Randomized Control Trials: Limitations for Explaining and Improving Learning Outcomes

Moses Oketch
UCL

1 Introduction

Randomized control trials (RCTs) have been advanced by some development economists as the "gold" standard method of inquiry for identifying solutions on how to resolve intractable challenges in low-income and lower-middle income countries. One such intractable problem in countries committed to universal primary education is how to improve learning outcomes, that is, to ensure that the enrolled students read and do math competently at the appropriate grade level. There are no simple solutions as numerous decisions are made along the following lines: increase textbooks, enhance the opportunity to learn (which in developing countries, Fuller et al. (1994) define as the actual time spent teaching the relevant subject matter), reduce class sizes by hiring more teachers, hold teachers accountable, improve school management, train teachers, and so forth. Yet none of these policy solutions seemed to have yielded straightforward outcomes, and few evaluations have produced evidence that would have inspired confidence in the chosen reform trajectories. Therefore, studies that use randomized control trials (RCTs) to assess impact have been framed and sold to governments as the only reliable tool that allows for evidence-based decision-making. Although the appeal of RCTs has remained strong among development economists and policy makers in recent years, there has been some pushback by some governments and education researchers. This chapter provides a critique of the claim that RCTs provide 'gold standard' evidence for policy makers. The argument is framed in relation to Banerjee and Duflos' (2011) claim about the superiority of RCTs compared to country comparisons. A summary of the East Africa Quality in Early Learning (EAQEL) study is provided, before moving to a discussion of the limitations of RCTs and suggestions for alternative approaches. The central argument is that RCTs frame out too much of what is important about, and for, system-wide reforms and thus are of limited value for policy making. Instead, there is a need to turn to political economy approaches that focus on institutions and political structures, because focusing solely on RCT evidence may distract from crucial macro-level factors (e.g. labour relations). This chapter

does not seek to provide a systematic critiquing of the RCT methodology itself (i.e. it is not an epistemological critique). Instead it focuses on the limitations of RCTs as a tool for policy.

1.1 The High Expectations Associated with RCTs

In the book *Poor Economics*, Banerjee and Duflo (2011) present their argument that simple country comparisons led by 'experts' have fallen short of yielding evidence that inspires confidence for solutions or an intellectual basis for finding solutions to intractable development problems in poor countries. They advocate randomized controlled trials as the 'gold' standard alternative approach to generating evidence of what works. One such intractable development problem facing low and lower-middle income countries is how to put an effective education system in place that is able to deliver equitable access and quality. Even though the number of students enrolled in schools has increased significantly over the past two decades, schools in low-income countries face a 'learning crisis' (Unesco, 2014). There is consensus that efforts should be made to analyze this learning crisis, otherwise it will lead to erosion of the gains made in access. A quick and right solution is not easy or even agreed upon because, firstly, there are multiple definitions of what quality means and even much disagreement about how it should be measured. In this context, the randomized controlled trial (RCT) approach to evaluate the impact of the various initiatives aimed at addressing this intractable problem of 'learning crisis' is being promoted, in countries such as Liberia, where recently the government has embraced a 'Partnership Schools' model of public-private partnership to support provision of education. The preliminary evaluation of this 'Outsourcing Model' in Liberia has been reported by Romero, Sandefur and Sandholtz (2017). They have observed in their preliminary report (Romero et al., 2017, pi): "After one year, public schools managed by private contractors in Liberia raised student learning by 60 percent, compared to standard public schools". This is unbelievably remarkable improvement and could only be attained where there was no learning at all. They qualify this finding as follows: "But costs were high, performance varied across contractors, and contracts authorized the largest contractor to push excess pupils and underperforming teachers onto other government schools" (Romero et al., 2017, pi). It is this qualifying statement that summarises the concerns raised about RCTs in education. Firstly, on the surface of it, this sounds unethical on the part of the contractors and the partnership itself, if the 60 percent results are difficult to scale up and replicate in real practice. This is one example of the challenges facing RCTs. This chapter aims to draw on another RCT approach in two East

Africa countries of Kenya and Uganda aimed at supporting raising learning outcomes. The initiative known as East African Quality in Education and Learning (EAQEL) was implemented by Aga Khan Foundation and independently evaluated through RCT by a team of researchers lead by the author of this chapter. For details of the evaluation see Oketch, Ngware, Mutisya, Admassu, Abuya, and Musyoka (2014) and Lucas, McEwan, Ngware and Oketch (2014), and additional related aspects of the study have been analysed and written up in Ngware, Abuya, Oketch, Admassu, Mutisya and Musyoka (2015) and Abuya, Oketch, Ngware, Mutisya, and Musyoka (2015). This chapter first presents the study in summarised form and then reflects on the lessons learnt and limitations of RCTs in education research.

1.2 Examples of Impact Evaluation Studies using RCTs

The design of EAQEL, presented in this chapter, was informed by prior studies in Sub-Saharan Africa on impact evaluation in education. Examples include the flip chart study by Glewwe, Kremer, Moulin, and Zitzewitz (2000); a merit scholarship program for adolescent girls by Kremer, Miguel, Thornton and Ozier (2005); Vermeersch and Kremer (2004) on the effects of subsidized school meals on school participation, educational achievement and school finance; the study on teacher incentives based on students' scores by Glewwe, Ilias and Kremer(2003); Glewwe, Kremer and Moulin (2007) on the impact of text books on test scores; and the study on the effect of deworming school children on school attendance (Miguel and Kremer, 2004). These studies demonstrate the traction that RCTs were gaining in the region, especially in Kenya, yet they had not provided answers to the intractable problem of low quality education at the systems level. EAQEL was clearly aiming to add evidence to this or reveal something new and different from similar RCT approaches.

The study on flip charts by Glewwe et al. (2000) used retrospective and prospective analyses of flip chart provision to assess the effect of flip charts on student scores in rural schools. The retrospective analysis showed an effect of up to 20 percent, after controlling for other learning inputs but the prospective analyses concluded that there was no effect.
Glewwe et al. (2003) examined teacher incentives and their effect on students' scores. Fifty schools were selected from a group of 100 schools that were considered by the Ministry of Education to be in need of assistance. On average, these schools performed more poorly in examinations than other schools in the area (Busia and Teso districts in Kenya). Schools were numbered alphabetically and the odd numbered schools were chosen to participate in the teacher incentive program. Teachers of grade 4 to 8 participated in the study – the incentive

was a 21 – 43% of the monthly salary award at the end of the year based on the best performing school and/or best improved schools in grade 4-8 district mock exams. The study examined the differences in test scores between the treatment and comparison schools using a random effect regression framework that allowed for the possibility that scores of students in the same grade and same school might be correlated due to unobserved characteristics of teachers and headmasters. The Kenya Certificate of Primary Examination scores were also used to independently evaluate the impact of the incentive. The study utilized the difference in test scores between treatment and comparison schools, and the difference-in-difference estimator of the effect of the program. Students in the incentive schools had higher test scores during the program period, due to the short-run test scores effect. There was no teacher effort aimed at increasing long-run learning. The study also found that teacher attendance did not improve, homework assignment did not increase, and pedagogy did not change. Here again was an example of less direct solution at systems level for the intractable challenge of low quality learning. Nevertheless, these studies provided a base for EAQEL, hoping that the design of EAQEL and its Core Model, and Core Model Plus, was unique and may yield meaningful and scalable impact (Oketch, et al. 2014)

2 The RCT of the East African Quality in Education and Learning

The summary of EQEAL presented here draws on Oketch et al (2014). EAQEL was a research and development initiative which was aimed at demonstrating effectiveness of a model for improving learning outcomes in reading and numeracy in early primary grades (1-3) in two districts (Kwale and Kinango) in Kenya and two (Amolatar and Dokolo) in Uganda. The initiative was implemented over a period of 16 months by the Aga Khan Foundation and independently evaluated through a randomized controlled trial approach under my leadership, whilst based at African Population and Health Research Centre as the head of education research. The EAQEL project tested an instructional approach known as the scaffolding model (Reading to Learn otherwise abbreviated as RtL), described as a systematic approach to the teaching of reading with subsequent impact on numeracy.

The project design had three components: 1) teacher preparedness and practice, 2) school leadership, and 3) classroom learning environments. These components were embedded into two separate but mutually inclusive modules: the “Core Model” and “Core Model Plus”. The Core Model involved early grade teachers being trained on the instructional approach, which

was child-centered, systematic and focused on social interaction. In addition, schools were supported to improve teachers' and pupils' access to and use of appropriate teaching and learning materials. Project technical staff worked with head teachers, key teachers and district education staff from decentralized teacher support resource institutions to train teachers and provide in-class mentoring support. The Core Model Plus included all of the aspects of the Core Model and a parental involvement component. The aim was to encourage literacy by establishing mini-libraries in selected homes, and encouraging parents to borrow books, read and tell stories to their children (Oketch et al. 2014)

The goal of the impact evaluation was to assess the effectiveness of the project in terms of: (i) whether it led to improved learning outcomes in numeracy and reading among children enrolled in primary grades 1, 2 and 3 as was intended; (ii) whether there was a critical difference in the learning outcomes of children enrolled in grades 1, 2 and 3 attributable to the two different treatment models (Core model and Core Model Plus) and; (iii) ascertain what were the key contributing factors that lead to improvements, which could be exploited for policy relevance and project scale up (Oketch et al. 2014).

2.1 The Context

Uganda and Kenya had been successful in the implementation of universal basic education in 1997 and 2003, respectively (Oketch and Rolleston, 2007; Oketch and Somerset, 2010). This had led to a remarkable growth in enrolment at primary level with negative consequences for learning. Therefore, the need for a project such as EAQEL was a natural one and required little justification. Here was an intractable problem of how to accelerate learning for many of these children who had gained a chance to enter schooling as a result of universal basic education policy implemented in both countries. No one could have reasonably argued against this idea of supporting learning improvement, involving parents and using locally available resources, and to search for evidence that it works as designed via the EAQEL project. As indicated earlier, the project was implemented in two districts in each of the countries. In Kenya, these were Kwale and Kinango districts in the coast (it should be noted here that since 2010, Kenya no longer has districts. Instead, Kenya now has devolved governance through 47 counties) and in Uganda in Amolatar and Dokolo districts in the northern region. Baseline research undertaken as part of the project design confirmed that learning levels were very low in these districts. In the Uganda baseline, for instance, learners in grade 3 scored only 16.8% on a written literacy test whereas those in Kenya had a much

higher score of 48.35%, but even this was below the acceptable standard of performance in such a test (Oketch, Ngware, Mutisya, Ciera, Abuya nd Musyoka, 2009). Scores were appallingly lower in Uganda at 1.63% and 5.08% for grades 1 of 2010 and 2009, respectively, for the two districts (Oketch et al., 2009; Oketch et al. 2014). Given these low baseline results, there was much expectation and even hope pinned on the results of the EAQEL project. Truly, these children were not school ready and whatever effort, be it through EAQEL as was the case here, was welcomed enthusiastically in the communities and the schools. However, EAQEL was not the first attempt to solve the intractable problem of low learning in the region, at least not in Kenya, where RCTs had been a common approach for some time. The next section highlights some of the RCTs that have been attempted in Kenya to address low levels of learning.

2.2 Design and Methods of the EAQEL RCT

EAQEL randomized control trial impact evaluation was designed to answer the following research questions (Oketch et al. 2009; Oketch et al. 2014):

1. Are children in lower primary grades (1, 2 and 3) able to read and do mathematics calculations more proficiently as a result of the Reading to Learn/scaffolding approach?
2. Are there differences in proficiency for children who have been exposed to parental involvement in the Reading to Learn Approach (Core Model Plus) compared to those exposed to the Reading to Learn Approach with no parental involvement (Core Model), and compared to control schools?
3. What are the key contributing factors to these improvements in numeracy and literacy in grades 1, 2 and 3?

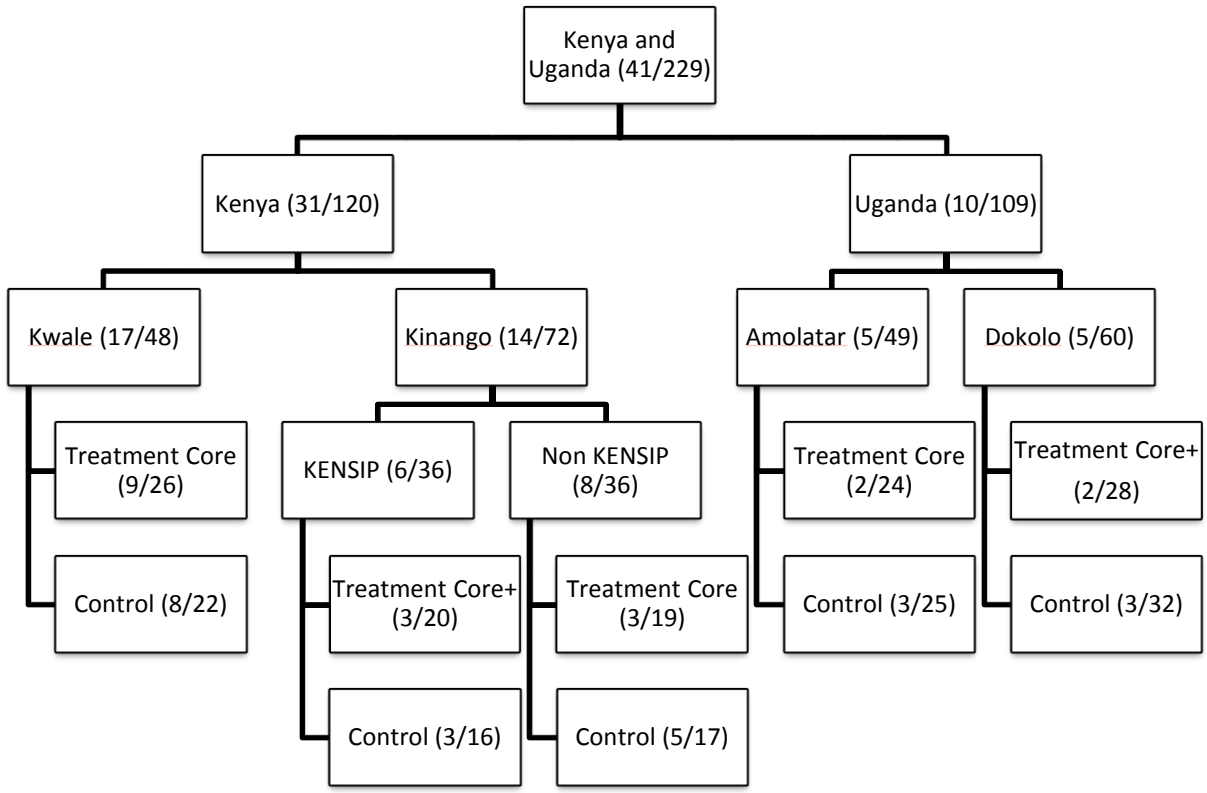
Figure 1 below shows the sampling designed. There were a total of 41 clusters in the study, with 31 in Kenya and 10 in Uganda. Kenya's clusters fall within two districts, Kwale and Kinango. The district of Kinango was further subdivided into clusters that did, or did not participate in the Kenya School Improvement Program (KENSIP) intervention. KENSIP was an earlier intervention undertaken by AKF whose effect needed to be isolated from the effect of EAQEL. Uganda's clusters also fall within two districts, Amolatar and Dokolo. The final randomization occurred within 5 strata (defined by 3 districts, plus one district divided between KENSIP and non KENSIP). Of the 41 clusters, 19 received the treatment (either

Core or Core Plus, depending on the district) and 22 were in the control group. However, one school in Amolatar and one in Dokolo were randomly assigned to a control cluster, but were later selected to be "model treatment schools" by AKF (a classic instance of experimental crossover between treatment and control conditions) (Oketch et al., 2014).

2.2.1 Sampling of Pupils

A random sample of 20 pupils was selected in each grade. The random sampling was done by first grouping pupils by sex and then selecting each sex based on their proportion in the class. Based on the baseline 1 experience, the sample was increased to 25 pupils for the 2010 grade 1 in the baseline 2 in order to allow for any possible attrition due to absenteeism and school transfers. The same pupils were followed at the endline survey that took place between June and July 2011. During the end line survey, pupil absenteeism presented a sample attrition problem. To address the attrition problem at endline, the pupils who were lost to follow-up were randomly replaced taking sex into consideration. This did not pose any methodological threat to the study because the intervention was administered at class level.

Figure 1: Sampling Frame for Kenya and Uganda



Note: The first number in parentheses is the number of AKF clusters in Kenya, or sub-counties in Uganda (i.e., the unit of randomized assignment). The second number in parentheses is the number of schools in all clusters/sub-countries.

2.2.2 Sampling of Parents

A sample of 180 parents was targeted for the focus group discussion (FGD) during the end line. A total of 106 parents turned up for the participation of the actual FGD. The selection of parents was first done by randomly selecting 10% of the schools in the core model plus districts. The selected schools were assigned to be either a male or female FGD. Then, 15 pupils in each of the sampled schools were randomly selected and provided with letters inviting their parents to participate in the FGD. Among the details in the letter to the parents included the venue, time and whether it was the father or the mother who was invited.

In total, 12 FGD's were conducted, 5 in Amolatar (3 treatments and 2 controls) and 7 in Kinango (4 treatments and 3 controls). The FGD's were held separately for men and women, except, in one school in Kenya where both male and female parents participated in the same FGD (Oketch et al., 2014).

2.2.3 Description of the Sample

A total of 229 schools distributed as shown in Table 1 participated in the study.

Table 1: Distribution of schools by district

Control			**Treatment**	
District	No	%	No	%
Kinango	33	45.83	39	54.17
Kwale	22	45.83	26	54.17
Amolatar	24	48.98	25	51.02
Dokolo	31	51.67	29	48.33
Total	110	48.03	119	51.97

2.2.4 Establishing the Baseline and Endline

For RCTs, determining the time period for the baseline and endline are crucial. The *baseline assessment* was carried out in two phases. The first phase was conducted in July and August 2009 and targeted 9,160 pupils in both grades 1 and 2. The second phase was carried out in

February and March 2010 for incoming grade 1 and targeted 5,725 pupils. This was in line with the intervention period and aimed to capture all the three grades (1, 2 and 3) in the impact evaluation. However, in the actual baseline test, the number (14,404) of pupils who were assessed was less than the target (14,885). The reasons for the difference between the target and actual were: 1) some classes had fewer pupils below the target sample size of 20 pupils in 2009 and 25 in 2010; 2) during the testing time, a few pupils disappeared from the test venues and some were also absent during call backs.

During the endline (follow- up), a total of 13944 pupils were captured. This consisted of 9397 pupils followed from the baseline (67.4% of the baseline sample). The rest were new pupils resampled to replace those lost mainly due to absenteeism. Teachers in treatment schools who were captured at the baseline irrespective of the grade they were teaching in 2011 as well as those currently teaching grades 1 to 3 were targeted. In total 445 teachers were interviewed, the distribution of the number of teachers who were interviewed is as shown in Table 3.

Table 2: Distribution of teachers interviewed

District	**Teachers(n)**	**%**
Kinango	135	30.34
Kwale	104	23.37
Amolatar	84	18.88
Dokolo	122	27.42
Total	445	100.00

2.3 Results

2.3.1 Treatment Effects on Numeracy

The impact evaluation results were based on the Difference-in-Difference (DID) analysis which is a straight forward and clear way of assessing the treatment effect of the intervention. DID compares outcome variables at two different assessment points to separate changes associated with intervention or reform. Table 3 shows the pooled DID data for Kenya and Uganda. The results indicate that there was no treatment effect on numeracy. One of the key highlights in Table 4 is that grade 1, 2010 shows a DID of negative 3.92 percentage points.

This shows that schools in the control group performed better in the case of Kenya. In Uganda, grade 1 of 2010 shows a positive DID of 6.45 percentage points in favour of the treatment group. The statistical significance shown in the two cohorts is possibly noise introduced by the low number of clusters and therefore, the pooled data indicating no treatment effect on numeracy in both countries is more reliable.

Table 3: Difference in difference (DID) in the numeracy assessment both countries

Group	**Grade 1, 2010**	**Grade 1, 2009**	**Grade 2, 2009**
Treatment (t_1-t_0)	9.44	18.81	22.37
Control (t_1-t_0)	9.39	17.63	19.70
DID	0.04	1.18	2.68

* p<0.1, ** p<0.05, *** p<0.01, t_1 is Endline, t_0 is Baseline

Table 4: Difference in difference (DID) in the numeracy assessment by country

Country	**Group**	**Grade 1, 2010**	**Grade 1, 2009**	**Grade 2, 2009**
Kenya	Treatment (t_1-t_0)	7.33	16.12	28.16
	Control (t_1-t_0)	11.25	14.75	25.38
	DID	-3.92**	1.37	2.78
Uganda	Treatment (t_1-t_0)	13.08	23.35	12.75
	Control (t_1-t_0)	6.63	22.31	11.50
	DID	6.45**	1.04	1.25

* p<0.1, ** p<0.05, *** p<0.01, t_1 is Endline, t_0 is Baseline

EAQEL did not have any overall effect in improving numeracy when the countries are combined nor in each country separately. This is a true reflection of absence of treatment effect in numeracy achievement as the DID takes in to account any differences that may exists between treatment and control groups at the baseline (Oketch et al. 2014)

2.3.2 Treatment Effects on Oral Literacy

Tables 5 and 6 present pooled and country specific treatment effects respectively of the EAQEL intervention based on the oral literacy scores. In Table 5, the treatment and control row entries show the mean difference between endline and baseline (score increases) for each of the groups, while the DID row presents the percentage point difference in difference between the treatment and control groups (i.e. the treatment effect).

Table5**:** Difference-in-Difference (DID) in the Oral Literacy assessment both countries

Group	**Grade 1, 2010**	**Grade 1, 2009**	**Grade 2, 2009**
Treatment (t_1-t_0)	17.51	21.19	18.51
Control (t_1-t_0)	15.69	19.53	16.02
DID	1.81	1.66	2.49

* p<0.1, ** p<0.05, *** p<0.01, t_1 is Endline, t_0 is Baseline

Table 6: DID between treatment and control in oral literacy by country

Country	***Group***	***Grade 1, 2010***	***Grade 1, 2009***	***Grade 2, 2009***
Kenya	Treatment (t_1- t_0)	18.60	21.88	19.03
	Control (t_1- t_0)	18.69	21.81	18.64
	DID	-0.09	0.07	0.39
Uganda	Treatment (t_1- t_0)	15.64	20.16	17.65
	Control (t_1- t_0)	11.19	15.62	12.20
	DID	4.45**	4.54**	5.45**

p<0.1, ** p<0.05, *** p<0.01; t_1 is Endline, t_0 is Baseline

In Table 6, the DID results in Uganda are positive and significant across all the three cohorts, whereas none is significant in Kenya's case. These results indicate that EAQEL/RtL had a positive treatment effect in Uganda and not in Kenya on oral literacy.

2.3.3 Written Literacy Treatment Effects

Table 7 and 8 respectively show pooled and country specific treatment effects (DID) of the EAQEL intervention based on the Written Literacy. The Written Literacy results show a positive treatment effect in Uganda across the three cohorts, whereas in Kenya, there is none.

Table7: Difference-in-Difference (DID) in the Written Literacy assessment both countries

Group	**Grade 1, 2010**	**Grade 1, 2009**	**Grade 2, 2009**
Treatment (t_1-t_0)	22.15	29.43	27.79
Control (t_1-t_0)	19.74	28.84	24.69
DID	2.40	0.59	3.10

* p<0.1, ** p<0.05, *** p<0.01, t_1 is Endline, t_0 is Baseline

Table 8: DID between treatment and control in written literacy by country

Country	**Group**	**Grade 1, 2010**	**Grade 1, 2009**	**Grade 2, 2009**
Kenya	Treatment (t_1-t_0)	27.29	31.95	27.89
	Control (t_1-t_0)	27.49	34.70	25.69
	DID	-0.20	-2.75	2.21

Uganda	Treatment (t_1-t_0)	12.83	25.48	27.90
	Control (t_1-t_0)	8.10	19.24	23.27
	DID	4.73**	6.24**	4.63**

* p<0.1, ** p<0.05, *** p<0.01

2.4 Interpretation of Findings: The Role of Program Implementation

Table 9 examines whether the treatment effects are sensitive to the degree of program implementation. Three separate coefficients indicate treatment effects among schools with high, medium, or low implementation as shown in the first column. What is noteworthy here is that implementation does not matter for numeracy effects, since all are small and statistically indistinguishable from zero. For both literacy assessments, the full-sample effects are highest among the high implementation category (19-22% of a standard deviation). In contrast, they are zero among the low category of schools.

Table 9: Treatment Effects in High, Medium, and Low Implementing Schools

	Global Mean			Heterogeneous Effects		
	Numeracy	Written Literacy	Oral Literacy	Numeracy	Written Literacy	Oral Literacy
Intention to Treat X High Uptake	0.033	0.191***	0.219***			
Intention to Treat X Medium Uptake	0.012	0.087	0.165**			
Intention to Treat X Low Uptake	-0.064	-0.018	0.036			
Intention to Treat X High Uptake X Kenya				0.029	0.097*	0.165**
Intention to Treat X Medium Uptake X Kenya				-0.099	-0.030	0.020
Intention to Treat X Low Uptake X Kenya				-0.200**	-0.111*	-0.060
Intention to Treat X High Uptake X Uganda				0.013	0.358***	0.300**
Intention to Treat X Medium Uptake X Uganda				0.147	0.233**	0.340***
Intention to Treat X Low Uptake X Uganda				0.138	0.118	0.179*
N	8,920	8,850	8,819	8,920	8,850	8,819

R-Squared	0.28	0.34	0.29	0.28	0.35	0.29

** significant at 10%, ** significant at 5%, *** significant at 1%.*

Table 9 shows results for the sample of students who completed the specified endline test and at least one baseline test. Endline test scores are standardized based on cohort, country, and grade at endline. All regressions include controls for all three baseline tests (students who did not take a particular test are given a score of 0), a dummy variable for each missing baseline test score, interactions between cohort and country, a dummy variable for sex, and district fixed effects separately by whether the school had participated in Kensip for all but one district in each country. Standard errors are clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda). The final columns of Table 9 confirm once again that in Uganda the treatment effects are uniformly positive for all levels of implementation. However, it should also be noted that effects are relatively lower in low-implementing schools. In Kenya, the results are particularly striking because they now show small and somewhat statistically significant treatment effects (10-17% of a standard deviation), but only in high-implementing schools.

In conclusion, the results suggest that implementation quality, as judged by the criteria and observation by implementer, is an important mediator of program effects. It even suggests that, among a subset of Kenyan schools, there were small positive effects (Oketch et at.2009; Oketch et al. 2014). The question that these results raise with regard to RCTs, is of what relevance is this mixed, qualified result to education policy decision making? It generated more questions than answers, and serves to demonstrate that RCTs in education, just as the ones highlighted in the summarised literature reviewed, are not a straight forward "gold" standard answer to the intractable problem of low quality education in low and lower income countries such as Kenya and Uganda. Drawing on the experience of having led a major RCT as presented by the EAQEL study, in the next section, I offer a few reflections.

3 Reflections and Conclusions

Unsurprisingly, learning levels and learning progress among students in many developing countries fall substantially short of those in higher income (e.g. OECD) countries and of those levels required to realise the potential of young people entering the labour market is extensively discussed in the literature. Moreover, low levels of learning progress, linked to poor quality education systems, deny young people and nations the full social and economic

benefits of quality education, for which strong evidence is also found in the literature. In countries where access to education has improved markedly following the Education For All (EFA) movement, such as in Kenya and Uganda in East Africa as shown in the reported EAQEL study above, concern regarding quality is often acute. However, the evidence available to policy-makers and policy analysts is sparse. There is little systematic information about the key drivers of educational quality at the system-level and especially with regard to the likely impacts of reforms and interventions intended to improve education quality and learning outcomes *at scale.* This has contributed to the current talk on the 'global learning crisis', generating pressure on governments to close the so-called "evidence gaps" in an attempt to remedy the crisis by means of better synthesis and integration of available research from across disciplines and paradigms; and through new empirical research employing multiple methods.

At the same time a majority of children in sub-Saharan Africa countries, most of which are low and lower middle income countries, now attend school, but access is still unequal and school quality is low on the measures available. There is also great variation in learning outcomes between countries at similar levels of income and per-pupil expenditure globally and in some cases countries that have higher per pupil expenditure post lower test-score results on measures used than those with lesser per-pupil expenditure. Turning to sub-Saharan Africa, among the SACMEQ countries, low-income Kenya outperforms both middle-income Botswana and South Africa, where per-pupil expenditure is considerably higher (Carnoy, Ngware and Oketch 2015). These two examples illustrate just the tip of the iceberg, that the evidence concerning the impacts of children's backgrounds and of a number of key features of teacher quality on learning outcomes appear consistent and fairly robust across a large volume of 'education production function' (EPF) studies, covering many countries, but much less is known about the macro-level drivers of system performance. Large differences between systems demand much greater attention to the macro-level, and to the inter-relationships between systemic factors and the micro and meso levels of pupil, class, teacher and school that is impossible through RCTs. Further, strong evidence is available for the impacts of individual interventions (such as conditional cash transfers) from randomised control trials, while the generalizability and 'scalability' of such interventions is highly dependent upon 'external conditions', namely context, system and reform-capacity; each of which is much less well understood. RCTs are usually advocated and undertaken by economists and criticised by educators and education planners with economics leaning, yet

insights may be gained from greater engagement and collaboration between economists, evaluators and educationalists and researchers in the field of international and comparative education; as well as through collaboration with 'operational researchers', especially those based in, and working with, ministries in a range of development fields. Insights from such collaboration, especially if it is genuine would suggest that RCTs do not offer robust evidence for addressing education system challenges.

A systems-approach to improving learning centres on the relationships between components of an education-system, including for example, teacher training and deployment, school-management and curricular design, within a comparative framework, including through a rigorous 'case-control' approach to understanding the internal and external conditions which enable 'better system performance'. While it is obvious, in one sense, that the quality of an education system cannot exceed the quality of its teachers, teachers with similar characteristics in one setting may 'produce' outcomes quite different from those in another, as has been shown in several studies of private versus public schooling, even at the within-country level; and in a number of cross-country studies which find a large unexplained 'country-effect' after accounting for differences in pupils' socio-economic background and in teacher quality and classroom conditions (Carnoy et al., 2015; Rolleston, 2014). Further, macro-level factors frequently resist analysis through reduction to simple proxy indicators. Measures such as per-pupil spending and teacher-pupil ratios explain relatively little about the differences between systems. Historical and political factors explain somewhat more. But these are less well understood and less readily quantified, highlighting the key role of analyses drawing on perspectives from political economy, which offer to shed light on the role of institutions and of formal and informal structures and mechanisms of decision-making and policy-implementation within and across contexts.

Moreover, diagnoses of systemic failure, usually an ambition embedded in RCTs, are not in themselves solutions. Research focused on the dynamics of systemic change is required to establish potentially successful reform pathways and to understand the blockages that stand in their way. Understanding the structural changes that accompany successful reforms and describing them comprehensively is considered a key early step towards improving system performance, leading ultimately to the construction of new theories of systemic change in education and the development of mechanisms to galvanise their uptake. This complex

education problem means answers cannot be found through RCTs. Instead, what it requires may be the following.

3.1 System Performance and Diagnostic Tools

Comparative education systems analysis is required to provide better evidence on both the efficiency and effectiveness of education systems. This forms the starting point for establishing summary 'performance indicators' at the system level. Differences in performance depend both on differences in outcomes and inputs, while the key to improving performance within a limited resource envelope is in improving the efficiency and effectiveness with which inputs are employed. The development of 'system metrics' would be an essential step in this. Apparently simple metrics, such as 'system cost-effectiveness' indicators (e.g. dollars per increment in pupil test scores) are currently rarely available, owing to the requirement for measures of 'value-added' in education systems. Measures of performance, efficiency and effectiveness often embedded and dominant in RCT studies do not provide explanations of how and why an education system 'is where it is' or of 'what works' to improve it. The subsequent development of 'system diagnostic tools' for which there is now initiative such as the RISE programme funded by DFID is a crucial second step for understanding the reasons for differences in system performance. These tools should be designed to identify strengths and weaknesses in systems. 'Weak links' in education systems are especially important, owing to the interdependence of components within a system. For example, poor school accountability may explain high levels of teacher absenteeism as well as poor compliance with a range of educational directives and reforms, and indeed the prevalence of 'corruption'.

While such system diagnostics provide a fuller understanding of the sources of good or poor performance at the present, they do not in themselves provide a 'way forward' with respect to specific reforms needed (or likely to be effective). For example, while poor school accountability may be a proximal cause of low learning progress, this leaves open the question of *how to improve* accountability. RCT based studies may offer some insight at small-scale on potential mechanisms of change, but where the 'blockage' lies at the macro-level, being related, for example to industrial relations via teacher unions, there is a danger of over-simplification or reduction when using experimental evidence such as RCT. With respect to macro-level questions, for example of curricular reform, alternative solutions such as decentralising curricular decisions or indeed centralising them as is often times the result

and recommendation from RCT studies could equally well improve or worsen system performance, depending on the institutional and political-economic context, so that understanding 'reform pathways' is a linked but separate research endeavour from understanding system performance and establishing appropriate diagnostic tools , both of which are not possible to answer through RCTs.

3.2 Reform Pathways, Blockages and Catalysts for Reform

The effectiveness of education reforms in respect of individual dimensions of the education system, such as curricula or teacher training, is often limited to a considerable extent by the 'next weakest link' in the system as considered above. Improving text-books may yield improvements in learning, but these improvements will depend upon teachers' knowledge and training being adequate to employ the new books effectively and on regular assessment of pupils' learning feeding back into teaching and learning. Many of these links are *learning opportunity processes*, rather than more readily measurable simple inputs, which require more complex indicators. 'Reform pathways' are more than mechanisms for change of individual features of a system (e.g. teacher absenteeism and its effect on learning as is usually the nature of questions addressed through RCT) and reflect the full chain of linkages required for sustainable system reform. Reform pathways describe routes from the present status quo to improved system performance based on a holistic system-oriented approach, which results from a thorough diagnosis of weaknesses and strengths plus a full understanding of the interdependence between mechanisms of change. The identification of reform pathways should begin with a 'situational analysis' of the status quo - an understanding of the reasons why the reforms which may be considered necessary have not been undertaken or have not succeeded to date (Oketch and Rollestone, 2007) something not answered by RCT because these questions relate largely to the political economy of individual education systems. What would be most useful are research designs that provide syntheses of the evidence across contexts and countries to enhance understanding of the nature of decision-making and implementation processes and their influences, providing a framework for understanding political economies of education and their linkages to both educational quality and learning outcomes. In short, a focus on questions relating to the conditions and circumstances under which effective education reforms are undertaken, and how these can be influenced can hardly be sufficiently answered through RCTs.

3.3 Policy Influence

Providing evidence for informed policy change in education requires technical analyses of system performance, diagnosis of priority areas for development and an understanding of the political economy of systemic change, so as to link potential reform mechanisms to the systems in which they are most likely to be effective. Further, the initiation of reform-oriented policies is dependent on the demand for such policies within states and societies. Accordingly, attention should be paid to understanding and improving the demand for reform and for the evidence required for evidence-based policy-making, usually a task that is not properly undertaken in RCTs which are driven by experts on this method. Moreover, better understanding of the demand for evidence and for evidence-based reform will provide further insights into why some countries and societies are apparently 'better than others' at driving change and at 'producing' learning progress through school systems. In addition, research on the political economy of evidence and policy change will enable a fuller understanding of the reasons for 'disconnects' in education policy (and relevant policies in other sectors) and how to remedy them. A systemic perspective is needed that draws together the evidence on the strengths and weaknesses of earlier approaches to research and policy influence, which have often focused very heavily on issue-specific areas such as early-grade reading (EGRA) in education or malaria in health, with less attention paid to the systemic impacts of approaches and limited integration within and between sectors. Moreover, lessons from successful programmes in one area are not systematically 'read across' to areas receiving less attention and a systems-approach should serve to develop this through a consideration of the characteristics of successful programmes and reforms and an analysis of their uptake and institutionalisation, all of which do not require RCTs.

3.4 The Need for Meaningful Empirical Studies

The proposed key areas for further empirical research include the following:

(i) Systematic and rigorous study of the 'natural variation' in education systems and system-performance, drawing on detailed case-studies in the focus countries and employing a mix of quantitative and qualitative methods, supplemented by quantitative analysis of available data for a larger selection of countries, would be much better than RCTs.

b) Rigorous evaluations of systems-issues within sub-systems of countries (e.g. at district level) by employing systematic methods excluding randomisation methods.

c) Research to better understand the nature of 'effective bureaucratic systems'. It is apparent that many of the most successful systems did not achieve their current success through the use of randomised trials but through systematic improvement enabled through bureaucratic systems. These systems and processes should be examined in detail employing a mixed-methods approach to learning from the most and least successful educational bureaucracies.

(d) Related to the above, effective systems typically make good use of measurement and evidence, but are also characterised by high levels of accountability and capacity for learning from experience in an evolutionary manner to 'ratchet up' performance across the sector.

ACKNOWLEDGEMENTS
The author thank Caine Rolleston and Luis A Crouch for conversations and input and two anonymous referees and Gita Khamsi for their comments. The chapter and its conclusions, however, are the sole responsibility of the author.

References

Abuya, B. A., Oketch, M., Ngware, W.M., Mutisya, M. and Musyoka, P.K. (2015) Experiences of parents with the Reading to Learn approach: a randomised control trial initiative to improve literacy and numeracy in Kenya and Uganda, *Education 3-13*, 43:5

Banerjee, A., and Duflo, E. (2011). Poor Economics: A radical rethinking of the way to fight global poverty. New York, NY: Public Affairs.

Carnoy, M., Ngware, M. and Oketch, M. (2015). The role of classroom resources and national education context in student learning gains: Comparing Botswana, Kenya, and South Africa, *Comparative Education Review*, Vol. 59, no. 2. P. 199-233 (Featured Article).

Glewwe, P., Ilias, N., & Kremer, M. (2003). Teacher incentives. Working Paper 9671, National Bureau of Economic Research. Accessed 17/11/2009 at: http://www.nber.org/papers/w9671

Glewwe, P., Kremer, M., & Moulin, S. (2007). Many Children Left Behind? Textbooks and test scores in Kenya. Accessed 19/11/2009 at: http://www.economics.harvard.edu/faculty/kremer/files/kntxtb18_2007July10.pdf

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2000). Retrospective Vs. prospective analyses of school inputs. The case of flip charts in Kenya. Working Paper 8018, National Bureau of Economic Research. Accessed 19/11/2009 at: http://www.nber.org/papers/w8018.

Kremer, M., Miguel, E., Thornton, R., & Ozier, O. (2005). Incentives to Learn, World Bank Policy Research Working Paper 3546. Accessed 20/11/2009 at: http://econ.worldbank.org

Lucas, A.M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda. *Journal of Policy Analysis and Management*, Vol. 33, p. 950-974.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impact on education and health on presence of treatment externalities. *Econometrica*, 72, 159-217

Ngware, W.M, Abuya, B.,Oketch, M., Admassu, K.,Mutisya, M and Musyoka, P. (2015) Randomized impact evaluation of education interventions: experiences and lessons from a reading to learn intervention in East Africa, *International Journal of Research & Method in Education*, 38:4, 430-451

Oketch, M., Ngware, M., Mutisya, M., Ciera, J., Abuya, B., & Musyoka, P. (2009) East African Quality in Early Learning (EAQEL) Baseline Findings Report. African Population and Health Research Center (APHRC), Nairobi.

Oketch, M., Ngware, M., Mutisya, M., Admassu, K., Abuya, B., Musyoka, P. (2014). When to Randomize: Lessons from Independent Impact Evaluation of Reading to Learn (RTL) Programme to Improve Literacy and Numeracy in Kenya and Uganda *Peabody Journal of Education*, Vol 89, p. 17-42.

Oketch M and Somerset, A. (2010). *Free Primary Education and After in Kenya: Enrolment impact, quality effects, and the transition to secondary school*, CREATE PATHWAYS TO ACCESS, Research Monograph No. 37. ISBN: 0-901881-44-9.Accesssed at : http://www.create-rpc.org/pdf_documents/PTA37.pdf p.1-31

Oketch, M. O. and Rolleston, R. (2007). Policies on Free Primary and Secondary Education in East Africa: Retrospect and Prospect, *Review of Research in Education*, Vol. 31 (1), p.131-158.

Rolleston, C. (2014) 'Learning Profiles and the 'Skills Gap': A Comparative Analysis of Schooling and Skills Development in Four Developing Countries' *Oxford Review of Education*, 40(1)

Rolleston, C. and Krutikova, S. (2014) Equalising Opportunity? School Quality and Home Disadvantage in Vietnam *Oxford Review of Education*, 40(1)

Romero, M., Sandefur, J. and Sandholtz, A.W. (2017). Can Outsourcing Improve Liberia's Schools? Preliminary results from Year One of a Three-Year Randomized Evaluation of Partnership Schools for Liberia. Working Paer 462, September 2017. Centre fr Global Development.

Unesco (2014). Teaching and learning: Achieving quality for all. Paris. Unesco.

Vermeersch, C. & Kremer, M. (2004). School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation. Accessed 22/11/2009

at: http://www.povertyactionlab.org/sites/default/files/publications/100_Kremer_School_Competition.pdf