-bones functions could manifest itself.

[10] Our notion on this is that the rank-and-file has a pact with the leadership—a pact that is not public and that, we would argue, most observers miss. The rank-and-file pay dues and in return expect the leadership to engage in collective bargaining to improve salaries and working conditions. But the leadership knows that collective bargaining requires a rank-and-file that is malleable and will respond to the call for collective action. The rank-and-file, or at least a significant portion of the rank-and-file (perhaps those with higher ability or willingness to work better for more pay) may not be opposed to ideas such as merit-based pay, or increased accountability to communities. But the leadership realizes that anything that differentiates among teachers, and increases teachers' allegiance to anyone other than their own collective and the leadership of the union, decreases the leadership's ability to collectively mobilize them and thus reduces the leadership's ability to fulfill their end of the bargain, and continue to "deserve" the dues (and honor and prestige and political influence).

Exhibit 2. Summary of bare-bones system functions and likely paths of resistance

Bare-bones functions	Political-economic obstacles likely to manifest themselves
Setting expectations for the outcomes of education	Resistance from leadership to publicizing results showing how poorly the system is doing
	Stakeholders still interested in promoting access as the highest priority
	Public pressure to address access to secondary education, rather than continuing to focus and improve the quality of primary education; expansion "up" of the access agenda
	Stakeholders invested in high-stakes public examinations, instead of potentially more useful assessments, as the measure of student/school/system performance
Monitoring and holding schools accountable for meeting those expectations	Prevailing culture of no accountability
	Principals and teachers (or, rather, their union/professional organization leadership) that will lobby against being held accountable for student performance
	Groups that believe simple measures of basic skills do not adequately capture the breadth of educational objectives schools should be promoting
	Public examinations interests that may object to other measures of student performance being introduced
Intervening to support the students and schools that are struggling and holding the system accountable for delivering that support	Administrators/school support providers who at present do not have to exert the effort needed to get out to schools
	Administrators who would not accept schools holding *them* accountable for delivering useful services/support
	Existing teacher-training interests that want the focus to remain on certification training rather than on using resources for other kinds of teacher support

What makes system change additionally challenging is that the benefits associated with the current arrangements tend to be concentrated and accruing to constituencies that are already organized—teachers and their unions, education administrators, political leadership. On the other hand, both the potential benefits and the beneficiaries of the reformed way of doing things are dispersed, and the constituencies that stand to gain from change (parents, children) have high relative costs associated with getting themselves organized. Any solution therefore needs to include mechanisms that facilitate and subsidize the organization of otherwise dispersed stakeholders. Classic community organizing and advocacy are designed to do just this. And the work of an organization such as Twaweza (leader of the Uwezo initiative) in East Africa is particularly instructive in how to generate data and use those data to both rally otherwise disconnected constituencies and create political pressure and advocate for reforms. We see this set of tools as essential to engaging the political-economic dimensions of systems change, namely marshalling data, inserting those data into well-facilitated deliberations and dialogue, conducting communications campaigns, and building networks among supportive organizations and reform-minded actors.

A growing body of evidence (as discussed in the first section of this paper) is demonstrating that learning outcomes can be improved significantly through pedagogically focused interventions on a manageable scale. The track record—where one exists—for implementing these kinds of interventions on a national scale is less impressive. All the examples of how developing country education systems fail at even the most basic logistics of introducing pedagogical innovations (for example, making sure new materials are delivered in time for the start of a school year, and that teacher training on those materials takes place) would make a very long, and sad, list. That is why we have focused on increasing the capacity of education systems to deliver the basic package of supports that schools need; then to monitor how those supports are used to improve teaching and learning; and to provide additional supports when a certain number of schools/teachers struggle to adjust to the new materials, methods, or other innovations.

Well-timed technical inputs to improve curriculum, materials, and training, and to help develop management and information and communication systems, will be needed. And strategically directed education reform support activities to inform and rally public opinion, combat opposition, and create coalitions of supportive stakeholders also will be needed. The authors contend that investing in supporting systems-level change by providing targeted assistance to address the technical, managerial, and political dimensions discussed here are potentially some of the highest-leverage options available to agencies wishing to support large-scale improvements in education. The alternative of trying to use "projectized" assistance (in the language of Pritchett, Samji, and Hammer 2013) to fund all the inputs necessary to ensure high-quality instruction on a national scale would be extremely costly, and yet still would not guarantee that the system would become capable of guaranteeing the three categories of bare-bones functions described here. And the alternative of using budget support and policy trigger points in nonproject loans and grants typically does not provide enough technical dialogue, nor does it take into account the political-economic reality of the changes needed.

4. A proposed implementation/evaluation approach

This section deals with the issue of how to make an actual systems reform project, on the ground, in a particular country, implementable in such a manner that it yields important evaluation lessons and, perhaps more importantly, that it actually results in impact. We want to start with evaluation because we believe that this can focus our attention on evaluable implementation, which is what we want.

A key assumption is that that even system reform can be projectized, a notion proposing that systems change can be thought of a project. That is, we assume that one can implement a reform as if it were a project: with a theory of change, logical frameworks, implementation plans (which have to be adaptive, of course), specific plans of evidence, and so forth.

Ever since this line of work started to be discussed by the UK Department for International Development (DFID) and a few others in mid-2012, a key question in our minds has been: How can you make something with *N* = approximately 5 (five countries) yield reliable knowledge about causality? Such a small sample will produce the same external-validity causality-inference threat as any other low-replication set of experiments.[11] One will also have the within-project causality-inference threat because each country experiment itself will have no counterfactual, and $N = 1$.

An opposing argument might be that, well, each country should engage in a vast number of little experiments, each one in a few dozen or a few hundred schools with good sample sizes and control groups, and thus a few hundred schools per experiment. That would at least deal with the internal causality-inference threat. But the problem, then, is that one does not really have an experiment in reforming systems. A system is not just a collection of experiences, and it is not enough to simply say "a better system is one that can implement the implications of all these little experiences." And by definition, "a system" means *one* system.

One could propose to experiment with many systems—for example, select a much larger set of countries than DFID has in mind; or, within countries, experiment with many provinces and use, say, different systems in each province. But this approach would not work either, for at least two reasons. First, it is unlikely any country would allow it to proceed. Second, it is doubtful that any experiment, particularly around the systems issue, could systematically vary just one simple thing and then have anything meaningful to say about "the" system of other things that may or may not need to be done. As an example, consider the idea of varying *only* the style of teacher pay, in a simple and specific way (that is, not using "teacher pay" as a label for a class of complex things) and then expecting variants of teacher pay regimes to have an effect on instructional quality. This point is made by Roberts (2004) as cited by Pritchett, Samji, and Hammer (2013).

One way out is perhaps not to think about how an experiment with $N = 1$ resembles or does not resemble an RCT or other rigorous quasi-experimental approaches, or even multivariate controls. That would seem to be a road that leads only to frustration. There are at least two sensible alternative approaches we can think of. One focuses on how institutions learn, and the other takes one step back and reasons in terms of causality. Moreover, as yet a third alternative, these could be combined.

The most relevant exposition of the first, *how institutions learn,* appears in Pritchett, Samji, and Hammer (2013). If we were planning to propose a country research case, we would adopt their recommendations.

The other complementary and sensible approach is to reason philosophically about causality ("what *is* causality, dude, hmmm?").[12] After all, RCTs are thought of as the Holy Grail of research because they profess to be able to tell us something about causality. In other words, they are perceived as highly valid

[11] And, at least in some philosophical traditions, lack of external validity is itself a causality problem. If an experiment is done once, very rigorously, one cannot say that one has discovered a causal path. Testing a cancer drug of amazing effectiveness in one single replication would not really permit a strong claim that the drug causes a cancer cure. So external validity is not a trivial issue.

[12] Though of course we will not ask that question. Our impression is that practical philosophers who worry about causality do not actually define it, except in terms of *how* one can deduce that something is causing something else. In any case, if we follow David Hume, who is perhaps the most important source of reasoning on causality, there is no way, even with the most exquisitely counterfactual and sophisticated research, to ever prove causality. All one is ever proving is, ultimately, correlation.

not just because, in an operational sense, they randomize and so on—those are paraphernalia; instead, it is the promise of causal reasoning that is attractive.[13]

We do believe it is important to think about causality. While Pritchett, Samji, and Hammer's experiential learning paper is very realistic about how institutions learn, and how projects can be designed so that implementers learn, it is also useful to think about whether the learning is about something causal. Other disciplines and other areas of life are just as interested in causality as development economics and educational development are, they face the same problems we claim our proposed system reform project would face, and they have had important things to say about the subject. In the end, what we propose is experiential learning as an implementation modality that can be supplemented with other forms of evidence that can come as close as possible to establishing causality within a project designed according to the experiential learning principles in Pritchett, Samji, and Hammer (2013).

So, let's return to thinking about determining causality for $N = 1$, through noncounterfactual experiments such as reforming a system. We could also extend the experimentation to a subsystem: It would still be effectively $N = 1$, and noncounterfactual. Note that it strikes us as unrealistic, however, to consider experimenting with different systems in different jurisdictions (i.e., not just different inputs, as in a multi-arm RCT). Instead of trying to compare this type of noncounterfactual experiment to what it would look like as an RCT, however, it seems to us that we have to start with the goal in mind. What do we want to have learned in the end? What is the ideal empirical claim we may want to make? Or, more scientifically, what are the hypotheses we want to test? We propose:

(1) It is possible and useful to develop a reasonably easy-to-use protocol for assessing how a country's systems compare to the "good enough" bare-bones approach.
(2) A system reform that focuses on the "bare bones" subsystems can be implemented purposively.
(3) By intervening with a clear theory of change, careful iteration, and careful documentation, local leaders and policy entrepreneurs can "cause" the system to change; that is, they can "cause" a reform. (And, from an evaluation point of view, it will be possible to say something about that causality.)
(4) And, lastly in the chain, a system reform can lead to (can cause?) (sharply?) improved learning outcomes (reasonably quickly?).

Before we proceed, it may be good to admit that we cannot think of an approach—other than replication—that would be able to deal completely with the problem of external validity. But in that sense we are no worse off than with, say, RCTs. That is why we have not, for instance, stated hypothesis (4) as "Certain system reforms will lead to…" The use of the word "a" in "a system reform" is purposeful. So the causality we are talking about here is a bit more limited. It is an argument more of the type: "Mr.

[13] This is not to agree with the somewhat uncritical tendency to "RCT everything." RCTs as applied in development projects or, indeed, in most social science, face serious threats in generalizing from medical or physical science. Examples are self-selection of implementers who bring unobserved inputs into the treatment, selective crowding (sometimes unobserved) of other inputs into treatment sites, compensatory reduction of other inputs in the treatment group, "resentful demoralization" of the control group, and others. Of course, one can guard for some of these, and to some degree, but not perfectly. For a good early list of cautions about excessive faith in RCTs, from the public health literature (and hence probably not familiar to economists or educators), see Valadez (1992).

A murdered Mr. B (caused his death) and here is how." Not, "people with the characteristics of Mr. A tend to murder people with the characteristics of Mr. B." So we are proposing that one can address causality in some sense, but still not be able to generalize as well as the ideal *multiple-replicate, multi-site* RCT might. Even so, more can probably be done than initially would meet the criticism of the experimentalist.

The approach we propose would use a mix of:

- The experiential learning approach proposed by Pritchett, Samji, and Hammer (2013)
- Insights on causality from the Bradford Hill conditions (Hill 1965)
- Insights on causality from Julian Simon (Simon 1969, 1970)
- Insights from explicit systems modeling
- A judicial or, more accurately, "congressional hearing" approach to using third-party "juries" to read the evidence produced according to the previous three bullets.

One of the frustrating things about our proposed approach is that the language of the monitoring and evaluation descriptions will have to be full of phrases such as "not too many" or "enough" and so on. These criteria lack the comforting reassurance of a *t*-test, as well as the relief of basing conclusions on a conventional 99% confidence interval. And, it is why a judicial or "congressional hearing" approach might be a good idea as an alternative: a judgment- peer-experts-based, and transparent way to decide on the "not too many" sorts of issues.

MeE too: Crawling the design and (implementation) space

In "It's All About MeE: Using Structured Experiential Learning ('e') to Crawl the Design Space," Pritchett, Samji, and Hammer (2013) proposed seven principles for developing projects from which one can learn. ***Exhibit 3*** summarizes the principles and then shows how we propose to apply them to both evaluation and implementation in a system reform project. In that paper, the authors were really talking about crawling both the design and implementation space, because the proposed approach to implementation is an iterative one. We are directly using the word "implementation" just for emphasis.

Lest this idea seem too slavish or incestuous given the leadership of RISE's ILT, we point out that it is actually relatively difficult to have truly original thoughts in this area. And we could admit that we have adopted a framework that is not that different from what others have done, but that is attractive precisely because it was (partially) created by the lead of the ILT. In fact, however, observers long predating the ILT (and some of them are current or former RTI colleagues), such as Rondinelli (1983) and Brinkerhoff (1990), presented similar pointers. Andrews, Pritchett, and Woolcock (2012)—whose paper is a companion to Pritchett, Samji, and Hammer (2013)—acknowledged a long list of papers that made points similar to those in the MeE paper. In fact, it is difficult not to figuratively smile at how similar some of the decades-old lists of recommendations are to the current lists. Quoting extensively from Rondinelli (1983), for example:

> "adjusting planning procedures and methods of administration to the process of political interaction through which policies are actually made and carried out; adopting a learning-

> based approach to planning and administration in order to cope with uncertainty and complexity; … ; decentralizing to the appropriate level authority for planning and administering development activities; simplifying analysis and management procedures; encouraging rather than suppressing error detection and correction; and creating greater flexibility for development administrators to manage complex and uncertain systems by offering incentives for innovation, risk-taking, and learning" (p. 120).

More recent authors not only have lists that are somewhat self-consciously similar, but even offer highly simplified project planning tools that are quite attractive (Faustino and Booth 2014).

Therefore, given what we see as the essential impossibility of saying anything new in this area, we propose adopting the MeE approach, but with some specifications for a systems reform project.

One final digression before we present Exhibit 3. In the "bare bones" approach we propose, implicitly, that there are indeed some *sine qua non*s for running an education system in which children learn. And we would go as far as to claim that that framework holds for just about any way to run an organized or coherent education system (redundancy intended). That is, the view that there is indeed a "bare bones" approach that comes close to being a *sine qua non*, is somewhat top down or context free. But we are also espousing the MeE approach and we are sympathetic to Andrews, Pritchett, and Woolcock (2012), who predicate a discovery and iterative approach.

Is there a deep contradiction there or just an apparent one? Perhaps a bit of both, but we are not sure. In the "deep" sense, a bare-bones approach that (1) focused on explicit goal-setting by stating the expected size of movement, (2) monitored goal-seeking, and (3) offered both support and (mutual) accountability for actions would resemble any *intentional* system that succeeds. Further, it would mirror the individual units (i.e., firms) of unintentional systems (markets) that succeed. Humans have broadly understood how this works ever since explicit or intentional systems for organizing their activity arose. Obviously, however, non-intentional systems (markets, ecosystems) do not have these characteristics at the system level.

How do we reconcile the contradiction? It seems to us, in two ways. First, we reiterate that what we have described as a "bare bones" system is so fundamental and bedrock to any successful form of organized human activity that, surely, the iterative approach cannot be suggesting that systems experiment with, say, not having goals, or not having accountability. This view is somewhat abetted by RISE Vision Document No. 3 (Center for Global Development 2015c) which, although the authorship is not explicit, we assume is associated with the leadership of the RISE ILT. (Specifically, see the bullet points on page 1, but with an emphasis on systems that can "do" the content of those bullet points.) Second, we note that the feasible crawling and iteration have to be around (1) the *details* that one builds into the bare bones, and (2) *how* a reform gets there—how the ministries implement both the bare bones and the detail, play the political economy, etc. Both of these should contain considerable iteration, nomination of the problem(s) locally, etc.

Now for the Exhibit 3 application of the implementation principles to an education system reform:

Exhibit 3. Pritchett, Samji, and Hammer's (2013) seven principles applied to system reform

Action	Applied principle
Reason back from a stated goal of a stated size	Note, however, that clever implementers may narrow the goal so as to achieve success and thus create a generalizability problem. If a reform (even a bare-bones one as we suggest) is working on too narrow a goal, it may create an external validity problem. What is the right breadth of goal so as to make sure one can generalize? If a reform has a large impact on children's reading, in a whole country, we have reason to suspect that the reform, or the style of reform, is at least capable of broader impacts, even if it has not shown them yet. But if the index metric was narrowed to, say, fluency or decoding ability as opposed to reading more broadly, generalizability and external validity of results could become problematic?. Getting the goals right will take some negotiating and thinking.
Reverse-engineer to the instruments	Ideally, we believe, reforms should employ to the *minimum* (bare bones) set of instruments that prior experience and literature lead us to believe we can get away with. This is for practical reasons: We believe the countries that most need reforms don't have the band-width to absorb much complexity.
Design a project	In this case, we mean *design a reform as a project.* In the case of a system reform rather than a pilot project, this will require paying attention to the flow chart that links all the individual instruments, as well as the applied political economy and management processes needed to activate them.
Design by crawling the design space, but also the *implementation* space	We would add to this principle that in a systems reform project (though to some extent also in a simpler technical project), it is crawling the implementation space that ultimately designs. However, borrowing Pritchett, Samji, and Hammer's analogy, while it is true that "No battle plan survives the first contact with the enemy," we suspect Napoleon did not really go into battle without plans.[14] To make the project evaluable in the judicial or "congressional-hearing" sense that we propose—and to have a hope of some external validity—the initial design or plan, the crawling, and the redesigns or re-strategizing that results all have to be carefully documented. We also note that in a system reform project, as opposed to a project aimed at a technical solution, one has to be ready with flexible strategies. This might mean that "using variations within a project to identify differentials in the efficacy of the project on inputs and outputs for real-time feedback into project implementation lowers evaluation

[14] Because, while it may be true that Napoleon said "Engage with the enemy and see what happens," he also said "Read over and over again the campaigns of Alexander, Hannibal, Caesar, Gustavus, Turenne, and Frederic the Great. This is the only way to become a great general."

Action	Applied principle
	cost and feedback loop time" (Pritchett, Samji, and Hammer, p. 35). But, it could also mean retreating, re-strategizing, re-marketing, rebuilding coalitions, etc., to get the technical content past the ideological and political opposition, as well as the inertia. Even technical projects run into managerial and political opposition. In a system reform project, if individuals and groups *don't* push back, probably nothing meaningful is under way.[15] A last point is that, in our experience, funders are often quite willing to be flexible if, upon execution, some aspects of the project do not work as well as intended. They may not be happy to be told in advance that the implementer is guessing about what will work. But they know it (to some degree) even if they will not admit it, and so will often be reasonable during implementation, especially if the cause of implementation problems is *force majeure*.
Specify the design space and select alternative designs	We are not too sure this is as important for a systems reform implementation project as for a technical project. (Incidentally, neither are we sure how this differs from an experiment with multiple arms, other than that the approach proposed is nimbler and less expensive—but perhaps weaker on causality?) In a system reform project, we'd propose to start with a fairly well-specified bare-bones system that, based on lots and lots of previous technical projects, we think can do the trick. The crawling, then, is around the implementation space: Who are the allies, who is likely the opposition, how will you market the ideas, how you will neutralize the opposition? There may be some semi-technical options in the design space as well, if options for the management of the process (how much technical assistance, of what type?) can be considered technical design.
Strategically crawl your design space: Pre-specify how implementation and learning will be synchronized	This strikes us as perhaps the most important aspect if one is to make the whole project evaluable using the judicial approach we propose. Also, the issue here is to document the redesign. The crawling is not just initial (as noted by Pritchett, Samji, and Hammer—we want to give this extra emphasis in the case of a systems reform project). The crawling happens over time and, since how the options branch out is not foreseeable, at each junction the nature of the options, and the decision taken, has to be documented, for a judicial approach to establishing causality to work.
Implement	No comment other than what has been noted elsewhere.

[15] As George Bernard Shaw put it: "A man never tells you anything interesting until you contradict him." The paper by Faustino and Booth (2014), on "working politically," has much useful evidence on how groups push back, in a variety of reforms.

The Bradford Hill conditions

Bradford Hill, who seems to be unknown to economists, was one of the most influential epidemiologists of the 20th century. He made pioneering contributions to the development and design of RCTs. But epidemiologists, of course, have special problems in dealing with causality. As "the other Luis Crouch," a biostatistician and epidemiologist, explained (when chided by his father about the fact that epidemiologists tell us one year not to eat eggs, and then the next year that it is ok): "Daaad, we do mostly correlational analysis, as opposed to mostly randomized trials, not because we *want* to, but because we *have* to. I can't give some women cadmium, and others not, to see what happens."

The proposed DFID systems research fits into the category of "The reason we don't do it is not because we don't want to, but because we can't." Stated another way, for various reasons we *can't* do high-*N,* high-replication, counterfactual studies of systems reforms in development settings, even if we *wanted* to. (Many researchers, such as Woolcock [2009], have documented those reasons.) Because of this imperative, what epidemiologists have to say about causality, especially when counterfactuals are not possible, is interesting. Hill (1965) summarized some conditions that have to be satisfied for one to begin to talk about causality, when counterfactual experimentation is not possible or desirable. These conditions have become known as the "Bradford Hill conditions." Their use as a checklist was foreseen and decried by Hill himself, but that will not stop us. What is true is that Hill himself did not consider them a substitute for counterfactuals with randomization but as the most reasonable make-do. Others have written about how to apply the conditions based on how they compare to the ideal of counterfactuals (Höfler 2005). ***Exhibit 4*** lists them (omitting one that we were not able to understand), and discusses which might be feasible in a one-off, whole-system experiment, and how.

Exhibit 4. Possible use of the Bradford Hill conditions in a systems reform project

Condition	Possible applicability to the project
Strength	The strength condition is generally meant to apply to high-*N* observational studies and to refer to truly large (or small) observations. If exposure to something (a chemical) is associated with a 200-fold increase in something else (cancer), one has some reason to suspect something other than correlation is going on. And, while a high correlation does not imply causation, a low one tends to argue against it. (This argument assumes the other conditions are analyzed.) In an *N* = 1 experiment, the strength of impact can still be noted. In the present project, one would hopefully be looking for serious impact—maybe not 200-fold, but certainly 1 SD and higher. This should be stated in the project's definition documents and its theory of change.
Consistency	Has the association been observed in many contexts? This, of course, is the same problem of external validity, even with RCTs. Hill notes that even in cases of low replication, sufficient strength (along with some of the other conditions here) can justify a causal conclusion. In the project we propose, one would do well to check whether the reform is having a consistent effect upon a variety of outcomes, based on an explicit ex-ante claim; and not just upon one outcome, or a predictably inconsistent impact.

Condition	Possible applicability to the project
Specificity	Specificity means one cause, one effect, but without overdoing the one-to-one correspondence (i.e., sometimes there may be more than one effect per cause). It seems to us that this condition is not usable in a systemic reform causal evaluation, almost by definition.
Temporality	Temporal factors are evident and relevant for a systems reform project. Were there pre-existing trends? Can one trace a causal path step by step over time?
Dose response or biological gradient	If there is evidence of a gradient, rather than a binary response, the presumption of causality is higher. This would seem inapplicable to a system reform project, but, in a manner similar to what is proposed in Pritchett, Samji, and Hammer (2013), natural variation in fidelity, and strength of application of some of the features of the reforms, might help. Again, in which aspects of the reforms, and for which outcomes, one should expect this, should be explicitly laid out, ex ante, in the project definition and theory of change. One need not necessarily lay out variations in intensity in the design; but one should specify in what processes and inputs variation is more likely, and which outcomes would respond in a gradient, so as to have a pre-stated hypothesis, not ex-post rationalization.
Biological (pedagogical in our case) plausibility and coherence[16]	Evidently, if one knows (or has collateral evidence about) how smoke affects lung cells, this strengthens observational studies, even in the absence of counterfactual experiments. In macroeconomics no one, to our knowledge, has experimented to see whether inordinate increases in the money supply cause inflation—yet the enormous majority of economists would agree with a causal version of this statement, because the logic is so clear. In an education reform, the changes would have to be plausible in various ways. We know, for instance, that, in learning certain skills, repetition and drilling are very helpful. Similarly, we know that branching out gradually but strongly and systematically from a base of practiced skills and solid knowledge is helpful (related to the "overambitious curricula" gradient issue discussed by Pritchett and Beatty 2012), and we now know the neurobiological reasons. Reforms that can trace a causal path from a system change (more frequent coaching of teachers that emphasizes the skills needed to drill children and start from where they are) would follow a plausible pedagogical path. For other aspects of the reform, one will be less able to sketch out the plausible pedagogical path. Incidentally, a sort of stepwise reasoning based on specific steps, each one of which is very causally evident but not in providing direct evidence, is similar to the "cause" determination in judicial proceedings. ("Mr. A did buy a hammer, he did bring the hammer to Mr. B's house, he was observed entering the house, no one else was observed entering, Mr. B was killed with a hammer, Mr. A was covered in blood when he left Mr. B's houses, and the blood found on the

[16] Hill presents these as separate conditions. We don't understand why. Some commentators claim to see a difference.

Condition	Possible applicability to the project
	hammer was Mr. B's.") The evidence can be circumstantial, but if there is enough of it, and is argued well enough, it can prove a case beyond a reasonable doubt.[17]
Experimentation[18]	This type of investigation means not simply taking advantage of natural variation, but instead carrying out purposeful and systematic experimentation, though short of using counterfactuals. It does not seem possible in a systems reform project, for the reasons noted above. *Natural* variation can be observed, recorded, and entered into the evidence stream, however, as noted above and in Pritchett, Samji, and Hammer (2013).

Julian Simon's four-point list

The economist Julian Simon wrote a couple of obscure papers on causality in 1969 and 1970. The papers are also somewhat pretentious and hard to understand. However, at the end, there is a very neat set of conditions for how one can judge causality in complex situations where counterfactual experimentation is not possible. These are fairly similar to the Bradford Hill conditions but are more succinct: (1) strength of correlation, (2) fewness of side conditions, (3) assessment of as many reasons as possible as to why the relationship could be spurious, and (4) consistency with a theory that has claims to other well-established causal relationships. The frustrating thing is that, after laying out a nice checklist, Simon concludes: "This heuristic definition is a checklist test against which one can compare a given relationship. If the relationship ***seems*** to meet ***most*** of the checklist criteria ***reasonably*** well, then you ***probably*** will (and ought to) call the relationship 'causal'; if not, not" (Simon 1969, p. 23). That is not as satisfying as a crisp, formal *t*-test from an experiment.

Insights from systems modeling and operations research

Noncounterfactual situations are frequently addressed via modeling. In economics, various thinkers have presented approaches that try to get at causality via formal modeling (e.g., Granger 1969; Orcutt 1952). The problem with such approaches for our situation, and indeed almost any approach based also on operations research and simulation modeling, is that one does not have the coefficients. One could argue that to get the coefficients one needs either engineering studies (as in operations research models) or good estimates of the coefficients, which would require an identification strategy, which in turn (at the limit) would requires RCTs, or some reasonably strong proxy. At the very least, serious

[17] Against the popular impression created by movies and thrillers (which seem to inform even the educated layperson), circumstantial evidence can be perfectly valid in creating a decision. In the United States, the Supreme Court has noted that "circumstantial evidence is intrinsically no different from testimonial [direct] evidence" (Holland v. United States, 1954). Obviously, though, the circumstantial evidence has to be good enough, as determined by the judicial process, and "must exclude every reasonable hypothesis as to the defendant's innocence," as presented by the evidence (Scheb and Scheb 2012, p. 197). In our case, a "jury" must be able to exclude every other reasonable explanation for impact upon presentation of all the possible evidence.

[18] Some commentators seem to interpret this as a call for counterfactuals. We disagree with this interpretation because otherwise a lot of the other conditions seem relatively unnecessary.

ordinary least squares (OLS) and time series work would be in order. So we would be more or less back where we started.

But there are approaches, not based on coefficients, that perhaps can offer some useful insights. Hanson (2015), at the ILT meeting on May 1–2, 2015, provided some useful examples of non-coefficient-based systems analysis from the health arena. Particularly interesting was a model of the decision to increase military physician salaries in Ghana, and the immediate and ripple effects this created (documented and modeled in Agyepong et al. 2012). We propose that ex-ante and ex-post modeling of this kind could perhaps be used as part of the theory of change and a somewhat formalized documentation of the crawling of the design space. Also as an input cited by Hanson, the UK's Medical Research Council (n.d.) recommended that a modeling approach can be useful in formalizing the theory of change:

> "the value of a causal modelling approach is that it makes explicit the choice of intervention points and associated measures along the causal pathway. This allows researchers to assess why interventions are effective or not, as well as how effective they are, and to have greater confidence in modelling long-term disease outcomes from shorter term changes in behavior" (p. 17).

A possibility: A congressional hearing or judicial approach

Non-experimental situations, as noted, do not allow for the crispness of statistical tests. The language of evaluation then becomes vague. And it uses (or should use) processes similar to the accumulation of circumstantial evidence in establishing causality in judicial proceedings and jury trials. So, what might be the possibility of generalizing from legal processes, and what might the steps be? Papers from practical evaluators outline a set of steps; we would innovate by adding the formal one of actually creating a formal panel to pass judgement on causality (perhaps "causality lite") and hence generalizability.

We realize that this is a far more complex proposition than simply running trials, but we agree with Woolcock (2009) that "a truly rigorous evaluation is one that deploys the best available assessment tools at intervals that correspond to the shape of a project's known (via experience, empirical evidence, or inferred on the basis of sound theory) impact over time" (p. 2) or, to quote extensively,

> "It is my central contention that a truly rigorous evaluation is one that deploys the full arsenal of social sciences methods as part of a strategy focused on achieving – within the prevailing political, economic and logistical constraints – an optimal match between these methods (or combination of methods) and the type of problem to which the project (or policy) is responding. The policy problem must generate the methodological response, not the other way around, just as the available policy/project 'solutions' should not determine which problems are addressed. Put differently, individual methods per se are not 'rigorous'; randomisation is not inherently a generator of superior data. Rather, methods become rigorous when they comprehensively generate valid and reliable data that is able to speak to the specific characteristics of the problem in question" (p. 5).

Thus, while Woolcock's paper emphasizes the problem of coming to conclusions too early, the paper also argues for a complete approach that includes as many forms of evidence as possible.

A summary of the steps by one of many authors (in this case, Mayne 1999) working in the field of complex evaluation makes clear the similarity of the process to that of establishing "good" circumstantial evidence:

> "A reasonable case that a program has indeed made a difference would entail:
>
> - well-articulated presentation of the context of the program and its general aims;
> - presentation of plausible program theory leading to the overall aims (the logic of the program has not been disproven, i.e. there is little or no contradictory evidence and the underlying assumptions still appear to remain valid);
> - highlighting the contribution analysis indicating there is an association between what the program has done and the outcomes observed; and
> - pointing out that the main alternative explanations for the outcomes occurring, such as other related programs or external factors, have been ruled out or clearly have only had a limited influence" (Mayne 1999, p. 16).

A more elaborate or detailed list is presented in Mayne (2011, p. 63) and could be borrowed/adapted.

To this we would add two aspects:

- Evidence that is as rigorous as possible, given the project design, the evidence of impact, the documentation of the crawling of the design space, interviews with key actors, etc. This means that, for any aspect whatsoever where quantitative evidence, experimental or quasi-experimental, can be adduced, it should be, and the country proposal in question should introduce in advance **all** the forms of evidence to be used.
- The formality of a true jury of peers to hear the evidence and pronounce on the four research hypotheses listed earlier. The jurors would be experienced in the subject matter and in evaluation, and would be peers of the implementers, but would be independent and a third party.

The first point above is fairly obvious so we will not elaborate on it. The second point is less so.

After we came up with the analogy of using a judicial or congressional hearing approach, knowing that it is truly difficult to have original thoughts, and being lazy, we searched the literature on the assumption that we would find some practical tips, or do's and don'ts. It turns out (not surprisingly) that others *have* thought and written about this type of third-party appraisal. It seemed to us that some of the early attempts were naïve, and were excessively modeled on the precise proceedings of a trial by jury, but that with some corrections, the basic idea could be useful. The approach was outlined or discussed by Owens (1973), Owens and Hiscox (1977), and Wolf (1979). Friendly critics of the approach, such as Worthen and Rogers (1980), have sounded certain cautions.

The approach would consist of the following steps or aspects, which we have adapted from the literature (by paying careful attention to the proposals, the experiences, and the critique):

(1) Ensure (as noted elsewhere) that the crawl of the design and implementation space is documented, with the theory of why each decision was made at each point laid out as carefully as possible.
(2) If it is available and relevant, bring in quantitative evidence, including RCTs. While some of the writers on the approach seem to eschew quantitative evidence, we believe that it can have value. Certainly a reform aimed at improving learning outcomes will have to collect evidence on learning outcomes. It seems there is little to dispute in that assertion.
(3) Empanel a group of experienced researchers and peers, including local experts as much as possible, *from the beginning*.
(4) The country research team (CRT) designs and then agree *ahead of time* with the panel of researchers and peers on the rules of permissible evidence*.* The various lists and ideas from noncounterfactuals can provide some sense of what is needed. In addition, of course, experimental and quasi-experimental evidence should be admissible.
(5) Link the panel to the advisory or delivery board that seems to be proposed for RISE: It could the local and more micro variant of such a board.
(6) At the end of year 1, year 2, year 4 and year 6, convene the panel to review the evidence, supply corrective input, and decide which lines of argument are plausible (or likely to be plausible if it is in years 1 and 2) and which are less so, and what aspects seem to permit generalizability.
(7) Interview key actors, rather than just reviewing of the documentary evidence.
(8) Make the process public, although the actual meetings and deliberations need not be so. But the idea has to be to release the results as quickly as possible.

Part of the reason to appoint an independent panel is that the CRT is a committed and engaged actor, although this point does depend on the type of model for the reform project chosen: If the CRT is a totally detached observer and researcher, then in some sense the CRT is the panel. A panel can bring objectivity as well as subtlety and a variety of viewpoints to the process of reading the evidence.

The process, it is argued by advocates (e.g., Owens and Hiscox 1977), can lead to:

- A clearer specification of the issues to evaluate because of the need for the researchers and implementers to present to a formal panel, with an advance plan.
- Public visibility and generation of support for the ideas that work.
- Better connection of evaluators and CRT to implementers.
- Support being provided in particular to interventions that are deemed somewhat controversial.

Critics, including friendly critics, have issued some cautions:

- Naively copying judicial proceedings, including a judicial approach to cross-examining testimony, can be counterproductive. As we have said and as many others also have suggested, a better analogy is that of congressional testimony.

- Judicial proceedings are adversarial by nature. It is not clear that an adversarial approach (two sides arguing, and the jury then choosing) is productive.
- A particular aspect to avoid is the idea that there is an indictment to be produced, or guilt to be found. Thus the need to really talk more about a congressional approach than a judicial one.

We think the idea is at least worth exploring even if some of the initial applications were a bit naive. Note that one of the initial practical applications—to an evaluation of Hawaii's statewide "3 for 2" (three teachers for two classrooms) team-teaching program—did win the prize for Evaluation of the Year from the American Educational Research Association (Worthen and Rogers 1980).

Conclusion

In the introduction to this paper, we posed four questions.

(1) Absent the means to fund national-scale direct interventions, are there system-level changes that could produce large effect size improvements in learning on a broad scale?

(2) Is there a limited set of system changes that are more directly linked to (and therefore more likely to lead to) improved learning?

(3) Is there a set of high-leverage policy support interventions that can increase the likelihood that system changes more directly linked to improved learning outcomes can be successfully implemented?

(4) Could a well-articulated causal chain and theory of action be used to document how reform support activities led to systems changes, which in turn result in improved learning outcomes? And could a research design that relies on a judicial approach be employed to verify the underlying causal relationships?

Our combined answers to these questions we believe offer some insight into how one could approach the research into education system improvement that is the objective of RISE. The research would require the following actions: carefully selecting the aspects of the system that need to change, working with the local actors who are intervening strategically to support the institutional reforms needed to improve system capacity related to those "bare bones" functions, explicitly mapping the anticipated causal chain that ties those system capacities to improved teaching and learning on a national scale, investing in documenting and evaluating how the pieces of that puzzle do or do not fall into place, and measuring learning outcomes all along the way. We feel that this research design would make it possible to learn a great deal about how system changes contribute to improved outcomes.

References

Agyepong, Irene, Augustina Kodua, Sam Adjei, and Taghreed Adam. 2012. "When 'Solutions of Yesterday Become Problems of Today': Crisis-Ridden Decision Making in a Complex Adaptive System (CAS)—the Additional Duty Hours Allowance in Ghana." *Health Policy and Planning* 27: iv20–iv31. http://dx.doi.org/10.1093/heapol/czs083

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2015. *Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets.* Policy Research Working Paper 7226. Washington, DC: Human Development and Public Services Team, Development Research Group, World Bank. http://dx.doi.org/10.1596/1813-9450-7226

Andrews, Matthew, Lant Pritchett, and Michael Woolcock. 2012. "Escaping Capability Traps Through Problem-Driven Iterative Adaptation (PDIA)." Faculty Research Working Paper Series, No. RWP12-036. Cambridge, MA: John F. Kennedy School of Government, Harvard University.

ASER Centre. 2015. *Annual Status of Education Report (Rural): 2014.* New Delhi, India. http://img.asercentre.org/docs/Publications/ASER%20Reports/ASER%202014/fullaser2014mainreport_1.pdf [All ASER reports from 2005 are available from http://www.asercentre.org/p/51.html?p=61]

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, and Michael Walton. 2012. *Effective Pedagogies and a Resistant Education System: Experimental Evidence on Interventions to Improve Basic Skills in Rural India.* Unpublished manuscript, Jamil Abdul Lateef Poverty Action Lab, Massachusetts Institute of Technology, Cambridge, MA.

Blimpo, Moussa P., David K. Evans, and Nathalie Lahire. 2015. *Parental Human Capital and Effective School Management: Evidence from The Gambia.* Policy Research Working Paper 7238. Washington, DC: Education Global Practice Group & Africa Region, World Bank. http://dx.doi.org/10.1596/1813-9450-7238

Bostock, Guy, and Rakusin, Mitchell. 2014. "Using Learner Literacy Data to Improve Early Grade Learning Outcomes in Zambia." Paper presented at the 32nd Annual Conference of the Association for Educational Assessment in Africa, Livingstone, Zambia, August 11–15, 2014.

Brinkerhoff, Derick W., Arthur A. Goldsmith, Marcus D. Ingle, and S. Tjip Walker. 1990. "Institutional Sustainability: A Conceptual Framework." In *Institutional Sustainability in Agriculture and Rural Development: A Global Perspective,* edited by Derick W. Brinkerhoff and Arthur A. Goldsmith, 19–49. New York: Praeger. http://pdf.usaid.gov/pdf_docs/PNABG576.pdf

Brown, Byron W., and Daniel H. Saks. 1986. "Measuring the Effects of Instructional Time on Student Learning: Evidence from the Beginning Teacher Evaluation Study." *American Journal of Education* 94 (4): 480–500. http://dx.doi.org/10.1086/443863

Bruns, Barbara, Deon Filmer, and Harry A. Patrinos. 2011. *Making Schools Work: New Evidence on Accountability Reforms.* Washington, DC: World Bank. http://dx.doi.org/10.1596/978-0-8213-8679-8

Bruns, Barbara, and Javier Luque. 2014. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington, DC: World Bank. http://dx.doi.org/10.1596/978-1-4648-0151-8

Center for Global Development. 2015a. *The Pivot from Schooling to Education.* Research on Improving Systems of Education (RISE) Vision Document No. 1. Washington, DC. http://www.ukcds.org.uk/sites/default/files/content/resources/RISE%20Vision%20document%201.pdf

Center for Global Development. 2015b. *Ambitious Learning Goals Need Audacious New Approaches.* Research on Improving Systems of Education (RISE) Vision Document No. 2. Washington, DC. http://www.ukcds.org.uk/sites/default/files/content/resources/RISE%20Vision%20document%202.pdf

Center for Global Development. 2015c. *Why Research into Education Systems Is Needed.* Research on Improving Systems of Education (RISE) Vision Document No. 3. Washington, DC. http://www.ukcds.org.uk/sites/default/files/content/resources/RISE%20Vision%20document%203.pdf

Chen, Dandan. 2011. “School-Based Management, School Decision-Making and Education Outcomes in Indonesian Primary Schools.” Policy Research Working Paper 5809. Washington, DC: Education Sector Unit, East Asia and Pacific Region, World Bank. http://dx.doi.org/10.1596/1813-9450-5809

Costa, Leandro Oliveira, and Martin Carnoy. 2015. “The Effectiveness of an Early Grades Literacy Intervention on the Cognitive Achievement of Brazilian Students.” *Educational Evaluation and Policy Analysis.* http://dx.doi.org/10.3102/0162373715571437

Crouch, Luis, and F. Henry Healey. 1997. *Education Reform Support. Volume One: Overview and Bibliography.* SD Publication Series, Paper No. 47; Advancing Basic Education and Literacy (ABEL) Technical Paper No. 1. Washington, DC: Office of Sustainable Development, Bureau for Africa, USAID. http://pdf.usaid.gov/pdf_docs/PNACA717.pdf

Cubberley, Ellwood. 1919. *Public Education in the United States.* Cambridge, MA: Riverside Press.

Dahal, Mahesh, and Quynh Nguyen. 2014. *Private Non-State Sector Engagement in the Provision of Educational Services at the Primary and Secondary Levels in South Asia: An Analytical Review of Its Role in School Enrollment and Student Achievement.* Policy Research Working Paper 6899. Washington, DC: South Asia Region Education Unit, World Bank. http://dx.doi.org/10.1596/1813-9450-6899

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. 2011. *School Inputs, Household Substitution, and Test Scores.* Policy Research Working Paper 5629. Washington, DC: Human Development and Public Services Team, Development Research Group, World Bank. http://dx.doi.org/10.1596/1813-9450-5629

DeStefano, Joseph. 2011. *Information for Education Policy, Planning, and Management: Summary of the Data Capacity Assessments Conducted in the Philippines, Ghana, and Mozambique.* Prepared for USAID under the Education Data for Decision Making (EdData II) Project, Task Order No. EHC-E-11-04-00004 (RTI Task 11). Research Triangle Park, NC: RTI International. https://www.eddataglobal.org/capacity/index.cfm?fuseaction=pubDetail&ID=342

DeStefano, Joseph, and Luis Crouch. 2006. *Education Reform Support Today.* Prepared for USAID under the Educational Quality Improvement Program 2 (EQUIP2), Cooperative Agreement No. GDG-A-00-03-00008-00. Washington, DC: Academy for Educational Development (AED). http://pdf.usaid.gov/pdf_docs/PNADQ913.pdf

Dowd, Amy Jo. 2014. "Succeeding Where Others Stumble? Lessons from the First Half Decade of Literacy Boost." Paper presented at the 2014 annual conference of the Comparative and International Education Society, Toronto, Ontario, Canada, March 10–15, 2014.

Dowd, Amy Jo, Elliott Friedlander, Jarret Guajardo, Noah Mann, and Lauren Pisani. 2013. *Literacy Boost Cross Country Analysis Results.* Washington, DC: Department of Education and Child Development, Save the Children. http://www.meducationalliance.org/sites/default/files/literacy_boost_cross_country_analysis_results.pdf

Elmore, Richard F. 1996. "Getting to Scale with Good Educational Practice." *Harvard Educational Review* 66(1): 1–27. http://dx.doi.org/10.17763/haer.66.1.g73266758j348t33

Evans, David, and Anna Popova. 2015. "How Systematic Is that Systematic Review? The Case of Improving Learning Outcomes." *Development Impact* (World Bank blog). http://blogs.worldbank.org/impactevaluations/how-systematic-systematic-review-case-improving-learning-outcomes

Faustino, Jaime, and David Booth. 2014. *Development Entrepreneurship: How Donors and Leaders Can Foster Institutional Change.* Working Politically in Practice Series, Case Study No. 2. London: Overseas Development Institute; and San Francisco: The Asia Foundation. http://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9384.pdf

Finnigan, Kara, and Jennifer O'Day. 2003. *External Support to Schools on Probation: Getting a Leg Up?* Chicago: University of Chicago Consortium on Chicago School Research. https://ccsr.uchicago.edu/publications/external-support-schools-probation-getting-leg

Foley, Ellen, Jacob Mishook, Joanne Thompson, Michael Kubiak, Jonathan Supovitz, and Mary Kay Rhude-Faust. n.d. *Beyond Test Scores: Leading Indicators for Education.* Providence: Annenberg Institute for School Reform, Brown University. http://annenberginstitute.org/pdf/LeadingIndicators.pdf

Gillies, John, and Jessica Jester Quijada. 2008. "Opportunity to Learn: A High-Impact Strategy for Improving Educational Outcomes in Developing Countries." Working Paper. Prepared for USAID under the Educational Quality Improvement Program 2 (EQUIP2), Cooperative Agreement No. GDG-A-00-03-00008-00. Washington, DC: Academy for Educational Development (AED). http://www.equip123.net/docs/e2-OTL_WP.pdf

Goyal, Sangeeta, and Priyanka Pandey. 2013. "Contract Teachers in India." *Education Economics* 21(5): 464–484. http://dx.doi.org/10.1080/09645292.2010.511854

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37(3): 424–438. http://dx.doi.org/10.2307/1912791

Hanson, Kara. 2015. *Researching Systems: Some Approaches from Health Systems Research.* Presentation prepared for the RISE program Intellectual Leadership Team meeting, May 1–2, 2015.

He, Fang, Leigh L. Linden, and Margaret McLeod. 2009. *A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trial.* New York, NY: Columbia University. http://www.leighlinden.com/Teach%20Children%20to%20Read.pdf

Hill, Austin Bradford. 1965. "The Environment and Disease: Association or Causation?" *Journal of the Royal Society of Medicine* 108(1): 32–37 [original publication: 58(5): 295–300]. http://dx.doi.org/10.1177/0141076814562718

Höfler, Michael. 2005. "The Bradford Hill Considerations on Causality: a Counterfactual Perspective." *Emerging Themes in Epidemiology* 2(1): 11. http://dx.doi.org/10.1186/1742-7622-2-11

Holland v. United States, 348 U.S. 121, 75 S. Ct. 127, 99 L. Ed. 150 [1954].

Jukes, Matthew C. H., Elizabeth L. Turner, Katherine E. Halliday, Sharon Wolf, Stephanie Simmons Zuilkowski, Simon J. Brooker, and Margaret M. Dubeck. 2015. "Teacher Professional Development and Text Messages Support for Improved Literacy Instruction in Kenya: An Experimental Evaluation." Unpublished manuscript.

Jung, Haeil, and Amer Hasan. 2014. *The Impact of Early Childhood Education on Early Achievement Gaps: Evidence from the Indonesia Early Childhood Education and Development (ECED) Project.* Policy Research Working Paper 6794. Washington, DC: Education Sector Unit, East Asia and the Pacific Region, World Bank. http://dx.doi.org/10.1596/1813-9450-6794

Lucas, Adrienne M., Patrick J. McEwan, Moses Ngware, and Moses Oketch. 2014. "Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda." *Journal of Policy Analysis and Management* 33(4): 950–976. http://dx.doi.org/10.1002/pam.21782

Mayne, John. 1999. “Addressing Attribution Through Contribution Analysis: Using Performance Measures Sensibly.” Discussion Paper. Ottawa: Office of the Auditor General of Canada. http://www.oag-bvg.gc.ca/internet/docs/99dp1_e.pdf

Mayne, John. 2011. “Contribution Analysis: Addressing Cause and Effect.” In *Evaluating the Complex: Attribution, Contribution, and Beyond,* Comparative Policy Evaluation Volume 18, edited by Kim Forss, Mita Marra, and Robert Schwartz, 53–96. New Brunswick, NJ: Transaction Publishers.

Medical Research Council [UK]. n.d. *Developing and Evaluating Complex Interventions: New Guidance.* Swindon, Wiltshire, UK. http://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance

Mourshed, Mona, Chinezi Chijioke, and Michael Barber. 2010. *How the World’s Most Improved School Systems Keep Getting Better.* London: McKinsey & Company. http://www.mckinsey.com/client_service/social_sector/latest_thinking/worlds_most_improved_schools

Mugo, John, Amos Kaburu, Charity Limboro, and Albert Kimutai. 2011. *Are Our Children Learning? Annual Learning Assessment Report.* Nairobi: Uwezo Kenya. http://www.uwezo.net/wp-content/uploads/2012/08/KE_2011_AnnualAssessmentReport.pdf

Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. “Contract Teachers: Experimental Evidence from India.” NBER Working Paper No. 19440. Cambridge, MA: National Bureau of Economic Research. http://dx.doi.org/10.3386/w19440

Nielsen, Dean. 2013. *Going to Scale: The Early Grade Reading Program in Egypt, 2008–2012.* Prepared for USAID under the Education Data for Decision Making (EdData II) project, Data for Education Programming in Asia and the Middle East (DEP-ASIA/ME), Task Order No. AID-OAA-BC-11-00001 (RTI Task 15). Research Triangle Park, NC: RTI International. https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=606

Orcutt, Guy H. 1952. "Actions, Consequences, and Causal Relations." *Review of Economics and Statistics* 34: 305–313. http://dx.doi.org/10.2307/1926858

Owens, Thomas R. 1973. “Educational Evaluation by Adversary Proceeding.” In *School Evaluation: The Politics and Process,* edited by Ernest R. House. Berkeley, CA: McCutchan Publishing.

Owens, Thomas R., and Michael D. Hiscox. 1977. “Alternative Models for Adversary Evaluation: Variations on a Theme.” Paper presented at the Annual Meeting of the American Educational Research Association. New York, NY, April 4–8, 1977. http://files.eric.ed.gov/fulltext/ED136425.pdf

Piper, Benjamin, Evelyn Jepkemei, Dunston Kwayumba, and Kennedy Kibukho. 2015. “Kenya’s ICT Policy in Practice: The Effectiveness of Tablets and E-Readers in Improving Student Outcomes.” Manuscript under review, *Forum for International Research in Education*.

Piper, Benjamin, and Medina Korda. 2010. *Early Grade Reading Assessment (EGRA) Plus: Liberia. Program Evaluation Report.* Prepared for USAID/Liberia under the Education Data for Decision Making (EdData II) project, Task Order No. EHC-E-06-04-00004-00 (RTI Task 6). Research Triangle Park, NC: RTI International. http://pdf.usaid.gov/pdf_docs/pdacr618.pdf

Piper, Benjamin, Stephanie Simmons Zuilkowski, and Abel Mugenda. 2014. “Improving Reading Outcomes in Kenya: First-Year Effects of the PRIMR Initiative.” *International Journal of Educational Development* 37: 11–21. http://dx.doi.org/10.1016/j.ijedudev.2014.02.006

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Armida Alishjabana, Arya Gaduh, and Rima Prama Artha. 2011. “Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia.” Policy Research Working Paper 5795. Washington, DC: Human Development Sector Department, East Asia and Pacific Region, World Bank. http://dx.doi.org/10.1596/1813-9450-5795

Pritchett, Lant. 2013. *The Rebirth of Education: Schooling Ain’t Learning.* Washington, DC: Center for Global Development. http://www.cgdev.org/sites/default/files/rebirth-education-introduction_0.pdf

Pritchett, Lant, and Amanda Beatty. 2012. “The Negative Consequences of Overambitious Curricula in Developing Countries.” Working Paper 293. Washington, DC: Center for Global Development. http://www.cgdev.org/files/1426129_file_Pritchett_Beatty_Overambitious_FINAL.pdf

Pritchett, Lant, Salimah Samji, and Jeffrey Hammer. 2013. “It‘s All About MeE: Using Structured Experiential Learning (“e”) to Crawl the Design Space.” Working Paper No. 322. Washington, DC: Center for Global Development. https://usaidlearninglab.org/sites/default/files/resource/files/its-all-about-mee_1.pdf

Roberts, John. 2004. *The Modern Firm: Organizational Design for Performance and Growth.* Oxford, UK: Oxford University Press.

Rondinelli, Dennis. 1993. *Development Projects as Policy Experiments: An Adaptive Approach to Development Administration.* New York: Routledge.

RTI International. 2011. *Task Order 7, NALAP [National Literacy Acceleration Program] Formative Evaluation Report, Ghana*. Prepared for USAID under the Education Data for Decision Making (EdData II) project, Task Order No. EHC-E-07-04-00004-00 (RTI Task 7). Research Triangle Park, NC. Washington, DC: USAID. https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=317

RTI International. 2013. *Report on the Pilot Application of LQAS in Ghana to Assess Literacy and Teaching in Primary Grade 3.* Prepared for USAID under the Education Data for Decision Making (EdData II) project, Task Order No. EHC-E-07-04-00004-00 (RTI Task 7). Research Triangle Park, NC. http://pdf.usaid.gov/pdf_docs/pa00k2dt.pdf

RTI International. 2014. *Research on Reading in Morocco: Analysis of the National Education Curriculum and Textbooks. Final Report – Component 1.* Prepared for USAID under the Education Data for Decision Making (EdData II project), Task Order No. AID-OAA-BC-11-00001 (RTI Task 15). Research Triangle Park, NC. https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=703

Scheb, John M., and John M. Scheb II. 2012. *Criminal Procedure.* Belmont, California: Wadsworth.

Schuh Moore, Audrey-Marie, Anne Smiley, Joseph DeStefano, and Elizabeth Adelman. 2012. "The Right to Quality Education: How Use of Time and Language of instruction Impact the Rights of Students." *World Studies in Education* 13(2): 67–86. http://dx.doi.org/10.7459/wse/13.2.06

Serra, Danila, Abigail Barr, and Truman Packard. 2011. "Education Outcomes, School Governance and Parents' Demand for Accountability: Evidence from Albania." Policy Research Working Paper 5643. Washington, DC: Human Development Economics Unit, Europe and Central Asia Region, World Bank. http://dx.doi.org/10.1596/1813-9450-5643

Simon, Julian L. 1969. "Untangling the Puzzle of Causality." Unpublished manuscript. http://www.juliansimon.com/writings/Articles/CAUSALI2.txt

Simon, Julian L. 1970. "The Concept of Causality in Economics." *Kyklos* 23(2): 226–254. http://dx.doi.org/10.1111/j.1467-6435.1970.tb02556.x

Ucelli, Marla R., and Ellen L. Foley. 2004. "Results, Equity and Community: The Smart District." *Voices in Urban Education* Fall: 5–10. Providence, RI: Annenberg Institute for School Reform, Brown University. http://vue.annenberginstitute.org/sites/default/files/issuePDF/VUE5.pdf

Valadez, Joseph. 1992. *Assessing Child Survival Programs: A Test of Lot Quality Assurance Sampling in a Developing Country.* Cambridge, MA: Harvard University Press.

Wang, Liang Choon. 2011. "Shrinking Classroom Age Variance Raises Student Achievement: Evidence from Developing Countries." Policy Research Working Paper 5527. Washington, DC: Human Development and Public Services Team, Development Research Group, World Bank. http://dx.doi.org/10.1596/1813-9450-5527

Wolf, Robert L. 1979. "The Use of Judicial Evaluation Methods in the Formulation of Educational Policy." *Educational Evaluation and Policy Analysis* 1(93): 19–28. http://dx.doi.org/10.3102/01623737001003019

Woolcock, Michael. 2009. "Toward a Plurality of Methods in Project Evaluation: A Contextualized Approach to Understanding Impact Trajectories and Efficacy." *Journal of Development Effectiveness* 1(1): 1–14. http://dx.doi.org/10.1080/19439340902727719

World Bank. 2015. "SABER: Systems Approach for Better Education Results. Strengthening Education Systems to Achieve Learning for All." Accessed June 8. http://saber.worldbank.org/index.cfm

Worthen, Blaine R., and R. Todd Rogers. 1980. “Pitfalls and Potential of Adversary Evaluation.” *Educational Leadership: Journal of the Association for Supervision and Curriculum Development* 37(7): 536–543. http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_198004_worthen.pdf

Yamauchi, Futoshi. 2014. “An Alternative Estimate of School-Based Management Impacts on Students’ Achievements: Evidence from the Philippines.” Policy Research Working Paper 6747. Washington, DC: East Asia and the Pacific Region Education Sector Unit, World Bank. http://dx.doi.org/10.1596/1813-9450-6747