International Development Working Paper **No. 2017-01 February 2017**

Doing Reform Differently: Combining Rigor and Practicality in Implementation and Evaluation of System Reforms

Luis Crouch
Joseph DeStefano
RTI International

RTI International
3040 East Cornwallis Road
P.O. Box 12194
Research Triangle Park, North Carolina 27709-2194
USA

International Development Working Paper

Doing Reform Differently: Combining Rigor and Practicality in Implementation and Evaluation of System Reforms

Luis Crouch
Joseph DeStefano

RTI International
3040 East Cornwallis Road
Post Office Box 12194
Research Triangle Park, NC 27709-2194
202-728-2058
lcrouch@rti.org

February 2017
No. 2017-01

Prior International Development Working Papers include:

- *No. 2013-01: Strengthening Local Councils in Uganda*
- *No. 2013-02: The Political Economy of Adopting Public Management Reforms: Patterns in Twenty Indonesian Districts*
- *No. 2013-03: Capacity Development for Local Organizations: Findings from the Kinerja Program in Indonesia*
- *No. 2014-01: Strategies for Improved Delivery of Basic Services: A Concise Framework and Selected Cases*
- *No. 2014-02: Does Better Governance Improve Service Delivery? Evidence and Lessons Learned from the Guinea Faisons Ensemble Project*
- *No. 2014-03: From Supply to Comply: Gauging the Effects of Social Accountability on Services, Governance, and Empowerment*
- *No. 2015-01: Social Accountability in Frontline Service Delivery: Citizen Empowerment and State Response in Four Indonesian Districts*
- *No. 2015-02: Systems Thinking and Institutional Performance: Retrospect and Prospect on USAID Policy and Practice*
- *No. 2015-03: Business Environment Constraints for Micro and Small Enterprises in El Salvador: Disparities between Male and Female Entrepreneurs*
- *No. 2015-04: Guide to Assessing Social Accountability Efforts Across Sectors*
- *No. 2016-01: Distance, Services, and the Decoupling of Citizen Perceptions of the State in Rural Africa*
- *No. 2016-02: RTI's Workforce Development Ecosystem Framework and Survey Tool: Findings from Caribbean Pilot Tests*
- *No. 2016-03: State Fragility, International Development Policy, and Global Responses*

TABLE OF CONTENTS

LIST OF EXHIBITS

ABSTRACT

This paper brings together two promising intellectual trends in development: Doing Development Differently (DDD), and whole-system reform. In addition, it provides a framework for evaluating system reforms, as rigorously as possible. Doing Development Differently proposes an approach to development in which many different things are tried, in smallish ways, involving stakeholders; in one sense it is a reaction against "blueprint" development. These approaches have much to recommend them. While it may seem counterintuitive that whole-system reform can benefit from an approach that tries many different things, we posit that this may be the best way to carry out reform. The proposed approach to system reform also borrows from the idea of Problem-Driven Iterative Adaptation (PDIA). An issue that arises is that evaluating policy reform—especially an approach to reform characterized by DDD or PDIA—is very difficult methodologically. The difficulty occurs because there is no counterfactual, and there is no one single intervention. Many attempts are made, and the package that emerges may be far from what was designed. The paper argues that, nevertheless, there are ways to interject more rigor into the evaluation of these reforms than at first might seem possible. The education sector is used as a case in point.

Key words: system reform, education reform, Doing Development Differently, political economy, reform evaluation

Doing Reform Differently:
Combining Rigor and Practicality in Implementation and Evaluation of System Reform[1]

1. INTRODUCTION

"Doing Development Differently" is a movement or initiative that tries to see development practice in a new way—that is, avoiding large-scale, top-down, "blueprint" planning. The initiative responds to the writings of scholars and practitioners (who might not all agree with each other) such as Andrews, Pritchett, and Woolcock (2012), Pritchett, Samji, and Hammer (2013), Faustino and Booth (2014), Greene (2014), and many others. The influence of others who may not be literally associated with the movement, such as Easterly (2006) can be discerned.

The Doing Development Differently initiative maintains a few key principles. Quoting and paraphrasing from the manifesto, DDD Manifesto Community (2017), "different" efforts:

- Solve local problems debated, defined, and refined by local people
- Are legitimized at all levels (political, managerial, and social)
- Are "locally owned" in reality (not just on paper)
- Work through local conveners who mobilize all those with a stake in progress
- Blend design and implementation through rapid cycles of planning, action, reflection, and revision
- Manage risks by making "small bets": pursuing activities with promise and dropping others
- Foster real results – real solutions to real problems that have real impact

Two important questions arise. First, the tenets outlined above relate to development practice in terms of specific projects. We are more interested in situations where specific solutions have been proven to work at the project level, but do not seem to go to scale. An implicit assumption here is that improving well-being may require working at scale, not staying forever at the level of thousands of localized solutions. The question then is: What if there is a stock of solutions, but nothing happens with them?

Second, what if the solutions imply upsetting powerful actors? Suppose that "doing development differently" processes identify some real solutions. In other words, the solutions have external validity, clearly work, have large impacts, are a clear improvement on the status quo, are well-evaluated, have gone through cycles of iteration, and so on. But now suppose that the solutions skipped some of the

[1] This paper adapts some concepts from the paper "A Practical Approach to In-Country Systems Research" written for the Research on Improving Systems of Education (RISE) Programme. This version turns more toward the Doing Development Differently movement rather than education systems reform. The original paper was presented at the first RISE conference in Washington, DC, June 18–19, 2015. RISE is financed by the British Department for International Development (DFID) and is guided by an Intellectual Leadership Team in which one of the authors participates. See RISE (2017) for a description.

principles. For instance, perhaps the solutions disturb a local political economic equilibrium, so it is impossible to get literally everyone to "own" the solution.

The education sector offers a good example of the not-quite-locally-owned dilemma. Imagine that the development process reveals that it is possible to produce books much more cheaply than what the local books oligopoly charges for them. Furthermore, these books are more effective than the books produced by the oligopoly. It is pretty clear that that oligopoly will have disincentives to "own" the solution. For the solution to go to scale, some kind of systemic reform or alternative agreement will be needed. Some local actors will need to be left behind, or, alternatively, sold on the solution.

The motivating question, then, is: Once there is clear evidence that reform elements have an impact, how does one assist with system reforms when powerful interest groups and bureaucratic inertia block the way forward? And, just as importantly, if a way is found, how does one evaluate whether such assistance is having impact? Randomized controlled trials (RCTs) on system reforms would not be an easy answer. How rigorously could one evaluate reforms where, by definition, no counterfactual is possible, and there is only one sample point (one country)? From a DDD perspective, how could one ensure rapid iterations, evaluations, and adaptations when trying to effect system reforms at scale? What does an at-scale "small bet" look like?

The paper is divided into four sections that build an approach to the questions posed above. In the first section, we present evidence of educational improvement solutions that have been proven to work in a variety of settings, and that have passed through processes of approximation and iteration. If there were no such evidence, there would be no raw material for reform. In the second section, we argue that those solutions can be generalized into a minimalist set of "system" lessons that can inform systemic change or reform. Third, we note the political-economic obstacles to such change. In the last section, we note how DDD principles can be used to both implement and evaluate complex reforms based on the experimentation that has already taken place and the generalizable minimalist or "core functions" system lessons.

2. ACHIEVING LARGE EFFECT SIZES

Is there evidence that certain interventions work? Yes, but with considerable variation among types of interventions. Experiments in some areas have produced apparent effect sizes on learning that, optimistically, ranged only to about 0.2 standard deviations (SDs). Among them are programs 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).

On the other hand, experiments that have focused extremely tightly and proficiently on certain pedagogical practices have had effect sizes ranging up to about 0.45 or even 0.50. Among these are

improved teaching methods that meet the children where they are, vastly improved textbooks based on rigorous research, use of the children's mother tongue, and all these in combination

There are systematic reviews of the impact of education interventions (Conn 2014; Glewwe, Hanushek, Humpage and Ravina 2010; Kremer, Brannen, and Glennester 2013; Krishnarante, White, and Carpenter 2013; McEwan 2015; Murnane and Gaminian 2014), and reviews of systematic reviews (e.g., Evans and Popova 2015). Rather than reprising the conclusions of that literature, and instead to provide a flavor of what it contains, we decided to search just 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. We recognized that some of the papers would represent reasonably rigorous but nonexperimental studies. The justification for this simple selection criterion was that the World 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 SDs. 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 0.1 level or better, the average effect size was 0.16 SDs. Muralidharan and Sundararaman (2013), doing research on contract teachers in India, found an effect size of around 0.16. 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 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 yields a median effect size of 0.17, with an interquartile range of 0.13 to 0.22. So, the optimistic or reasonably high expectation is around 0.22.

Our World Bank search identified a few papers on pedagogical interventions. 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.

We narrowed the non-Bank selection by looking at interventions on reading in the early grades, some of which have shown relatively high effect sizes.

- 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 (He, Linden, and McLeod 2009; Banerjee et al. 2012; 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 have followed 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. 2016).

Programs that were less tightly focused than those summarized above produced statistically significant impacts, but the effect sizes tended to be smaller. See, for example, an evaluation by Costa and Carnoy (2015) of the Brazilian program “Pact for Literacy 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 much better than those from the accountability interventions, with the medians almost double, and the bottom of the interquartile range being about the same as in the accountability interventions.

In summary, there have been interventions that were piloted in ways that allowed for cycles of iteration consistent with DDD. And those interventions have been fine-tuned iteratively to achieve the effect sizes summarized above. In addition, we contend that narrowing the focus of interventions—as per the DDD principle of making "small bets"—can, counterintuitively, contribute to meaningful system-level change. Tackling specific, key elements of the system that impact the quality of service provision can stimulate rapid cycles of learning and, in fact, can lead to "doing systems change differently." Again, we offer the example of improving the teaching and learning of reading in the early grades.

Therefore, we start with, and reason outward from, pedagogical subsystems that do seem to work. We then ask what specific accountability structures and other systems features might best leverage the desired pedagogical practices. The next section examines what those subsystems might be.

3. A SYSTEM FOCUS ON A CORE SET OF FUNCTIONS

Almost all education systems across the developing world have implemented reform efforts over the past three decades. Indisputably, during that time, access to schooling has expanded. But at the same time, most measures of outcomes have indicated that actual learning is all too elusive for many children who have gained entry to school (Pritchett, 2013). 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 can attest that the complexity of those efforts, combined with an excessively blueprint approach (in direct contradiction of DDD principles), 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? In short, comprehensive approaches such as SABER might not be DDD, and might not be effective.

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. We are making a distinction here 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, albeit in a very non-DDD manner—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. Ministries (or their decentralized units) will need to be able to manage the most basic day-to-day operations in good DDD fashion, rapidly iterating and learning from "small bets" on the types of supports that best aid teachers in adopting new instructional practices.

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

First, the ingredients for good teaching and learning of basic skills in primary school are well known (see Kim et al. 2017). 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. A simple view of learning is that *learning = learning time × rate of learning.* The "learning time" portion is a matter of clock or calendar time (and accountability for using it well). The "rate" portion is related to method: By definition, better methods, appropriately adapted to learners' needs, produce better rates of learning. This algorithm is what informs, for example, 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 simple math, precisely because they are not able to guarantee the basic requirements named above. Studies of allocated versus effective instructional time have shown 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 (Bruns and Luque 2014; Schuh Moore et al. 2012). Observations of classrooms across numerous countries also have 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 (Abadzi, 2007; Bruns, De Gregario, and Tau 2016; Schuh Moore et al. 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 translate them into coherent lessons, and partly because teachers are not trained explicitly in how to maximize the value of materials during teaching (Crouch, 2007; Schuh Moore et al. 2012). Additionally, instruction falls far short of what is needed because the 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. 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.[2] Most schools operate in isolation from the

[2] For a general discussion of problems with implementing education reforms at scale, see Elmore (1996).

system. They may receive some resources, such as curriculum and materials, but at the same time, they may not even be guaranteed that staff salaries will be paid regularly. Teachers may participate in workshops or other professional development activities, but those are usually divorced from classroom practice, delivered through a highly ineffective cascade, and devoid of ongoing support that would allow the teachers to apply what they learned. Suppose, for example, that a well-intentioned reform leads to research-based improvements in the structure of the reading curriculum for the initial grades of primary school. However, teachers receive only a smattering of poorly delivered training about the new curriculum, literacy materials are delivered late if at all, and the teachers have no one to turn to for advice on how to align their daily practice to the new program of study.[3] It is no wonder if such reforms fail to alter student outcomes.

We therefore do not see the challenge of "systems change" in terms of trying to catalogue all the policies, procedures, and institutional capacity needed (say, like SABER). Instead, the challenge is identifying the core things a system needs to do to ensure that teaching and learning improve in all schools. This requires stripping down the notion of an education system to its first principles, so that one can think small enough—like the DDD principles.

Realizing large-scale improvements in learning outcomes hinges on one key challenge: removing bureaucratic administrative structures afflicted by isomorphic mimicry[4] and replacing them with an explicit management imperative to achieve results and ensure equity. This shift involves moving away from a bureaucracy designed to handle routine administration, and toward a system of management relationships designed to carry out DDD-like rapid cycles of planning, action, reflection and revision.

In the early 2000s, the Annenberg Institute for School Reform set out to determine how school districts in the United States 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 the very basic components of what a school district should do: 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 struggling (Ucelli and Foley, 2004).

Likewise, we are reverting to the basic notion of how a system adds value to schools, namely by performing three core functions (discussed more in depth below) that can enable a DDD-like approach to management to occur:

- Setting expectations for the outcomes of education,

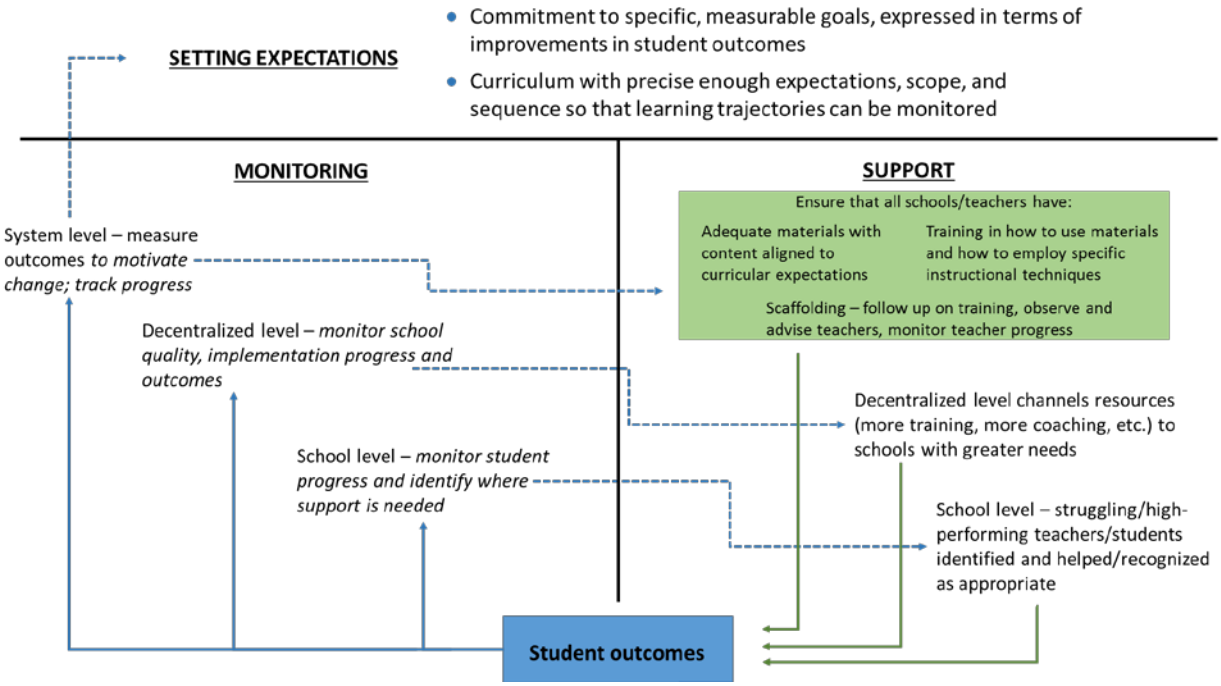
[3] As a sample case, many issues like these arose recently during implementation of the National Literacy Acceleration Program (NALAP) in Ghana. For a detailed account and evaluation, see RTI International (2011).

[4] Pritchett (2013) defines "isomorphic mimicry" as 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.

- Monitoring and holding schools accountable for meeting those expectations,
- Intervening to support the students and schools that are struggling to meet expectations, 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 whether the system can influence schools and students to achieve better learning outcomes. Each function is described below and the relationships among them are depicted in ***Exhibit 1.***

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



3.1 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. Any attempt to take a DDD approach to managing the delivery of education services by purposefully iterating and adapting the types of supports provided to teachers, schools, and students, also requires a clear understanding of what outcomes should be obtained, as well as the means to monitor and measure them, as described below.

3.2 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 and for iterative management.

The system also needs to monitor 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 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 or 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 to 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 for two types of monitoring. The first type is to monitor implementation and impact. The second is, in a decidedly DDD way, also to revisit and test initial assumptions along the way by making smaller attempts at change, evaluating those, and re-strategizing constantly.

Accountability for results is 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. The system needs to have the capacity to respond to monitoring information that indicates that basic inputs and supports are reaching schools—and to iteratively evolve "solutions" to those issues as needed.

System capacity to monitor school-level implementation and results is decidedly different from what typical management information systems (in education, EMIS) are structured to do. We depict decentralized monitoring as 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 in 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 iterative decision-making regarding strategies for providing the supports which schools need to improve.[5]

Developing leading rather than only lagging indicators of improved outcomes is another aspect of how monitoring at the decentralized level can improve (Foley et al., n.d.). 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 a 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 and local actors to come up with 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, adaptive help: more/different coaching, additional/different training, or supplementary materials or other inputs. 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 and in response to local conditions

[5] RTI has piloted the use of LQAS-based approaches to school monitoring in Ghana (RTI International, 2013), Zambia (Bostock and Rakusin, 2014), and Tanzania (RTI International, 2015).

and local input. These kinds of interventions, whether flowing as system responses to differentiated need or arising from local adaptation to contextual variations, represent the kind of DDD "small bets" that would enable rapid cycles of experimentation, evaluation, and adaptation in system functions, not just school-level solutions.

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 (for example, see Coyle 2009) 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. And that shift is at the heart of a deliberately adaptive and iterative approach to teaching and school management.

Children (especially young children) will learn at different rates, 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 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. And the system should ensure the monitoring that allows these local solutions to be evaluated and adapted as needed.

We conclude Section 3 by noting that from the successful experiences described in Section 2, one could distill a set of core functions. This core could be the focus of system improvement and should be less challenging to implement than a complete system reform. Focusing on a relatively minimalist set of three core functions also could allow a DDD sort of approach to implementing a reform: Focus on smaller things, and iterate, without everything impacting everything else.

4. PUTTING IN PLACE THE CORE SYSTEMS FUNCTIONS: POLITICAL-ECONOMIC ASPECTS

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.

Existing arrangements are not an accident.[6] 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. 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).[7]

In contrast, reforms needed now 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. The usual challenges of changing human behavior are compounded by the vested interests that 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 core functions we are advocating would have to confront exactly these kinds of obstacles.

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 that 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

[6] 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).

[7] 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 when systems are trying to put in place the core functions we write about here.

still in early stages of experimentation, but standardized approaches are being developed for sampling, defining indicators, and training school directors and district administrators to collect the necessary data (RTI International 2016). 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. While a DDD approach might argue that locally generated and locally owned solutions are the key, we contend that it is inherently unjust to leave it to every school/community to "discover" on its own what, for example, years of research tell us are effective ways to teach reading. The local input, adaptation, and ownership can come into play when schools and communities are determining how best to support and effectively manage the introduction of research-based approaches to instruction.

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 jobs as monitoring quality, not as making sure quality improves. Having these administrators and inspectors become 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 those roles, and even of their definition of what success means will need to change.

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 only need look 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 the findings 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 the bureaucracy, as opposed to accountability to the leaders themselves.[8] 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 three core functions could manifest itself.

Exhibit 2. Summary of likely resistance to improving the core system functions

Core 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

[8] 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 union dues (and honor and prestige and political influence that often come with holding high offices in unions).

Core functions	Political-economic obstacles likely to manifest themselves
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) face high organizational costs.

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 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. In this manner, significant in-country ownership for a different vision of education system operation can emerge.

This section has argued that investing in supporting systems-level change by providing targeted, opportunistic 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. We emphasize the term “opportunistic.” Being consistent with a DDD approach of gradual iterations based on making smaller attempts, evaluating those attempts, and resetting strategy (or “crawling the design space” in Andrews, Pritchett, and Woolcock’s [2012] terminology) requires some adaptation to these political-economic realities. The notion is to proceed in a manner that makes many different little sorties, evaluates resistance to reforms, and looks for ways around the push-back.

One alternative would be trying to use "projectized" assistance (in the language of Pritchett, Samji, and Hammer 2013) to fund the inputs necessary to ensure high-quality instruction on a national scale. This route would be extremely costly, and still would not guarantee a system capable of ensuring the three core functions described here. And yet another alternative—external funders using budget support and policy trigger points in nonproject loans and grants—typically does not generate enough technical dialogue. Nor does it take into account the political-economic reality of the changes needed, leaving us to argue for a DDD-like approach to large-scale change that strategically explores how to navigate the technical, conceptual, and political challenges described here.

5. A PROPOSED IMPLEMENTATION/EVALUATION APPROACH: ITERATIVE (DDD-LIKE) REFORM

This section deals with the issue of how to make an actual system reform project 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.

The basic idea here is that implementing a system reform ought to be as DDD-like as possible: being led by local actors; using local energies to find solutions, particularly to the political-economy problem; and, above all, constantly iterating in implementation when things do not go as planned, as they will not. How does one evaluate, in a comprehensive mode, something so inherently messy?

A key assumption we are making is that even system reform can be projectized. In other words, 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. And our attention to a limited, core set of system functions is in part predicated on trying to make system reform more project-like. But taking it a step further, we believe this "projectization" of reform should obey DDD-like principles in both its implementation and its evaluation. We are not proponents of traditional blueprint sorts of projectization when it comes to system reform.

An approach to implementation that is consistent with DDD would use a mix of:

- The experiential learning approach proposed by Pritchett, Samji, and Hammer (2013), where iterative implementation, and design "as one goes along" are emphasized
- Insights on causality from the Bradford Hill conditions (Hill 1965), needed in nonexperimental situations
- A juried approach to using third-party peers and experts to read the evidence produced according to the points immediately above.

5.1 Crawling the design and implementation spaces

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. These are consistent with DDD, which is not surprising, given the links between

Pritchett and colleagues and the DDD effort. ***Exhibit 3*** summarizes the principles and then shows how we would 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 limit ourselves to the word "implementation" just for emphasis.[9]

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 core-functions reform is working on too narrow a goal, it may create an external validity problem, as well as short-changing broader, and more important, goals. What is the right breadth of goal so as to make sure one can generalize, and be significant? If a reform had a large impact on children's reading, we would have reason to suspect that the reform, or the style of reform, needed to sustain those impacts, was capable of broader impacts, even if it had not shown them yet. But if the index metric were narrowed to, say, fluency or decoding ability as opposed to "reading," then generalizability and external validity of results could become problematic. Getting the goals right would take some negotiating and careful thinking.
Reverse-engineer to the instruments	Ideally, reforms should employ a *minimum* set of instruments (or core functions) that prior experience and literature lead us to believe are plausible. This is for practical reasons. If we want to implement iteratively, having too many subsystems as the object of reform will make it difficult to figure out who (or what) is pushing back, and why—and thus how to iterate around the obstacles.
Design a project	In this case, we mean *design a reform as a project.* For a system reform rather than a pilot project, this will require paying attention to the flow chart that links all the individual instruments (core functions), 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	In a system 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 (2013) 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.[10] To make the project evaluable in the judicial or "parliamenary-hearing" sense that we propose (see below)—and to have a hope of some external validity—the initial design or plan, the crawling, and the redesigns or re-strategizing that result all have to be carefully documented. 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

[9] We are adapting a framework similar to what others have proposed in the past. Rondinelli (1983) and Brinkerhoff (1990) presented similar pointers long ago. 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. These papers emphasize planning iteration rather than using blueprint planning, decentralizing decisions, not suppressing negative findings, and creating more flexibility in development agency programs. More recent authors not only have lists that are somewhat self-consciously similar, but also even offer highly simplified project planning tools that are quite attractive (Faustino and Booth 2014). The principles espoused in Doing Development Differently strike us also as another way of expressing the same underlying concepts.

[10] 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
	evaluation 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 narrowly 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.[11] 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.
Specify the design space and select alternative designs	This action may not be as important for a system reform implementation project as for a technical project. In a system reform project, we would propose to start with a fairly well-specified system of core functions that, based on lots and previous technical projects, could do the trick. The crawling, then, is around the implementation space: Who are the allies, who is likely the opposition, how will one market the ideas, how one 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 is 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 system 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.

5.2 Causality: Evaluating complex reforms

It is extremely difficult to evaluate a system reform using experimental or quasi-experimental means. Evaluating a reform implemented using DDD principles would be even harder, because there would be no single intervention: Every iteration in some sense would represent a different intervention. Even unsuccessful iterations would be building toward (possible) final success, because they would inform later iterations. In an experimental evaluation, failed attempts tend not to be visible as part of the "treatment" that is provided. But in an implementation based on DDD principles, or on crawling the design space, the learning from failed iterations would be very much part of the "treatment."

What forms of evaluation might be relatively rigorous, and yet not sacrifice the principles of DDD design? We propose borrowing from epidemiology, a field of knowledge where there is great interest in causality, but where controlled experiments are not possible, and where the accumulation of evidence is itself often iterative. Using the Bradford Hill conditions is a good place to start.

[11] 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.

Bradford Hill 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.

For various reasons, researchers *cannot* carry out high-*N*, high-replication, counterfactual studies of system reforms in development settings, even if they *wanted* to, as many researchers have documented (e.g., Woolcock 2009). 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 establish causality when counterfactual experimentation is not possible or desirable. These conditions have become known as the "Bradford Hill conditions." Hill himself considered them not a substitute for counterfactuals with randomization, but the most reasonable alternative when experimentation is not possible. Others have written about how to apply the conditions based on how they compare to the ideal of counterfactuals (Höfler 2005). ***Exhibit 4*** lists the conditions, and discusses which might be feasible in a one-off, whole-system reform experiment (as opposed to an evaluation of a collection of experiments).

Exhibit 4. Possible use of the Bradford Hill conditions in a system 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 impacts (or lack thereof, to make the opposite case). If human exposure to something (a chemical) is associated with a 200-fold increase in something else (cancer), there is reason to suspect something other than correlation. (This argument assumes the other conditions are analyzed.) In an *N* = 1 experiment, the strength of impact can still be noted. In a reform attempt, one would hopefully be looking for serious impact—maybe not 200-fold, but certainly 1 SD and higher. This should be stated in the reformist project's definition documents and its theory of change.
Consistency	Has the association been observed in many contexts? This is the 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 approach 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.
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). This condition is not usable in a systemic reform causal evaluation, almost by definition.
Temporality	Temporal factors are evident and relevant for a system 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, 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 in attributing causality. Again, in which aspects of the reforms, and for which outcomes, one should expect this, should be explicitly laid out, ex ante, in the reform 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.

Condition	Possible applicability to the project
Biological (pedagogical in our case) plausibility and coherence[12]	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 an education reform, the changes would have to be plausible in various ways. We know, for instance, that, for humans to learn 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 (e.g., more frequent coaching of teachers that emphasizes the skills to instill automaticity in their students, and to start learners from where they are) would follow a plausible pedagogical path. For other aspects of the reform, one might be less able to sketch out the plausible pedagogical path.
Experimentation[13]	This type of investigation means not simply taking advantage of natural variation, but instead carrying out purposeful and systematic experimentation, albeit short of using counterfactuals. It does not seem possible in a system 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).

5.3 Putting together the evidence

Non-experimental situations, as noted, do not allow for the crispness of statistical tests. Also as noted, it would be very difficult to evaluate a reform implemented according to DDD principles, because of the many iterations that would make "treatment" hard to specify in binary fashion even ex-post. The language of evaluation then would become vague. And it would (or should) use processes similar to the accumulation of circumstantial evidence in establishing causality in judicial proceedings and jury trials.[14] 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 and generalizability.

Forming panels would be 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). While Woolcock's paper emphasized the problem of coming to conclusions too early, the paper also argued for a complete approach that includes as many forms of evidence as possible.

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

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

[14] 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 the case of systems reform, upon presentation of all the possible evidence, a "jury" must be able to exclude every other reasonable explanation for impact.

A summary of steps proposed by one of many authors (but a widely cited one, in this case: Mayne 1999, p. 16) 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."

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 data on impact, the documentation of the crawling of the design space, interviews with key actors, etc. The proposed approach is not an excuse to lower the bar. This means that, for any aspect whatsoever where quantitative evidence, experimental or quasi-experimental, can be adduced, it should be, and one 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 engaged through a third party.

The approach has been 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 available from the reform process itself or prior in-country or worldwide experience, 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.

(3) Empanel a group of experienced researchers and peers, including local experts as much as possible, *from the beginning*.

(4) Design 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 non-counterfactuals can provide some sense of what is needed. In addition, of course, experimental and quasi-experimental evidence should be admissible.

(5) Assuming a long-term reform process, 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 for years 1 and 2, likely to be plausible) and which are less so, and what aspects seem to permit generalizability.

(6) Interview key actors, rather than just reviewing the documentary evidence.

(7) 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 by involving stakeholders in the deliberation itself.

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 connections among evaluators and 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. A better analogy might be that of parliamentary or 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 parliamentary or congressional approach than a judicial one.

We think the idea is at least worth exploring even if some of the initial applications of the concept 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).

6. CONCLUSION

In the introduction to this paper, we posed four questions, relevant to a DDD-like approach to sectoral reform, using the education sector as a case in point.

Has there been sufficient DDD-like experimentation in any given region or sector to solve specific problems, through iterative processes? If not, the case for systemic reform would be weaker because the "content" of the reform would be unclear.

(1) Is there a limited set of system changes that, because limited, could be subjected to iterative implementation based on DDD principles? If not, then it would be difficult to iterate, given the complexity of trying to reform a whole system.

(2) What are the political-economic constraints that particularly require a DDD-like approach? These must be considered as they are often the most important determinant of why one has to reassess, retool, and iterate.

(3) Can something as inherently messy as system reform, especially system reform implemented using a DDD approach rather than a blueprint approach, be subjected to rigorous implementation and evaluation? While not all things that are important can be measured, it would seem professionally irresponsible not to be as rigorous as feasible, just because the issue is complex.

Our combined answers to these questions, we believe, offer some insight into how one could approach reform applying DDD principles in at least one important sector. The lessons probably generalize to other sectors—if they don't, however, education is important enough, as a sector, to merit the needed attention. The approach would require the following actions: (1) carefully selecting the aspects of the system that need to change, (2) working with the local actors who are intervening strategically to support the institutional reforms needed to improve system capacity related to those core functions, (3) explicitly mapping the anticipated causal chain that ties those system capacities to improved teaching and learning on a national scale, (4) anticipating political-economic issues and possible push-back, (5) iterating gradually and in a manner where the iterations are endogenous to how the reform is going, (6) investing in documenting and evaluating how the pieces of that puzzle do or do not fall into place, and (7) measuring learning outcomes all along the way. We feel that this implementation design would make it possible to implement system reforms more effectively, in a manner that would borrow from DDD principles.

BIBLIOGRAPHY

Abadzi, Helen. 2007. "Absenteeism and Beyond: Instructional Time Loss and Consequences." Policy Research Working Paper No. 4376. Washington, DC: Independent Evaluation Group, World Bank. https://openknowledge.worldbank.org/handle/10986/7569

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. https://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. https://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. https://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. https://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. https://doi.org/10.1596/978-0-8213-8679-8

Bruns, Barbara, Soledad De Gregario, and Sandy Tau. 2016. "Measures of Effective Teaching in Developing Countries." Research on Improving Systems of Education (RISE) Working Paper 16/009. http://www.riseprogramme.org/content/rise-working-paper-16009-measures-effective-teaching-developing-countries

Bruns, Barbara, and Javier Luque. 2014. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington, DC: World Bank. https://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. https://doi.org/10.1596/1813-9450-5809

Conn, Katherine. 2014. "Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Rigorous Impact Evaluations." Dissertation, Columbia University. https://doi.org/10.7916/D898854G

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.* https://doi.org/10.3102/0162373715571437

Coyle, Daniel. 2009. *The Talent Code. Greatness Isn't Born. It's Grown. Here's How.* New York: Bantam Dell.

Crouch, Luis. 2007. *Toward High-Quality Education in Peru: Standards, Accountability, and Capacity Building.* World Bank Country Study. Washington, DC: International Bank for Reconstruction and Development / World Bank. https://openknowledge.worldbank.org/handle/10986/674

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. https://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. https://doi.org/10.1596/1813-9450-5629

The DDD Manifesto Community. 2017. "Doing Development Differently: The Manifesto." Accessed January 18, 2017. http://doingdevelopmentdifferently.com/the-ddd-manifesto/

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://globalreadingnetwork.net/eddata/information-education-policy-planning-and-management-summary-data-capacity-assessments

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

Easterly, William. 2006. *The White Man's Burden: Why The West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good.* Oxford, UK: Oxford University Press.

Elmore, Richard F. 1996. "Getting to Scale with Good Educational Practice." *Harvard Educational Review* 66(1): 1–27. https://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

Glewwe, Paul, Eric Hanushek, Sarah Humpage, and Renato Ravina. 2010. "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010." NBER Working Paper No. 17554. http://www.nber.org/papers/w17554

Goyal, Sangeeta, and Priyanka Pandey. 2013. "Contract Teachers in India." *Education Economics* 21(5): 464–484. https://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. https://doi.org/10.2307/1912791

Greene, D. 2014. "Building Women's Leadership in the Most Difficult Places (Pakistan): Case Study for Your Comments," *From Poverty to Power* (blog), June 2, 2014, http://oxfamblogs.org/fp2p/building-womens-leadership-in-pakistan-case-study-for-your-comments/

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]. https://doi.org/10.1177/0141076814562718

Höfler, Michael. 2005. "The Bradford Hill Considerations on Causality: a Counterfactual Perspective." *Emerging Themes in Epidemiology* 2(1): 11. https://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, Margaret M. Dubeck, Katherine E. Halliday, Hellen N. Inyega, Sharon Wolf, Stephanie Simmons Zuilkowski, and Simon J. Brooker. 2016. "Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial." *Journal of Research on Educational Effectivess,* 1–33. https://doi.org/10.1080/19345747.2016.1221487

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. https://doi.org/10.1596/1813-9450-6794

Kim, Young-Suk Grace, Helen N. Boyle, Stephanie Simmons Zuilkowski, and Pooja Nakamura. 2017. *Landscape Report on Early Grade Literacy.* Washington, DC: USAID. https://globalreadingnetwork.net/publications-and-research/landscape-report-early-grade-literacy-skills

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340(6130): 297–300.

Krishnaratne, Shari, Howard White, and Ella Carpenter. 2013. "Quality Education for All Children? What Works in Education in Developing Countries." 3ie Working Paper 20. New Delhi: International Initiative for Impact Evaluation. http://www.3ieimpact.org/en/publications/working-papers/working-paper-20/

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. https://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

McEwan, Patrick J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85(3): 353–394. https://doi.org/10.3102/0034654314553127

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. https://doi.org/10.3386/w19440

Murnane, Richard and Alejandro Ganimian. 2014. "Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations." NBER Working Paper No. 20284, National Bureau of Economic Research, Cambridge, MA.

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://globalreadingnetwork.net/eddata/going-scale-early-grade-reading-program-egypt-2008-2012

Orcutt, Guy H. 1952. "Actions, Consequences, and Causal Relations." *Review of Economics and Statistics* 34: 305–313. https://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." *Forum for International Research in Education* 2(1): 3–18. http://preserve.lehigh.edu/fire/vol2/iss1/2

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. https://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. https://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

Research on Improving Systems of Education (RISE). 2017. "What Is RISE?" Accessed January 18, 2017. http://www.riseprogramme.org/content/what-rise

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. http://pdf.usaid.gov/pdf_docs/PDACU018.pdf

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. http://pdf.usaid.gov/pdf_docs/PA00M2WK.pdf

RTI International. 2015. *Lot Quality Assurance Sampling (LQAS) Pilot in Tanzania: Final Report.* Prepared for USAID under the EdData II project, Measurement and Research Support to Education Strategy Goal 1, Task Order No. AID-OOA-12-BC-00003 (RTI Task 20, Activity 5). Research Triangle Park, NC: RTI. http://pdf.usaid.gov/pdf_docs/PA00M8DK.pdf

RTI International. 2016. *Toolkit for the Local Education Management Approach (LEMA).* Prepared for USAID under the Education Data for Decision Making (EdData II) project, Task Order No. No. AID-OAA-12-BC-00003 (RTI Task 20, Activity 5, Project No. 0209354.020). Washington, DC: United States Agency for International Development. https://globalreadingnetwork.net/sites/default/files/eddata/LEMA%20Manual_23Nov2016_SubmittedtoUSAID_Final.pdf

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. https://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. https://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. https://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. https://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. https://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. https://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, 2016. 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. https://doi.org/10.1596/1813-9450-6747