Draft Prepared for the Handbook of Field Experiments Conference (10-11 April 2015)

PRELIMINARY & INCOMPLETE VERSION FOR COMMENTS ONLY – PLEASE DO NOT CITE OR CIRCULATE

Field Experiments in Education in Developing Countries

Karthik Muralidharan[1]

1. Introduction

Perhaps no field in development economics in the past decade has benefited as much from the use of experimental methods as the economics of education. The rapid growth in high-quality studies on education in developing countries (many of which use randomized experiments) is perhaps best highlighted by noting that there have been *several* systematic reviews of this evidence aiming to synthesize findings for research and policy in *just the past two years*. These include Muralidharan 2013 (focused on India), Glewwe et al. 2013 (focused on school inputs), Kremer et al. 2013, Krishnaratne et al. 2013, Conn 2014 (focused on sub-Saharan Africa), McEwan 2014, Murnane and Ganimian (2014), Evans and Popova (2015), and Glewwe and Muralidharan (2015).[2] While these are not restricted to experimental studies, they typically provide greater weight to evidence from randomized controlled trials (RCT's).

Given the large number of recent reviews of the evidence on the impacts of various policy interventions to improve education in developing countries, this chapter will *not* seek to review the literature based on experiments conducted in education in developing countries in recent years. Rather, it aims to highlight key theoretical and design considerations in conducting field experiments in education in developing countries, and to illustrate design choices and trade-offs. In particular, given the well documented strengths of experimental methods, this chapter will pay more attention to their limitations and discuss how many of these weaknesses can be mitigated by better design of experimental studies and research programs. The goal of the chapter is to serve as a ready reference for students, researchers, and practitioners to guide the design and implementation of high-quality field experiments in education in a way that maximizes what we learn from them.

The chapter is organized as follows. Section 2 provides a conceptual overview of field experiments in education. Section 3 highlights some key findings from field experiments in education in developing countries in the past decade. Section 4 presents an extensive discussion on the limitations of field experiments for policy-relevant inference and discusses way in which these can be addressed by better design. Section 5 discusses considerations in the design and analysis of field experiments. Section 6 discusses approaches for future research to maximizing the extent of learning from future experimental work in education. Section 7 concludes and discusses areas for future research.

[1] UC San Diego, J-PAL, NBER, and BREAD. E-mail: kamurali@ucsd.edu

[2] There is also a systematic Campbell review in progress on the same theme (see Snilstveit et al. 2014)

2. Field Experiments in Education - a short overview

2.1. Background

Education and human capital are widely considered to be essential inputs for both aggregate growth and development (Lucas 1990; Barro 1991; Mankiw, Romer, and Weil 1992), as well as for enhancing the capabilities and freedoms of individuals, and thereby enabling them to contribute to and participate in the process of economic development (Sen 1993). Thus, improving education outcomes in developing countries has been an important policy priority for both national governments as well as for the international development and donor community. The importance of education in the development agenda is perhaps best reflected by the fact that two of the eight United Nations Millennium Development Goals (MDG's) pertained to education (achieving universal primary education, and achieving gender parity at all levels of education - both by 2015).

Further, education (especially school education) in most countries is typically provided publicly by the government and financed by taxpayer contributions. It is beyond the scope of this chapter to deeply analyze *why* this is the case, but there are at least three broad sets of reasons for the preponderance of publicly financed and provided education. First, there is plenty of evidence to suggest that a variety of supply and demand side constraints may prevent optimal education investments by households and optimal provision by markets, and that outcomes can be improved by a social planner. Second, it is widely believed that education generates positive spillovers to society beyond the returns that accrue to individuals, which would also suggest an active role for governments in education financing and production.[3] Finally, an important non-economic reason for publicly-provided education may be states' desire to control curriculum and the content of education, which affect the formation of preferences and civic identity.

Thus, financing and producing education is an important policy priority for most countries, and public spending on education is typically one of the three largest components of government budgets (the other two being defense and healthcare). However, while this is true for most countries, developing countries face especially acute challenges in achieving universal quality education. They have lower levels of school enrollment and completion, much poorer learning outcomes, and also have fewer public resources to spend on education. Thus, spending

[3] Models with complementarity in worker human capital in production (such as Kremer 1993) predict spillovers. Lucas (1988) argues that human capital spillovers may be large enough to explain most of the long-run differences in per-capita income between high and low-income countries. Moretti (2004) provides evidence of such spillovers in the context of US manufacturing workers and plants. One direct channel of spillovers for which there is evidence in a developing country context is that education promotes technology adoption, and that non-adopters (who may be less educated) learn from adopters, which is a positive spillover from education that would not have been accounted for in the decision-making of individually optimizing agents (Foster and Rosenzweig 1995).

scarce public funds effectively is especially important in developing countries, where the opportunity cost of poor spending is higher.[4] As a result, a major area of focus for research on education in developing countries has been to understand the effectiveness (both absolute and relative) of various policy options to improve education outcomes.

2.2 The main research questions

The majority of experimental research on education in developing countries in the past decade has focused on one (or both) of two main policy questions. First - how should we increase school enrollment and attendance, and second - how should we improve learning outcomes?[5] The two are closely related because increased enrollment and attendance are likely to be necessary pre-conditions for improving learning outcomes. Nevertheless, it is useful to think about the two problems distinctly because the school attendance decision is typically made by parents, whereas the extent to which increased school participation translates into improved learning outcomes is more likely to be affected by school-level factors.

On school participation and attendance, a simple model of optimizing households in the tradition of Becker (1962) and Ben Porath (1967) yields the result that households will only invest in an additional year of education for their child if the present discounted value of the expected increase in benefits exceeds the costs of doing so. Thus, policies that seek to improve school participation typically aim to increase the immediate benefits to households of sending their children to school or to reduce the costs of doing so. The magnitude of the impact of these policies will in turn depend on the distribution of the household-child specific unobservables that determine whether a given child enrolls in or attends school, and the extent to which the policy helps make it more attractive to do so.

On quality of learning, a standard education production function with certain additional assumptions (see Todd and Wolpin (2003) for a detailed exposition of these assumptions), allows the lagged test score to be treated as a sufficient statistic for representing prior inputs into learning, and for the use of a value-added model to study the impact of changing

[4] In principle, governments in developing countries should be able to borrow to undertake any investment where the social rate of return is greater than the cost of borrowing. In practice, financial markets typically constrain the extent of government borrowing which places a hard budget constraint on public expenditure.

[5] Note that an alternate way of framing these questions is to ask: "What are the determinants of school participation and learning outcomes?" The difference in emphasis is that the latter approach is more that of a "scientist" trying to "understand" the world, whereas the approach in the text is more than of an "engineer" trying to "improve" outcomes and "solve" problems (see Mankiw 2006 for a discussion of a similar distinction in approaches to macroeconomics). The choice of emphasis is deliberate, because most of the experimental research in education in recent years has been motivated by policy questions of how best to improve education outcomes in specific settings. I return to the distinction between these approaches and ways of better bridging the divide in the section on external validity. Further, as I discuss later, in the absence of measuring changes in other inputs that may change in response to an experimental change in one set of inputs, experimental estimates typically yield "policy" effects and not the more scientific primitive of a "production function" effect (see section 2.3).

contemporaneous inputs into education on test scores. Specifically, the typical value-added model takes the form:

$$T_{i,t} = \gamma T_{i,t-1} + \boldsymbol{\beta} \boldsymbol{X}_{i,t} + \varepsilon_{i,t} \quad (1)$$

where $T_{i,t}$ represents test scores of child *i* at time *t*, $T_{i,t-1}$ represents the lagged test score, and $\boldsymbol{X}_{i,t}$ represents a full vector of contemporaneous home ($\boldsymbol{H}_{i,t}$) and school ($\boldsymbol{S}_{i,t}$) inputs. While the production function above is linear in $\boldsymbol{X}_{i,t}$ and is typically estimated this way, the specification does not have to be as restrictive, because $\boldsymbol{X}_{i,t}$ can include non-linear terms in individual inputs, and also include interaction terms between specific sets of inputs.

Given budget constraints in public education, and the almost unlimited set of ideas for inputs and interventions that may improve education outcomes, an optimal policy approach to allocating scarce resources across the set of potential inputs would be to estimate the marginal return to providing a specific input and to compare it with the marginal cost of doing so (since these inputs are typically provided publicly) and to prioritize investments in diminishing order of the estimated return per dollar spent. Since cost data is relatively easier to obtain,[6] the main practical challenge is one of estimating the marginal returns to different inputs. The economics of education literature has correspondingly devoted a lot of attention to doing this and produced hundreds of papers across several developing countries trying to estimate these returns for various inputs (see Glewwe and Hanushek 2013 for a review).

2.3 The value of experiments in education and reasons for their growth

The main challenge for non-experimental studies (that use observational data) is the concern that variation in the specific input being studied ($X_{i,t}$) is correlated with the unobserved error term ($\varepsilon_{i,t}$), yielding biased estimates of β. In practice, this is quite likely to be true. For instance, communities and parents that care more about education are likely to be able to successfully lobby for more school inputs, and are also likely to provide unmeasured inputs into their children's education, which would lead to an upward bias on β estimated in cross-sectional data. In other cases, governments may target inputs to disadvantaged areas to improve equity in which case areas with increases in $X_{i,t}$ may be negatively correlated with $\varepsilon_{i,t}$, yielding downward biased estimates of β.

Thus, the value of experimental evaluations in this setting is quite clear since random assignment of the input (intervention) of interest solves this identification problem by ensuring that variation in $X_{i,t}$ is orthogonal to variation in $\varepsilon_{i,t}$, thereby yielding unbiased estimates of β

[6] In practice, even obtaining cost estimates of specific interventions is non-trivial (especially when it involves aggregating expenditure across multiple levels of government), but in principle they can be reconstructed from government budget documents.

(with some caveats as noted in section 4).[7] The importance of accounting for omitted variable bias in the evaluation of education interventions is starkly illustrated by Glewwe et al. (2004) who compare retrospective and prospective studies of the impact of classroom flipcharts on learning outcomes. They find using observational data that flipcharts in classrooms appear to raise student test scores by 0.2σ. However, when they conduct a randomized controlled trial of flipcharts in classrooms, they find no impact on test scores at all, suggesting that the non-experimental estimates were significantly biased upwards (even after controlling for other observable factors). These results underscore the value of field experiments for program evaluation in developing countries and Glewwe et al. (2004) can be considered analogous to LaLonde (1986) in the US program evaluation literature, which showed that non-experimental methods were not able to replicate the estimates from experimental evaluations of the impact of job training programs.

While field experiments have improved causal inference in most topics in applied micro-economics, they have been particularly prevalent in the economics of education (especially in developing countries) in recent years. There are several reasons for this. First, interventions in education are typically "modular" and therefore feasible to randomize at the student, classroom, or school level. Second, the outcome variables are quite well defined and there is a fair bit of agreement on the key outcomes that programs should aim to improve (enrollment, attendance, and test scores[8]). Third, the large number of non-profits and Foundations that work on education has made it feasible for researchers to find implementation partners who can design and deploy the interventions being studied. Fourth, since non-government implementation partners typically cannot (and are not expected to) work "everywhere", it is politically and practically feasible for them to use a lottery to determine where their programs will first be rolled out since this ensure fairness in program access in addition to enabling experimental evaluations of impact. Finally, evidence from nationwide surveys like ASER in India and Uwezo in East Africa showing that the large increases in spending on education have not led to improvements in learning outcomes has increased the demand from policy makers and funders for evidence on the impact of the programs they are funding, and for cost-effective ways of improving learning.

This confluence of factors has led to several high-quality experimental studies in education in developing countries that have both contributed evidence on the effectiveness of specific programs and also promoted a deeper understanding of the barriers to improving education outcomes in these settings. These include studies on interventions to improve parent and

[7] See companion chapters in this volume for more details (Banerjee and Chassang 2015; Glennerster 2015).
[8] While test scores are not the ultimate outcomes that a social planner cares about, evidence using long-term longitudinal data find that interventions that raise test scores in school (such as having a better teacher in primary and middle school) also lead to better long-term labor market outcomes (Chetty et al. 2011).

student demand for education, interventions to improve the inputs and resources to schools and children, interventions to improve classroom organization and pedagogy, and those aiming to improve school governance and teacher effort (with some interventions combining features across this broad classification). When put together (with further caveats), this body of research also enables comparison of marginal costs and benefits across different kinds of education spending and can guide policy priorities over allocation of limited resources. A non-exhaustive summary of the main results from this body evidence is presented in the next section to illustrate both the breadth of the recent research in this area and to summarize important insights that have been gained.[9]

3 Selected summary of field experiments in education in developing countries

3.1 Demand side interventions

On the demand side, perhaps the most widely studied intervention using RCT's has been the idea of "conditional cash transfers (CCT)" (with eligibility often targeted to poorer households) whereby households receive a regular cash supplement (typically monthly) as long as their children are enrolled in school and maintain a minimum attendance rate. While CCT programs aim to provide income support to the poor more generally (and not just increase demand for education), they have been found to have significant positive impacts on school enrollment and attendance across most settings where they have been evaluated using an RCT (see Fiszbein and Schady 2009 for a summary of several studies).

RCT's have also been used to study variants on the idea of cash transfers to see the extent to which modifying the design of cash transfer programs has an impact on outcomes. Prominent examples include: (a) Baird et al. (2011) who study the importance of conditioning the cash transfers on school enrollment by comparing a standard CCT to an unconditional cash transfers (UCT) in Malawi and find that conditioning the cash transfers do improve school enrollment relative to UCT's, but that the UCT's do better at protecting vulnerable girls by providing them with income even if they drop out of school, (b) Benhassine et al. (2014) who find that labeling a UCT as being for education (in Morocco) achieved significant gains in school participation, and that adding conditionality did not yield any additional gains in schooling (though it added additional costs of monitoring and enforcing the conditionality), and (c) Bertrand et al (2011) who find that postponing part of the monthly transfer in a standard

[9] As mentioned earlier, this section does not seek to conduct a comprehensive review of every field experiment in education in developing countries or to present a "meta analysis" of the result (which are better done by other review papers). Rather, the aim is to provide a flavor of the ranges of topics studied and to summarize some of the more influential experimental studies (note that we use the terms "experiment", "randomized controlled trial", and "RCT" interchangeably).

CCT to the time when school reenrollment has to take place (which is when fees need to be paid) significantly raised school enrollment rates relative to a standard CCT in Colombia.

In addition to studies evaluating the impact of CCT's on school participation, the randomized roll-out of CCT programs across individuals and communities (most notably PROGRESSA-Opportunidades in Mexico) has also enabled well-identified studies on important determinants of education participation including peer effects (Bobonis and Finan 2009, Lalive and Cattaneo 2009), consumption smoothing across households within communities (Angelucci and De Giorgi 2009), and the role of income controlled by women/mothers on children's consumption and education (Bobonis 2009). Finally, Todd and Wolpin (2006) and Attanasio et al. (2011) combine a structural model with PROGRESSA's experimental variation to generate predictions on schooling impact of the program under different values and design of the CCT program. Overall, CCT's have been one of the most highly researched and deployed policy options to improve demand for schooling in developing countries, and have been a poster child for the value of carefully randomized program rollouts in generating high-quality evidence on both program impact as well as on deeper determinants of household schooling investments.

A second prominent category of demand side interventions studied experimentally relate to the provision of better information about education to students and parents. Since education decisions are taken on the basis of *perceived* as opposed to actual returns, households may make sub-optimal decisions on education investments if they misperceive these returns. Jensen (2010) shows using household survey data in the Dominican Republic that the perceived returns to high-school are much lower than the actual returns, and shows experimentally that simply providing students in randomly selected schools better information on the higher measured returns to secondary schooling led to a significant increase in the years of school completed. In a variant of this experiment, Jensen (2012) shows that providing randomly selected villages in northern India with information on the job opportunities available to educated young women and helping them access these opportunities (without in any way changing the qualifications for being hired) led to a significant increase in female education, and to delays in marriage and fertility.[10]

However, an interesting contrast is seen in results from a similar experiment in China. Loyalka et al. (2013) conduct an experimental evaluation of the impact of providing information

[10] The information in this case was provided by providing job recruiting services in randomly selected villages and demonstrating that educated girls from the village could find jobs in call centers. The recruiting firms would not usually go to villages on their own since the expected number of successful candidates in a village would be too low to cover the fixed cost of visiting the village. One concern with the approach in Jensen (2010) is that the information provided on Mincerian returns to education may be incorrect on average (due to omitted variable bias) and also not be correct for the marginal student. The approach in Jensen (2012) is less susceptible to this concern because the standards for hiring candidates did not change (and hence no potentially misleading information was provided) but new information was provided by the recruiting firms.

on returns to education, and providing education and career counseling services (in separate non-overlapping treatments) to junior high school students in China. They find that the information treatment had no impact on high-school participation, and also find that the career counseling treatment actually increased school dropouts and reduced test scores. The authors attribute this surprising negative result to the fact the wages of unskilled workers were rapidly rising in China in this period, and the possibility that the counseling services may have made academically weaker students decide that the academic requirements of higher education were too onerous and that it may make more sense for them to drop out and join the labor force.[11] The difference in the results of similar interventions across country contexts highlights the importance of caution in comparing results across contexts (a point I will discuss extensively in the section on external validity).

RCT's have also been used to study the impact of providing information on school quality in competitive education markets with multiple providers (both public and private). Andrabi et al. (2014) use a large-scale RCT across 112 education markets (villages) in Pakistan to study the impact of providing parents with detailed student and school-level report-cards with information on test scores, and find that the intervention increased mean test scores by 0.11σ, reduced mean private schools by 17%, and also increased primary school enrollment by 4.5%. They also find that the mechanism for these results was an improvement in quality among the lower quality schools and a reduction in price among the higher-quality schools, which is consistent with the predictions of models of optimal endogenous pricing and quality choice by providers in settings of asymmetric information (and how these should change in response to provision of better market-level information on quality)[12]. This study highlights the capacity of RCT's to yield *market level* insights on how the provision of information can affect parental demand and increase competitive pressure on schools and change outcomes.

A final category of demand-side interventions with promising experimental evidence on positive impacts is the provision of student-level incentives for better academic performance. Two prominent studies include Kremer, Miguel, and Thornton (2009) and Blimpo (2014). The former conduct an RCT of a girls merit scholarship in Kenya and find significant positive effects on girls test scores and also find that teacher absence is reduced in treatment schools. They also find positive impacts on girls with low baseline test scores alleviating the concern that merit scholarships for high-performing students may de-motivate students at the lower end of the test score distribution. In a similar vein, Blimpo (2014) conducts an RCT of three different types of student incentives for better performance in Benin (one based on individual student incentives and two based on team student incentives), and finds that all three variants of the

[11] Note that this result is similar to that seen non-experimentally by Atkin (2014) who shows that Mexican high-school students were more likely to drop out from school during a period of increasing demand for unskilled labor.
[12] They also verify that parents' knowledge of school quality did improve as a result of the intervention.

student incentive program had a significant positive impact on the high-school exit exam test scores. The average impact across treatments and subjects was 0.3σ, with the difference across variants of the incentives not being significant.

Overall, the experimental evidence on the impact of demand-side interventions suggests that there are likely to be important demand side market failures that may lead to sub-optimal investments in education by parents and children. For instance, the large positive impacts of relatively small student prizes and incentives (which are especially small relative to the lifetime returns to completing schooling) suggest that students may misperceive the returns to education (or discount the future at a significantly higher rate than a social planner would). The results also suggest that well-designed demand side interventions are likely to be a promising avenue to explore for improving education outcomes in developing countries.

3.2 School Inputs

To be completed

3.3 Pedagogy

To be completed

3.4 Governance

To be completed

4 Limitations of experiments in education

As the discussion above suggests, field experiments in education in developing countries have yielded important insights about effectiveness (or lack thereof) of several possible education interventions. At the same time, there are also important limitations to what we can learn from field experiments in education. Note that most of the limitations discussed in this section apply to almost *all empirical research*, and should not therefore be seen as weaknesses of experimental methods in particular. But it is important to be clear about what problems experiments do and do not solve, and doing so can improve the quality of policy-relevant inferences made from individual studies, and may also help guide future research in ways that mitigate these challenges. In addition to discussing the limitations, I briefly describe ways of addressing each of the limitations in this section itself, and return to some of these themes later in the chapter.

4.1 Production Function versus Policy Parameters

The discussion in section 2.3 highlighted the value of experimentally varying $X_{i,t}$ in estimating the causal impact of $X_{i,t}$ on education outcomes. Note however, that even random

assignment of $X_{i,t}$ may not yield the production function parameter β outlined in Eq. (1). This is because the production function parameter β is a partial derivative ($\partial T_{i,t}/\partial X_{i,t}$) holding *other inputs constant*. In practice, other inputs at the school or household level may endogenously respond to exogenous changes in $X_{i,t}$, and the estimated parameter should therefore be more accurately interpreted as a policy parameter, which is a total derivative ($dT_{i,t}/dX_{i,t}$) that accounts for re-optimization by agents in response to an exogenous change in $X_{i,t}$.

The extent to which an experimental estimate reflects re-optimization will depend critically on the duration of the study. A clear illustration is provided by Das et al. (2013), who study a randomly-assigned school grant program in India over a two-year period and find significant positive effects on test scores at the end of the first year, but find no effect in the second year even though the grant was provided again in the second year, and was spent on very similar items in both years (books, school supplies, and classroom learning materials). They show that the most likely explanation for this result is that household spending on books and school supplies did not change across treatment and schools in the first year (when the school grant was unanticipated), but that households in treatment schools sharply cut back their spending on these categories in the second year (when the school grant was anticipated and could be accounted for in household decision making), and that this reduction offset around 80% of the per-student value of the grant.

The authors therefore argue that the "first year" effect of the program is more likely to represent the "production function" effect of providing the school grant (since other factors did not have time to adjust), whereas the "second year" effect is closer to the "policy parameter" (which reflects household re-optimization). The example highlights the importance of measuring as many intermediate inputs as possible to have a better idea about the mechanisms of program impact. However, in practice, it will be difficult to measure *all* possible intermediate inputs, and the extent to which they may have changed in response to the exogenously-varied treatment. Thus, it is perhaps most accurate to interpret the "causal estimate" of β from experimental studies as the "policy effect" of $X_{i,t}$ at the point when the outcomes are measured.

Note that this limitation is also present for non-experimental methods, and is therefore not a criticism of experiments per se. But it an important limitation to highlight because experimental estimates are often implicitly interpreted as representing production function parameters based on Eq. (1). This may well be true over short time periods where other agents may not have re-optimized behavior, but it is (a) difficult to confirm that this is true on every dimension of potential behavior modification, and (b) much less likely to be true over longer

horizons.[13] One advantage of well-identified evaluations using large administrative data sets (based on regression discontinuity designs for example) is that it may be possible to observe the policy effects at longer time horizons at much lower marginal cost than in experimental studies (since the cost of conducting follow up surveys on experimental samples can be quite large, and the challenge of differential attrition grows over time). A good example of this is provided by Bharadwaj et al (2013) who can measure the impact of early childhood interventions several years later using administrative data in both Chile and Norway. Longer-term follow ups of experimental interventions are relatively rare, but should be a higher priority for funders and researchers.

4.2 Interpreting zero effects

A second challenge in conducting inference from experimental studies is that of interpreting zero effects. In theory, this should simply mean that the estimate of β in Eq. 1 is zero and that the marginal impact of increasing $X_{i,t}$ is insignificantly different from zero. In practice, however it is important to distinguish between four different interpretations of a zero result. These include (a) poor implementation of the intervention, including corruption or administrative failures, (b) substitution away of other inputs by agents (including governments, schools, teachers, and households) in response to the treatment, (c) the intervention did not alleviate a binding constraint to education outcomes in the context being studied, and (d) absence of complementary inputs/reforms that may be needed for the intervention to be effective. Note that reasons (c) and (d) are consistent with the interpretation that the marginal impact of increasing $X_{i,t}$ on outcomes is zero in a production function sense, but reasons (a) and (b) are not. Further, the distinction between (c) and (d) also matters for policy because the policy implication of (c) may be to not prioritizing increasing $X_{i,t}$, whereas the policy implication of (d) would be to increase $X_{i,t}$ as long as the complementary input is also increased.

These possibilities are illustrated across four different randomized evaluations of the impact of providing books and materials to students (in some cases, these were directly provided as books and in other cases, the input was a grant to schools that was substantially spent on books and materials). Each of the four studies find zero average impacts of providing books and materials, but point to different possible reasons for the zero effects. Sabarwal et al (2014) find no impact on test scores from the provision of textbooks to schools in Sierra Leone and attribute this to the fact that schools actually stored the textbooks instead of distributing them to students (which is a form of poor implementation). Das et al (2013) described above also find no net impact on test scores from the provision of a school grant (that was mostly spent on

[13] While the discussion may suggest that experimental estimates may be lower bounds of production-function parameters and upper bounds of policy parameters, this need not be true if the unmeasured inputs are complements to the experimental intervention as opposed to substitutes (as was the case in Das et al 2013).

books and materials) in India, but attribute it to households offsetting the intervention by reducing their own spending on these inputs.

Glewwe et al (2009) also find no impact on test scores from providing textbooks to students in Kenya. But they find positive impacts on students with the highest baseline test scores and suggest that their results are consistent with the fact that the majority of children could not read the English language text books to begin with, and thus could not benefit from the textbooks (whereas those who could read *did* benefit). Thus, in this case, the non-impact is interpreted as suggesting that textbooks did not alleviate the binding constraint to learning in this context (which was the lack of reading).

Finally, Mbiti and Muralidharan (2015) also find no impact on test scores from the provision of a large capitation grant to schools in Tanzania (the largest item that the grant was spent on was textbooks). However, their study was explicitly designed to test for complementarities with teacher effort (which was boosted by a separate intervention that paid teachers bonuses based on student performance) using a cross-cutting design with a sample size large enough to test for complementarities, and they find that the interaction effect of the school grant and teacher performance pay was significantly positive. In other words, the school grant on its own had no impact, but had a significant impact when provided in conjunction with a teacher performance pay intervention. Thus, it is likely that the performance pay treatment contributed to teachers making more effective use of the additional materials, but it is also true that having the materials allowed teachers to significantly improve student outcomes relative to teachers who only increased effort due to performance-linked pay.

The larger point here is that each of these experiments with zero results are useful results in and of themselves, and yield an important policy conclusion that the marginal impact of providing books and learning materials to students may be very low on their own. On the other hand, the fact that four papers with the same result point to four different reasons for this non impact suggest that a "black box" experiment on its own may yield limited insights into the nature of the education production function and the true binding constraints to learning. Doing so requires either more ambitious experimental designs to test complementarities (which can quickly get prohibitively expensive) or significant investment in collecting data on intermediate processes and inputs and arguing (non-experimentally) that these factors (such as poor implementation, substitution, or impacts in a subset of the population) explain the observed "reduced form" estimates of program impact.

Thus, it is good practice for researchers running field experiments to think ex ante about how they would interpret a zero effect, and to collect as much data as possible on implementation quality as well as intermediate inputs and processes to enable better interpretation of finding no effects of a program.

4.3 External Validity

Perhaps the most widely commented on limitation of experiments is that of the external validity of their results beyond the specific setting where they are carried out. The formal way of thinking about this problem is to recognize that though the random assignment ensures that unobservables are distributed identically across treatment and control groups and that the treatment is not correlated with these unobservables, the estimated program effects are for not for the treatment *alone*, but rather for the treatment *interacted* with the unobservable characteristics in the study sample. If these unobservable characteristics vary between the study sample and the universe to which we seek to extrapolate the findings to, then the estimated treatment effects may not be valid because the interactions may change. There are several variants of this concern that are worth spelling out distinctly, because the strategies for mitigating them are different.

4.3.1 External Validity Concerns in the Same Context

There are at least four limitations to generalizing experimental results even in the *same context* where the experiment was carried out. First, there is a concern of external validity even in the context of the evaluation because most experiments are carried out within a universe of schools that agree to participate in the experiment. If these schools are different from those who do not agree to participate (perhaps their leadership is more open to trying out new ideas), then the results might have limited external validity (Heckman and Smith 1995).[14] The ideal experimental protocol to address this problem is to try as hard as possible to first draw a representative sample of schools/students from the universe that the study is trying to extrapolate to, and then randomly assign these schools into treatment and control groups. Such a protocol provides much more external validity than studies carried out in a "convenience sample" of schools and allows policy makers to be more confident that the experimental estimates apply to the relevant universe of interest.[15] While this may not always be possible, experimental studies should at least discuss their sampling procedure in more detail and show tables comparing the study sample and the universe of interest on key observable characteristics (similar to tables showing balance on observable characteristics across treatment and control units).

Second, a further concern with external validity even in the same context comes from the fact that many RCT's in education evaluate interventions implemented by committed

[14] A variant of this concern is seen in the US charter-school literature where well-identified estimates are only available on the causal impact of over-subscribed charter schools (which are likely to be the higher quality ones) as opposed to the universe of charter schools, which is the policy parameter of interest (unless the over-subscribed schools are able to expand without compromising quality and the schools that are not over-subscribed shut down).
[15] Examples of such an approach include Muralidharan and Sundararaman (2011, 2013) in the Indian state of Andhra Pradesh, de Ree et al. (2015) in Indonesia, and Mbiti and Muralidharan (2015) in Tanzania.

implementation partners (often highly motivated NGO's). Thus experimental estimates of programs implemented by NGO's may not translate if the same program is implemented by the government (as shown in Bold et al. 2013). The differences in the results they report between government and NGO implementation of a contract teacher program largely reflects the fact that the program itself was very poorly implemented by the government. So, it does not negate the results found under NGO-implementation, but it does highlight that programs are not just an "intervention", but rather an intervention *and* an implementation protocol. This is not a problem per se, but suggests that evaluations of NGO-led implementations should be seen as efficacy trials and not effectiveness trials.[16] It also suggests that when successful NGO-implemented interventions are being scaled up, there may be a strong case for conducting further RCT's at larger units of implementation and when implemented by the entity that will eventually scale up the intervention (typically a government).

Third, even if a government wishes to use experimental results in a given context to guide policy in the same context, it is extremely unlikely that the policy chosen will be exactly the same as the one evaluated. For instance, even if a CCT is found to have a positive impact, the value of the CCT may be changed later for political or budgetary reasons. Similarly, even if a teacher performance pay program or a student incentive program is found to be effective, a policy maker would care about the elasticity of the outcome of interest to the magnitude of the incentives in order to better calibrate the value of the incentives. These questions are harder to answer within the context of an experiment because it is often politically and administratively difficult to vary the magnitudes of such incentives within an actual program. While it may be possible to do this, a more promising approach would be to combine experimental methods with structural modeling to allow more credible out of sample predictions than either of the two approaches could on their own.

Good examples of this are Todd and Wolpin (2006), and Attanasio, Meghir, and Santiago (2012) who combine structural models of school participation with observed impacts of the Progressa CCT program to enable predictions of program impact under alternative values of the cash transfer. Another good example is Duflo, Hanna, and Ryan (2012) who use the non-linearities in the compensation schedule of an experimentally evaluated teacher incentive program to identify parameters in a dynamic model of teacher labor supply and use the model to estimate cost-minimizing compensation policies to achieve a desired level of teacher attendance. However, these additions of structural models to enable out of sample predictions have mostly been done ex post and were not designed ex ante into the study, which may have

[16] These terms are standard in the medical literature and refer to the difference between impacts under high-quality implementation that is closely monitored (efficacy trial), and impacts under typical implementation that allows for typical patient behavior including non-compliance with dosage frequency and complementary instructions (effectiveness trial).

limited the extent to which the experiment could be used to identify parameters in the structural model of interest. Future experimental research in education is likely to have greater impact if the ex-ante design of the study includes careful thinking about the model that the experiment can be used to identify.

A fourth and final concern regarding external validity in the same context is that experiments cannot typically capture the general equilibrium effects (both political and economic) that may accompany attempts to scale up successful smaller scale experiments. In the words of Acemoglu (2010), "Political economy refers to the fact that the feasible set of interventions is often determined by political factors, and large counterfactuals will induce political responses from various actors and interest groups. General equilibrium and political economy considerations are important because partial equilibrium estimates that ignore responses from both sources will not give the appropriate answer to counterfactual exercises".

A good example of this is the case contract teachers, where existing experimental and non-experimental evidence suggest that locally-hired contract teachers who are less educated, less trained, and paid much lower salaries than civil-service teachers are at least as effective (if not more) at improving learning outcomes in rural primary schools in both Kenya and India (Duflo et al. 2014; Bold et al. 2013; Muralidharan and Sundararaman 2013). Thus, expanding the use of contract teachers on the current margin would appear to be a very promising and cost-effective policy for improving education outcomes in developing countries. Nevertheless, scaling up contract teacher programs has been difficult politically because forward looking officials are aware that hiring a large number of contract teachers will lead to them getting unionized and creating political pressure to get "regularized" as civil-service teachers, which is very difficult for politicians to ignore.

This is a problem that is difficult to solve empirically with an experiment, but the discussion above highlights the importance of treating positive results from an experimental evaluation of an intervention as just one of many inputs into the policy-making process. Finding positive technical results from an RCT of an intervention can be a good starting point for considering the administrative and political challenges of scaling up and designing implementation protocols that take these into account, but it would be naïve to recommend "scale ups" based on RCT evidence alone. It is perhaps not a coincidence that the leading example of policy scale up based on RCT evidence is de-worming, which is administratively easy and politically costless. On the other hand, other interventions with robust evidence of positive effects (like the use of

contract teachers) have been much more difficult to scale up. The experimental evidence is only the starting point for the policy conversation in such cases.[17]

4.3.2 External Validity Concerns across contexts

Obtaining external validity across contexts is even more challenging, given that the unobserved covariates (that would interact with the treatment of interest to produce the average treatment effect) are likely to be different across contexts. This problem is well known among academic researchers, but is often under-stated in "systemic reviews" that compare interventions and estimates across contexts (see Pritchett and Sandefur 2013 for a discussion). There is no good solution to this problem beyond conducting more studies and gathering more evidence by replicating evaluations of similar (if not 'identical') interventions in many settings.

Despite the many successful field experiments in education in developing countries in the past decade, the overall experimental research agenda in education in developing countries is still at an early stage. Some pieces of evidence seem quite robust across several contexts (such as the lack of impact of providing books and materials to students), and others have been replicated in multiple sites in the same country (such as "teaching at the right level" across states in India), but most other interventions do not have enough replications across contexts to enable confident claims of their impacts across contexts. To the extent that donors and development agencies seek summaries of evidence (as seen by the eight summaries written in the last 2 years), attempts to calculate comparative cost effectiveness of interventions conducted across several contexts should be interpreted cautiously (Dhaliwal et al. 2013).

5. Design of Experiments in Education

5.1. Choosing and designing interventions worth evaluating

While experimental methods can help with credible estimation of the causal impact of interventions, it is important to ensure that adequate thought is given to determining whether the intervention being studied is worth evaluating and the extent to which the findings of a study generate more generalizable knowledge (especially given the non-trivial time and effort costs of setting up an RCT). In particular, it is not uncommon to see competently implemented and analyzed experiments in education, where the underlying intervention is rather ad hoc and not adequately theorized, which limits what we learn from the evaluation. In this section, I offer some (personal) guidelines for informing this decision.

Three useful questions to ask before deciding if an intervention is worth evaluating are: First, are governments spending large amounts of money doing things whose effectiveness we

[17] See Muralidharan (2013) for an example of a policy proposal that takes the results from evaluations of contract teacher programs and

do not understand (examples include infrastructure, class size reductions, teacher training, teacher salary increases, school feeding programs, school grants)? Second, is there a clearly documented market failure (or other source of inefficiency) that the proposed intervention hopes to solve? Third, is the *design* of the intervention itself sensible and in line with theoretical first principles?

It is not necessary that all three questions be answered in the affirmative before embarking on an RCT. It is possible that there are simple interventions that satisfy the second and third criteria but do not satisfy the first, and these may be the most useful interventions to evaluate because finding positive effects would make a case for allocating public funds for expanding the program. A good example is the provision of de-worming tablets to school children, which was found to be a much more cost effective way of increasing school attendance than other spending on other student inputs (Miguel and Kremer 2004, Dhaliwal et al. 2013).

On the other hand, if the policy/program being evaluated is something that governments spend a lot of money on, an RCT can be very useful even if the researcher has ex ante reasons to believe that the intervention is not well designed or may not be effective because the money will be spent on the program anyway, and it is very useful for policy to understand whether the program was effective. Further, to the extent that the program (as designed) reflected conventional wisdom that it would have a positive impact, the evaluation could shed light not just on the "program" as implemented but also on the hypothesis underlying the design of the program. A good example of such an evaluation is de Ree et al. (2015) who study the impact of an unconditional doubling of teacher pay in Indonesia. While we did have a prior expectation that this doubling may not have much impact on student learning (or at least that the same money could have been much better spent), the evaluation was still worth conducting because (a) the policy was very expensive, and (b) many education advocates believed that increasing teacher pay would improve teacher motivation, effort, and student learning.

On the other hand, there are several cases of new programs (that have not been scaled up) that are not well designed, in which case even an experimental evaluation may not contribute much to learning. One common mistake is to rush into an RCT before the intervention has been adequately piloted, codified, and stabilized. If the intervention is being modified during the study, it is difficult to interpret the findings. Thus, it is essential for programs and interventions to be "standardized" and "easily replicable" before embarking on an RCT. A related challenge occurs with RCT's of "composite" interventions that include components that are not easy to codify, which makes it difficult to interpret the results of an RCT. Note that a "composite" intervention per se is not a problem since it is often possible that there are complementarities across components of the package and the "package" may be the intervention that we need to

evaluate. Rather, it is the inability to codify and replicate the "package" that limits the learning from an evaluation of composite interventions.

Another pitfall to be aware of is that of studying "gold-plated" interventions that have high unit costs. The risk in such a setting is that a high-quality (but high-cost) intervention gets evaluated and is found to have a substantial positive impact, but is difficult to scale up because of a lack of financial resources to sustain the program.[18] A further risk is that a diluted version of the high-cost prototype is scaled up (on the basis of the evaluation), but that this version may not have any impact.

One option for imposing discipline in this regard is to not only have a "pure" control group (that does not get any additional intervention), but to have other comparison groups that are provided an equivalent amount of resources, which enables a direct cost-effectiveness comparison against reasonable policy alternatives. An example of this is provided by the Andhra Pradesh Randomized Evaluation Studies (APRESt) where the impact of teacher performance pay was evaluated not just against a pure control group, but against comparison groups that received an equivalent valued school grant or extra contract teacher.[19]

There are also more subtle ways in which opportunity costs may not properly accounted for in the design of evaluations. Consider the case of modifying curriculum by teaching new content. It is crucial to also specify what is being replaced in the existing curriculum to make way for the new additions and to test if there are negative impacts on learning of subjects that may have had their instructional time reduced to make way for the new content. On the other hand, if the new materials are being taught over and above the existing content, it is important to price the opportunity cost of student and teacher time. This may be low or high, but needs to be accounted for.

A good example of the importance of accounting for time use in schools is provided by Linden (2008) who finds that a computer enabled instruction program had positive effects on student test scores when offered as an after school supplementary program, but had negative impacts when it was used to substitute existing teaching activity. A second example is provided by Muralidharan and Sundararaman (2015) who find in their study of school vouchers that there was no impact on math and native language test scores of winning a voucher and attending a private school. However, they also find that private schools spend much less

[18] One manifestation of this is the phenomenon of donor-financed pilot projects being abandoned once the donor financing is over. Of course, these programs typically do not have credible impact evaluations to inform the decision on whether developing country governments should continue to spend on them out of their own budgets. But scaling up of high unit-cost interventions would be difficult even with positive evidence of impact.

[19] Blattman and Niehaus (2014) make a similar point with regard to evaluations of anti-poverty programs in general, proposing that the benchmark should not just be a pure control group, but rather an unconditional cash transfer that is equal in value to the full cost of the program being evaluated.

instructional time on math and native language and use the time saved to teach other subjects (where they strongly outperform the public schools as may be expected). The most important general lesson here is that the inference on the relative productivity of public and private schools would have been incorrect if the differential patterns of time use had not been accounted for.

Thus, an important lesson for the design of evaluations of education interventions in general is to account for all costs of the intervention - including *time* and financial costs, and being clear about whether the impact on test scores is coming from additional time on task (either home or school work) or from using existing time more efficiently (either by organizing pedagogy more efficiently or by reducing slack during the school day).

5.2. Choosing the unit of randomization

As discussed earlier, the "modular" nature of many education interventions makes it feasible to randomize them at relatively small unit levels, including at the student, classroom, and school levels. The main consideration in determining the unit of randomization is the trade-off between statistical power[20] for a given budget (which is higher for lower units of randomization) and the possibility of spillovers, which may bias the experiment and negate the power gains from randomizing at a lower level.

The most striking example of this trade-off is perhaps seen in the difference between results from studies of the impact of deworming school children that randomized the treatment at the student level, and the results reported in Miguel and Kremer (2004) who randomized the provision of deworming tablets at the school level. While earlier studies typically found very limited impacts of deworming on education outcomes, Miguel and Kremer (2004) found large positive impacts of doing so, and also find strong evidence of spillovers from treated to non-treated students. The spillovers would under-state the impact of deworming in studies with student-level randomization in two ways – first, it would under-estimate the impact on the treated students (because the control group also gained), and second it would not count the impact on control students who also benefited. Of course, randomizing at the school level significantly increases the sample size required for adequate power, and correspondingly increase the cost of the study. On the other hand, it is not clear that the gain in power from student-level randomization is worth it if it biases the treatment effect itself.

Note however, that the case of de-worming may be an outlier in terms of the extent of spillovers across students. Several studies within the REAP program have utilized student-level randomization within schools in China to study the impacts of interventions ranging from peer-

[20] Note that this chapter does not spend too much time on generic issues of experimental design such as power calculations, for which there are many other references. Rather the focus is on design choices that are especially relevant to research in education.

tutoring to student incentives. Berry (2013) uses student-level randomization to study the impact of different combinations of student and parent incentives. Nevertheless, given that students interact with each other every day in the classroom, it is difficult to credibly claim that there would not be any spillovers (especially for treatments implemented in schools as opposed to households), which may limit the extent to which we can learn from such experiments. One important exception is cases where the treatment is at the school level and control students are not in treated schools, as happens in cases where students receive vouchers (or charter-school admission) by lottery and do not interact with control students during the course of the school day. But overall, student-level randomization designs should be used with caution and limited to situations where the focus of the intervention is outside the school setting (such as households or after-school programs).

A less extreme set of concerns applies to designs that randomize at the grade or classroom level as opposed to the school level. Again, the main reason for doing so is power and cost. Several prominent studies have used classroom-level randomization designs to study the impact of remedial instruction (Banerjee et al. 2007), tracking of students according to initial ability (Duflo et al. 2011), and comparing contract and regular teachers (Duflo et al. 2013). The spillover concerns are less severe in this case because most instructional activity as well as peer interactions between students happens at the classroom level and not across classrooms (which is where the spillovers would have to happen).

However, there is still a concern that interventions that provide a significant increase in resources to some classrooms, may lead a head-teacher to offset some of the impact of the treatment by reallocating some other resources to control classrooms. This could happen either due to norms of fairness within the school or because an optimizing head-teacher would reallocate resources to equalize their marginal product across classrooms.[21] Such behavioral responses could contaminate the experiment and the inference.

This is why my personal preference has been to randomize at the school-level to the extent possible for studying interventions ranging from teacher performance pay, across the board teacher salary increases, provision of school grants, provision of diagnostic feedback to schools, and also the provision of an extra contract teacher. The overall logic of this approach is that a policy maker can typically target resources at the school level, but cannot control how those resources are allocated within the school by an optimizing head-teacher. Thus, the policy-relevant parameter of interest in these cases is the impact of a school-level provision of an intervention. The cost of this approach of course is that the samples need to be much larger and the studies cost more.

One rule of thumb for choosing between school and grade/classroom level randomization may be to consider the size of the school. In smaller schools (as is the case in rural India where

[21] For instance the model in Lazear (2006) predicts that more disruptive students will optimally be assigned to smaller classrooms

most primary schools have less than 100 students across 5 grades and the modal school has only 2 teachers teaching in multi-grade classroom settings), it is difficult to convincingly argue that a within-school randomization protocol that assigns a program to just some grades will be adhered to without re-adjustment. On the other hand when schools are much larger and have several hundred students and dedicated teachers in each grade (as is the case in many African settings), the fidelity of within-school experimental protocols may be more reasonable to assume because teachers spend all their time with one grade as opposed to teaching multiple grades at the same time. Further, the costs of school-level randomization may be prohibitive in settings with such large schools.

Overall, my view is that it is less of a problem for an intervention to be targeted at specific grades within a school (as opposed to the entire school) as long as the control group is the same grade in a *different* school as opposed to other students (especially) within the same school. In such a setting one may still worry about spillovers attenuating the treatment (if resources are diverted to other non-treated grades in treatment schools), but at least the control group will not be contaminated. Studies that use within-school controls should have the burden of proof for demonstrating that the controls were not contaminated by spillovers.

5.3. Cross-cutting (interaction) designs

To be completed later.

The main issue is that you need to pre-commit to a view that interactions are not important if you want to double-count the sample with multiple treatments under both treatments (to increase power within a given measurement budget). But in practice, interactions *are* likely to be important, and there is risk of overstating treatment effects if you ignore the interactions. The first wave of experiments regularly used cross-cutting designs and ignored interactions which was a sensible way to proceed given tighter budgets and the need to have adequate power for the first order questions.

But overall, I am not such a big fan of this, especially as we generate evidence that interactions do matter, and are detectable with designs that treat them as equally important to the main effect and allocate adequate sample size to detect them. Since evaluation budgets are growing (as they rightly should with programs increasingly being expected to set aside funds for evaluation), I would be in favor of cleaner designs with single treatments. Note that the treatment can be a "composite" one if that is the treatment of interest. But ignoring interactions feels like an ad hoc assumption that may not be justified.

5.4. Data Collection and Analysis

To be completed later. Topics to be covered include:

- Data collection (outcomes, implementation/first-stage, intermediate mechanisms, spillovers, and substitution – both time and resources and at home and school)
    - Main message – think really hard about possible mechanisms of impact and try to collect the data for testing/showing these
    - In general, experiments that help shed light on mechanisms and pathways of impact are much more useful than those just give us black box treatment effects
    - Think about the full distribution of potential treatment effects and what you would learn from each of them and what intermediate data you need to be able to tease apart different explanations for your results.

- Analysis
    - Pre-analysis plans
    - In addition to standard treatment effects, education interventions are particularly well-suited to non-parametric analysis
    - Quantile treatment effects; Non-parametric TE as a function of baseline scores
    - These are not the same because of non-preservation of rank over time

6. How do we maximize learning from Field Experiments in Education?

While the rapid increase in field experiments in education in the past decade has produced several high-quality individual studies and yielded important insights, there are also many limitations and caveats to what we have learnt (as discussed in section 4). In this section, I discuss ways of mitigating these limitations and present some thoughts on structuring field experiments to maximize what we learn from future research. I first outline the range of evaluation scenarios that experimental researchers on education in developing countries typically face and highlight the strengths and limitations of each of these settings.

6.1. Characterizing the typical evaluation scenarios

Experimental evaluations in education can broadly be classified into three types of operational scenarios. The first scenario is one of evaluating the impact of existing government programs that have been designed by government officials without much (or any) inputs from the researchers/evaluators and which have limited scope to be modified by the researchers, but where it is possible to randomize the roll out of the program. A good example of such an evaluation is the Indonesian teacher certification law of 2005 that instituted a process by which teachers who got "certified" would be eligible for a doubling of their base pay. The law was passed by the Indonesian parliament and could not be modified, but researchers were able to work with the National Ministry of Education to randomize the roll out of program eligibility in a representative sample of schools and conduct an experimental evaluation of the impact of unconditionally doubling teacher pay on student learning outcomes (de Ree et al. 2015).

The second scenario is of an intervention designed by non-government implementing partners, who have programmatic and field-based inputs into the design of their programs, and work with researchers to evaluate these interventions. This setting typically allows researchers to offer more inputs into the design of the intervention and the evaluation strategy, but the degrees of freedom may be limited by operational considerations. The ideal situation provides for a deep ongoing partnership between researchers and implementing partners to iterate the design of programs based on feedback from ongoing evaluations. A good example of such a partnership is the one between JPAL and Pratham in India spanning over a decade, where early positive evidence of the positive impact of a remedial education program in primary schools (Banerjee et al. 2007) led to further refinement and evaluation of different ways of effectively delivering such programs at a larger scale (Banerjee et al. 2015).

The third scenario is one where even the design of the intervention is researcher led. These evaluations also typically have implementation partner organizations who may offer inputs based on operational feasibility considerations, but the projects are researcher-led and are more likely to be focused on testing theories about constraints and enablers of better education outcomes than on evaluating the impacts of specific programs per se. Examples of such studies include Muralidharan and Sundararaman (2011), and Muralidharan and Sundararaman (2015). In both cases, the studies mapped into very specific policy questions regarding the positive (and potentially negative) impacts of teacher performance pay and school choice respectively. Other examples of such studies are Jensen (2010) on the impact of better information on returns to education in the Dominican Republic, Duflo et al. (2011) on the impact (positive or negative) of tracking classrooms by initial levels of student achievement in Kenya, Duflo et al. (2012) on the impact of paying daily piece rates to teachers based on attendance (verified by cameras), and Andrabi et al. (2014) on the impact of providing market-level information on school performance in Pakistan. The salient feature of these studies (across very different locations and research questions) is that the design of both the *intervention* and the evaluation in these cases was completely led by the researchers and the implementing partner did not have a direct stake in the results of these studies.

Each setting offers both advantages and challenges. In the first scenario, good studies have the potential to provide highly-credible estimates of “as is” implementation of major government policies (as in the case of de Ree et al. 2015). Since the programs are government-designed (and typically intended to be implemented at scale), these studies may be the most directly policy relevant. On the other hand, such settings offer limited scope for researchers to consider and evaluate alternative designs that may have improved the effectiveness of the spending incurred under the program. In particular, if the intervention was "poorly designed", finding that there was no impact is still a useful result to know about that specific program, but may not yield much knowledge beyond that.

The second scenario offers the promise of combining researcher and implementer inputs to design and evaluate interventions that are both theoretically promising and likely to be *implementable* at a moderate scale (at least at the level of the partner agency).[22] The main challenges of this setting include: (a) the implementing agency often does have a stake in the success of the program being evaluated and may want to iterate the design more quickly than the evaluator (who would want a stable design to evaluate), (b) the range of ideas that can be tested may be limited by pre-determined programmatic areas/priorities of the implementing partner, and (c) programs implemented by NGOs may still be too small relative to that of those implemented by governments and evaluations of these programs may not yield the relevant policy parameter of interest when implemented in a truly scaled up setting.

The third case is often the most attractive for researchers since it gives them much greater control on the design of the intervention to both maximize the theoretical possibility of positive impact and to mitigate negative impacts. It is perhaps no coincidence that many of the most influential academic papers[23] in the past decade in this area have followed this model, because the researcher-led designs typically yield the most insights in terms of identifying (a) specific binding constraints in the status quo, and (b) promising ways of alleviating these constraints. On the other hand, the main limitation of these types of studies is that they mostly constitute a "proof of concept" or an "efficacy" trial that demonstrates the impacts of the intervention studied under high-quality implementation, but cannot be the basis for assuming that they will achieve similar impacts under scaled up implementation by governments (Vivalt 2014).

Broadly, there appears to be a continuum of trade-offs across these scenarios between the validity of the results at a policy-relevant scale on one hand (where the first scenario does best) and researcher-control and ability to test theoretically-motivated intervention designs on the other (where the third one does best). Each type of study is potentially valuable, and will typically provide useful building blocks in understanding how to improve education outcomes. However, the trade-offs discussed above suggest that it will be difficult for a single study to satisfy the twin objectives of being amenable to researcher inputs in the design of the intervention as well as being evaluated at a policy-relevant scale. This trade-off points to the importance of research *programs* and not just individual studies.

6.2. Designing research programs in education

[22] The KiuFunza project in Tanzania is a good example of such a partnership where a theoretically sub-optimal teacher performance pay scheme was deliberately implemented and studied because of the primacy placed on having a design that could be implemented at scale (see Mbiti and Muralidharan 2015).

[23] This is based on an informal review of papers in the economics of education in developing countries that have been published in "top 5" economics journals in the past decade and on their citation counts. I will formalize this in the revision.

As discussed in section 4 concerns of external validity are a central limitation in our ability to make policy-relevant inference from experimental research in education. I argue here that many of these challenges can be mitigated by focusing on research programs as opposed to individual studies. In particular, while individual researchers are more likely to be involved with a few studies at a time, funding agencies should try to prioritize research programs with the following characteristics. [The discussion of the points below is in an outline format]

6.1 Measurement

Creating a common scale for comparing outcomes across studies in different parts of the world. Can work with TIMSS/PISA/PIRLS to access question banks from past tests and to encourage individual projects/studies to include questions from this question bank to allow IRT-based calibration of treatment effects on a common scale. It is then easy to take the items in disparate studies around the world and also calibrate them on the same scale.

6.2 Studying the same intervention in multiple settings

There is great value in repeating the same intervention in multiple settings to understand the extent to which treatment effects generalize across contexts. This only makes sense for interventions that have proven highly effective in at least some contexts. Candidates include:

- Teaching at the right level
- Contract teachers
- Well-designed teacher performance pay systems

6.3 Studying the same intervention in multiple settings

There is also great value in evaluating multiple interventions in the same setting. Cost-effectiveness calculations are much more convincing when the distribution of unobserved covariates is the same across interventions. Examples include:

- Kenya studies
- AP studies

6.4. Designing for long-term follow up

- Very important, but also very expensive, and difficult to do well because differential attrition challenges mount over time
- Perry pre-school; Jamaica Early Childhood interventions; Chetty et al with Tennessee STAR have all been extremely valuable

- Strong case for experiments in systems with high-quality administrative data that would enable long-term follow ups

6.5. Section summary

At its best, such a research program will feature partnership between researchers and implementing agencies (including governments) that can provide for an iterative approach to intervention design, evaluation, and scale up. Such a program would feature (a) researcher *inputs into initial program design* based on theory, and an analysis of existing data and prior research to identify binding constraints to education attainment and achievement that can be mitigated with an intervention, and articulating a clear mechanism by which the intervention is expected to have an impact; (b) pilot-level projects to refine the intervention to be evaluated and "stabilize" the implementation model that is going to be evaluated (since it is not very useful to conduct an RCT of an intervention that is changing during the evaluation); (c) an RCT at a smaller unit of randomization (such as a school) to evaluate the "efficacy" of the intervention when well implemented; (d) refining program details based on evaluation data on both processes and outcomes; (e) building systems needed to implement the interventions at progressively larger scales if it is found to have a significant positive impact and be cost effective based on the evaluation results (including working with governments for building systemic implementation capacity); and (f) conducting additional RCT's of the intervention at larger units of implementation to assess the "effectiveness" of the intervention.[24]

In practice, such a program will be challenging to set up, will require a long time horizon, and may be beyond the scope of an individual researcher to set up, and there is no good precedent to point towards as a model. The closest example of such a collaboration is perhaps the one between JPAL South Asia and Pratham to test variants of programs that aim to "teach at the right level" and to partner with state governments in India to implement and evaluate at a larger scale than those of the earlier studies that informed the design of the program.[25] Nevertheless, building on the past decade of successful individual experiments in education and bridging the gap between "efficacy" and "effectiveness" studies will require such an approach. Funders of education programs in developing countries would do well to consider building in such an iterative approach to evidence-based policy making in their program design.

7. Conclusion

To be completed

[24] This is the model being followed by the author in an ongoing evaluation of a state-level school quality improvement program in the Indian state of Madhya Pradesh.

[25] I will describe the specific states and studies in more detail later in this footnote (because many of these studies are still not in working paper form though slides/analysis do exist).

References (very incomplete):

ANDRABI, T., J. DAS, and A. KHWAJA (2014): "Report Cards: The Impact of Providing School and Child Test-Scores on Educational Markets," Harvard Kennedy School.

ANGRIST, J., E. BETTINGER, E. BLOOM, E. KING, and M. KREMER (2002): "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92, 1535-1558.

ANGRIST, J., E. BETTINGER, and M. KREMER (2006): "Long-Term Consequences of Secondary School Vouchers: Evidence from Secondary School Records in Colombia," *American Economic Review*, 96, 847-862.

BAIRD, S., C. MCINTOSH, and B. OZLER (2011): "Cash or Condition: Evidence from a Cash Transfer Experiment," *Quarterly Journal of Economics*, 126, 1709-1753.

BANERJEE, A., R. BANERJI, E. DUFLO, R. GLENNERSTER, and S. KHEMANI (2010): "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," *American Economic Journal: Economic Policy*, 2, 1-30.

BANERJEE, A., R. BANERJI, E. DUFLO, and M. WALTON (2012): "Effective Pedagogies and a Resistant Education System: Experimental Evidence on Interventions to Improve Basic Skills in Rural India," MIT.

BANERJEE, A., S. COLE, E. DUFLO, and L. LINDEN (2007): "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics*, 122, 1235-1264.

BANERJEE, A., and E. DUFLO (2011): *Poor Economics*. MIT Press.

BARRERA-OSORIO, F., M. BERTRAND, L. LINDEN, and F. PEREZ (2011): "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia," *American Economic Journal: Applied Economics*, 3, 167-195.

BARRO, R. J. (1991): "Economic Growth in a Cross Section of Countries," *Quarterly Journal of Economics*, 106, 407-43.

BHARADWAJ, P., K. LOKEN, and C. NIELSON (2013): "Early Life Health Interventions and Academic Achievement," *American Economic Review*, 5, 1862-1891.

BLIMPO, M. (2014): "Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin," *American Economic Journal: Applied Economics*, 6, 90-109.

BOBONIS, G., and F. FINAN (2009): "Neighborhood Peer Effects in Secondary School Enrollment Decisions," *Review of Economics and Statistics*, 91, 695-716.

BOLD, T., M. KIMENYI, G. MWABU, N. A. A. ALICE, and J. SANDEFUR (2013): "Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education," Washington DC: Center for Global Development.

BURDE, D., and L. LINDEN (2013): "Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools," *American Economic Journal: Applied Economics*, 5, 27-40.

CHAUDHURY, N., J. HAMMER, M. KREMER, K. MURALIDHARAN, and F. H. ROGERS (2006): "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *Journal of Economic Perspectives*, 20, 91-116.

CHETTY, R., J. N. FRIEDMAN, and J. E. ROCKOFF (2011): "The Long-Term Impact of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," Harvard.

CONN, K. (2014): "Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Rigorous Impact Evaluations," Columbia.

DAS, J., S. DERCON, J. HABYARIMANA, P. KRISHNAN, K. MURALIDHARAN, and V. SUNDARARAMAN (2013): "School Inputs, Household Substitution, and Test Scores," *American Economic Journal: Applied Economics*, 5, 29-57.

DE REE, J., K. MURALIDHARAN, M. PRADHAN, and H. F. ROGERS (2015): "Double for Nothing: Experimental Evidence on the Impact of an Unconditional Teacher Salary Increase on Student Performance in Indonesia," UC San Diego.

DHALIWAL, I., E. DUFLO, R. GLENNERSTER, and C. TULLOCH (2013): "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education," in *Education Policy in Developing Countries*, ed. by P. Glewwe: University of Chicago Press.

DUFLO, E., P. DUPAS, and M. KREMER (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101, 1739-74.

— (2012): "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools," National Bureau of Economic Research. Working Paper 17939.

DUFLO, E., R. HANNA, and S. RYAN (2012): "Incentives Work: Getting Teachers to Come to School," *American Economic Review*, 102, 1241-78.

EVANS, D. K., and A. POPOVA (2015): "What Really Works to Improve Learning in Developing Countries ? An Analysis of Divergent Findings in Systematic Reviews," World Bank Policy Research Working Paper 7203.

FISZBEIN, A., and N. SCHADY (2009): *Conditional Cash Transfers: Reducing Present and Future Poverty*. Washington DC: World Bank.

FOSTER, A., and M. R. ROSENZWEIG (1995): "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture " *The Journal of Political Economy*, 103, 1176-1209.

GLEWWE, P., E. HANUSHEK, S. HUMPAGE, and R. RAVINA (2013): "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010," in *Education Policy in Developing Countries*, ed. by P. Glewwe: University of Chicago Press.

GLEWWE, P., N. ILIAS, and M. KREMER (2010): "Teacher Incentives," *American Economic Journal: Applied Economics*, 2, 205-227.

GLEWWE, P., M. KREMER, and S. MOULIN (2009): "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, 1, 112-135.

GLEWWE, P., M. KREMER, S. MOULIN, and E. ZITZEWITZ (2004): "Retrospective Vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," *Journal of Development Economics*, 74, 251-268.

KREMER, M. (1993): "The O-Ring Theory of Economic Development," *The Quarterly Journal of Economics*, 108, 551-75.

KREMER, M., C. BRANNEN, and R. GLENNERSTER (2013): "The Challenge of Education and Learning in the Developing World," *Science*, 340.

KREMER, M., and E. MIGUEL (2004): "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72, 159-217.

KREMER, M., E. MIGUEL, and R. THORNTON (2009): "Incentives to Learn," *Review of Economics and Statistics*, 91, 437-456.

KREMER, M., K. MURALIDHARAN, N. CHAUDHURY, F. H. ROGERS, and J. HAMMER (2006): "Teacher Absence in India," Harvard University.

KRISHNARATNE, S., H. WHITE, and E. CARPENTER (2013): "Quality Education for All Children? What Works in Education in Developing Countries," Internationl Initiative for Impact Evaluation (3ie) Working Paper 20.

LALIVE, R., and A. CATTANEO (2009): "Social Interactions and Schooling Decisions," *Review of Economics and Statistics*, 91, 457-477.

LUCAS, R. (1988): "On the Mechanics of Economic Development," *Journal of Monetary Economics*, 22, 3-42.

— (1990): "Why Doesn't Capital Flow from Rich to Poor Countries?," *American Economic Review*, 80, 92-96.

MANKIW, G., D. ROMER, and D. WEIL (1992): "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, 107, 407-437.

MCEWAN, P. (2014): "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments," *Review of Educational Research*.

MURALIDHARAN, K. (2013): "Priorities for Primary Education Policy in India's 12th Five-Year Plan," *India Policy Forum*, 9, 1-46.

MURALIDHARAN, K., and V. SUNDARARAMAN (2011): "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, 119, 39-77.

— (2013): "Contract Teachers: Experimental Evidence from India," NBER Working Paper 19440.

MURNANE, R., and A. GANIMIAN (2014): "Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations," NBER Working Paper 20284.

SEN, A. (1993): "Capability and Well-Being," in *The Quality of Life*, ed. by M. C. Nussbaum, and A. Sen. Oxford: Clarendon Press, 30-53.

SNILSTVEIT, B., E. GALLAGHER, D. PHILLIPS, M. VOJTKOVA, J. EYERS, D. SKALDIOU, J. STEVENSON, A. BHAVSAR, and P. DAVIES (2014): "Education Interventions for Improving the Access to, and Quality of, Education in Low and Middle Income Countries: A Systematic Review," Campbell Collaboration Systemic Review Title Registration (Extended Proposal).

TODD, P. E., and K. I. WOLPIN (2003): "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal*, 113, F3-33.

TODD, P. E., and K. I. WOLPIN (2006): "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *American Economic Review*, 96, 1384-1417.