RISE
RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

RISE Working Paper 19/034
November 2019

Evaluating Large-Scale Education Reforms in Ethiopia

John Hoddinott, Padmini Iyer, Ricardo Sabates, Tassew Woldehanna

Abstract

Understanding whether wide-ranging education reforms effectively improve learning outcomes is of crucial importance in many low and middle-income countries where, in spite of gains in school enrolments, learning outcomes remain poor. Ethiopia is one country experiencing the global 'learning crisis', and since 2008 the Government of Ethiopia has embarked on an ambitious package of reforms – the General Education Quality Improvement Program (GEQIP) – in order to improve the quality of basic education in the country. In this paper, we consider a suitable quantitative strategy for investigating the impact of large-scale, complex education reforms, with particular reference to the latest phase of GEQIP in Ethiopia. Quantitatively, randomised control trials (RCTs) have been considered the 'gold standard' to evaluate interventions; however, in the field of education, many scholars have acknowledged the limitations of RCTs particularly with respect to their external and construct validity. In light of these and other concerns, we outline our approach to assessing the impact of GEQIP-E on learning outcomes – a longitudinal design which incorporates variations across time and space that we are likely to observe over the course of reform implementation. This design allows us to understand both the impact and, importantly, the processes of implementation, which is essential if we are to understand not only whether but why certain elements of large-scale reforms may – or may not – lead to improved learning outcomes.

Keywords: Randomised Control Trials, Sampling, Systems Research, Longitudinal Research; Ethiopia

UKaid from the British people
Australian Aid
Bill & Melinda Gates foundation
Oxford Policy Management
Blavatnik School of Government
University of Oxford

Evaluating Large-Scale Education Reforms in Ethiopia

John Hoddinott
Cornell University

Padmini Iyer
NatCen Social Research

Ricardo Sabates
REAL Centre, University of Cambridge

Tassew Woldehanna
Addis Ababa University

Acknowledgements:

We are grateful to Professor Lant Pritchett for his insight and comments on an earlier version of this paper.

This is one of a series of working papers from "RISE"—the large-scale education systems research programme supported by funding from the United Kingdom's Department for International Development (DFID), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:
Hoddinott, J., Iyer, P., Sabates, R. and Woldehanna, T. (2019). Evaluating Large-Scale Education Reforms in Ethiopia. RISE Working Paper Series. 19/034.

Introduction

Improving students' learning outcomes is central to many education reforms in both developing and developed countries. In Ethiopia, emphasis on this began in 2008 under the General Education Quality Improvement Programme (GEQIP). GEQIP was introduced to improve teaching and learning conditions in both cycles of primary education (Grades 1-4 and Grades 5-8), and to strengthen education institutions and service delivery at federal and regional levels of government (World Bank, 2008). The first phase, GEQIP-I, was implemented between 2008 and 2013 and focused on providing essential inputs for improving teaching and learning. The second phase, GEQIP-II (2013-2018), continued the resource provision and improvement activities in GEQIP-I and added support for information and technology. The third phase, GEQIP-E (2018-2022), continues many of the prior reforms but adds a stronger focus on equity, aiming to address education-related challenges facing girls, children from pastoralist communities and children with special needs. Given that over $400 million is being spent on GEQIP-E (World Bank 2018) and the importance of improving of human capital formation including school-related outcomes in Ethiopia, understanding the impact of the current GEQIP package on learning outcomes represents an important research and policy question. RISE (Research on Improving Systems of Education) Ethiopia, a five-year, DFID-funded programme of research, aims to assess the impact of GEQIP-E on equitable learning in Ethiopia.

While RISE Ethiopia as a whole is adopting an interdisciplinary, mixed-methods approach to examining the GEQIP-E reforms, in this paper we focus on the quantitative component of our impact and process evaluation, with specific reference to the sources of variation to estimate impact but importantly on the potential variations between implementation inputs which may explain variations in learning. We begin with a brief description of the GEQIP-E reforms and the research questions we seek to address. We then consider some of the challenges of a RCT design to study our research questions; these challenges motivate our choice of an alternative strategy which aligns more broadly with our ultimate goal, which is to understand why GEQIP-E reform is or is not working to improve learning outcomes in Ethiopia.

Systems Research: The case of GEQIP-E

GEQIP-E

GEQIP-E has four principal results areas – improving internal efficiency, equitable access, quality and system strengthening in terms of planning and policy formation – which collectively all aim to improve learning outcomes. A bundle of interventions will be implemented; nationally, school grants, performance-based awards, textbooks, and support for children with special needs will be provided to all government primary schools. In regions where support to education has historically been limited, the 'emerging regions' of Afar, Ethiopia Somali and Benishangul-Gumuz, equity-enhancing interventions focusing on gender, children with disabilities and children of pastoralist groups will be implemented. Finally, a phased

approached to enhance early grade provision (O-Class), mother tongue instruction in early grades of primary school (i.e. grades 1 and 2) and support to improve teaching of literacy and numeracy in the upper grades of primary school (i.e. grades 7 and 8) will be rolled out: in 5% of woredas (districts) in academic year 2018-19, 25% of woredas in 2019-20 and 50% of woredas in the 2020-21 academic year.

Table 1 shows the assumptions made under each of the components of GEQIP-E reform in order to improve efficiency, equity, quality and strengthen the overall system. To improve efficiency, for instance, O-Class component aims to increase access to pre-school age children and also to reduce repetition in Grade 1. Equity is achieved by gender-sensitive interventions, particularly in geographical areas where female enrolment is below male enrolment. Quality is assumed to be achieved with several interventions, including in-service teacher training, increasing the quality of candidates admitted to teacher training, making school grants available on time, and improving the quality of teaching in the early years. All these interventions are geared at raising learning outcomes in the primary school cycle, from pre-school to the last grade of primary school (Grade 8 in the Ethiopian education system).

Table 1. Examples of Assumptions of GEQIP-E

Efficiency	• **O-Class**
	  o Inputs to O-Class will lead to greater internal efficiency and remove bottlenecks in the system.
	  ▪ Increasing student attendance will help to reduce inefficiency
	  ▪ Addressing the issue of repetition in Grade 1 will help to reduce inefficiency
	• **PfR**
	  • Performance-based awards will incentivise schools to increase internal efficiency (i.e. G2/G1 enrolment ratio and G5 survival rate).

Equity	• **School Improvement Programme/School Grant**
	  • Gender sensitive school improvement plans in Afar, Somali and Ben. Gumuz will lead to increased female enrolment in these regions and in turn increased GPI in G8.
	  ▪ Separate latrines
	  ▪ Life skills training
	  ▪ Girls' clubs
	  • Improved availability of basic school grant and additional school grant will lead to:
	  ▪ Better quality/increased equity in the emerging regions.
	  ▪ Better provision of education for students with special needs.
	• **Community Involvement**
	  • Addressing socio-cultural gender practices will help to increase girls' enrolment in emerging regions (by addressing issues of marriage and FGM).
	  • O class will help to increase inclusive education in schools.

Quality	• **Teachers**
	  o Improved instructional activities for teachers will lead to increased teacher quality and better student outcomes in Phase 1 schools.
	  o Improving teacher training programmes will improve teacher quality and improve learning outcomes.
	  o Increasing the quality of candidates admitted to teacher training will increase teacher quality.
	• **School Improvement Programme/School Grants**
	  o Timely availability of school grants and textbooks will lead to greater education quality and better student outcomes in Phase 1 schools.

	○ Evidence-based SIPs will lead to greater education quality and better student outcomes in Phase 1 schools. ○ Improving the quality of the SIP and aligning it with inspection guidelines will help to improve the learning environment. • **O Class** • Inputs to O-Class will increase the quality of pre-primary education.
System Strengthening	• **Performance for Results** ○ Shifting to a results-based approach will help to address the remaining challenges in the education system. ○ Improved availability, quality and use of data will strengthen the education system. • **Teachers** ○ Improved pre- and in-service teaching will strengthen the education system. • **Technical Assistance** • TA assistance provided by the World Bank will help to strengthen the education system.

Note: The World Bank is specifically supporting mother tongue education in grades 1 & 2 as well as enhancing teaching quality of literacy and numeracy in grades 7 & 8.

The Research Questions

Undertaking a systems approach to our research requires contextualisation of how the GEQIP reforms are embedded within Ethiopia's education system (Andrews, Pritchett & Woolcock, 2017). In other words, it is essential to understand how actors make sense of, and respond to the incentives, opportunities and barriers they encounter. Identifying the fidelity of implementation of GEQIP components from the federal through to the local level, as well as the pathways that enable or hinder implementation, also provides the information that is required for the formulation of research instruments geared at examining the effectiveness of the GEQIP-E reform. According to Andrews, Pritchett and Woolcock (2017) fidelity of implementation can sometimes be more important than policy or programme design in explaining learning outcomes and therefore key aspects of the implementation need to be explicitly captured to explain the impact of such a reform or intervention.

With these issues in mind, two key research questions for RISE Ethiopia are as follows:

1. What is the impact of the GEQIP-E bundle of interventions on learning outcomes in primary school?
2. Why, how, and for whom is the GEQIP-E reform working?

An RCT for GEQIP-E?

In both economics and in education, a popular approach to assess the impact of interventions aimed at improving learning outcomes is a randomised control trial (RCT). The great strength of an RCT design is that, provided certain conditions are met, they generate unbiased estimates of the impact of the intervention being studied (Imbens, 2010; Gertler, et al. 2016). But an unbiased estimate by itself does not explain why a given policy, program or intervention studied by an RCT works or does not work. The careful systematic

review undertaken by Glewwe and Muralidharan (2015) illustrates this concern.[1] Glewwe and Muralidharan review of the evidence emerging from 118 high quality RCTs focused not just on synthesising but interpreting the evidence for generalisable lessons. Consider their discussion of four RCTs on the impact of providing books and materials to students. These found zero effects on learning outcomes and had four different explanations for this result: (1) books were stored and not used by students, (2) books are provided to students but households offset the intervention by reducing their support at home; (3) books are too advanced for students; (4) teachers' were not using the books properly. Each of these potential explanations have key policy and implementation lessons for how books should be produced, delivered, utilised, and enhanced via use at home in order to improve learning outcomes, but it is not possible from these RCTs alone to determine which explanation(s) are most plausible. More generally, there is also a challenge with evidence emerging from RCTs with respect to the large variations on the causal estimates of interventions, which arise from the large differences on what is been captured by the causal evidence of these interventions. For example, improvements in learning outcomes from interventions to enhance ICT in schools may be due to the provision of equipment, but also through teacher training or better connectivity (potentially all of these plus other conditions are necessary!). In turn, these considerations lead to concerns regarding external validity of RCTs (Carthwright, 2012, 2016; Deaton, 2010; Deaton and Cartwright, 2018; and Ravallion, 2018), even external validity within the same country (Bold, et al., 2013).

Systematic reviews of RCTs might seem to be one way of addressing some of these concerns. But a review of systematic reviews by Evans and Popova (2015) points to additional concerns with the evidence emerging from RCTs in education. First, systematic reviews starting from the same potential studies and following strict methods for selection of robust causal evidence are not included in all systematic reviews in a systematic way. In other words, of the 227 studies with RCTs that look at learning outcomes only three are included in all six systematic reviews analysed by Evans and Popova. Since each of the systematic reviews include different evidence, Evans and Popova highlight the second concern which is that these systematic reviews are leading to different conclusions. Hence, it is unsurprisingly that a 2016 systematic review by Snilstveit, et al, focused almost exclusively on what has been learned in educational research from 216 programmes, all evaluated using RCTs, the authors concluded: "there are no magic bullets to ensure high quality education for all" (Snilstveit, et al, 2016, page 1).

Estimates of causal impacts from RCTs usually focus on a single treatment arm. While multi-treatment arms are possible, there are limits to how they can account for the complex interactions that are required to improve learning. RCTs which find that one aspect of an intervention is effective cannot generalise that this is the only arm of the intervention that works for improving learning. For instance, Snilstveit, et al (2016) provide evidence from 37 RCTs assessing the effects of programmes implemented to enhance or

[1] Other systematic reviews focusing on education include McEwan, 2012, Kremer, Branner and Glennerster, 2013; Glewwe et al., 2014; Glewwe and Muralidharan, 2015; Snilstveit, et al, 2016.

change the governance or financing of education, a more systems-oriented approach. But these were evaluated exclusively on the simple one arm treatment and control and provide no explanation as to why some reforms worked and others did not. Hence, evidence from RCTs has also been limited with respect to unpacking the potential reasons why educational interventions work or not (Pritchett, 2017). Pritchett (2017) emphasises that even when the empirical evidence emerging in education from RCTs are based on sound principles for estimating unbiased causal estimates, they are too inadequately theorised to be able to present effective ideas for accelerating learning outcomes. Pritchett highlighted "more of the same empirical research is unlikely to be of much help or add up to a coherent action or research agenda as it faces massive challenges of external and construct validity" (page 1).

These challenges - the complexity of the causal process and the many intervening factors outside of the intervention which could influence learning; the exact nature of the intervention which is evaluated; how different are the beneficiaries and whether they could all respond in a similar way to such an intervention; and the likelihood of correlates induced by the reform coming into effect after randomisation - are not unique to RCTs, they speak to two key arguments that need to be considered in the context of efforts to measure the impact of the GEQIP-E reforms: the complexity of the reform as designed, and an inability to ensure nationwide implementation fidelity in a resource-constrained system.

Table 1 above outlines the key components of the GEQIP-E reform. Under a well-designed RCT, each component and each combination of components would be identified as treatment arms of the intervention, and a control or 'business as usual' group, in which no treatment arm is implemented would be identified. Ethiopia is divided into regions, which are sub-divided into woredas (districts); in the case of education, woredas are the lowest administrative level and so the relevant unit of randomization. There are at least 10 individual treatment arms of the GEQIP-E reform identified in Table 1. To determine which components of the reform most effectively improve system efficiency, an RCT design would randomly allocate schools within woredas where the only intervention is O-Class, schools within woredas where the only intervention is performance-based awards, and so on. Additionally, an RCT would need to identify schools where combinations of these different components are introduced. Put simply, there are not enough woredas in Ethiopia for the number of possible treatment arms under GEQIP-E – as currently designed, the reforms are too complex for an RCT research design.

More significantly, even within Table 1, one has to clearly specify the nature of the intervention. One of the particular aspects of GEQIP-E is to improve teacher training (as indicated in Table 1). Popova, Evans and Arancibia (2016) reviewed evidence on the exact nature of interventions used in education to enhance learning via in-service teacher training (all of which have been evaluated using RCTs). Their first observation of the authors is that not all RCTs provided enough details into the nature of the intervention. In other words, it was unknown who provided the training, who was learning, how the training was

delivered, for how long, in which location, among other factors. Once this information was gathered from the original interventions, Popova, Evans and Arancibia found 51 possible indicators which capture the dimensions in which programmes may differ. For the case of GEQIP-E, suppose we could randomise exposure to teacher training and obtained an unbiased estimate of the impact. The questions remain as to what exactly is this estimate capturing (there are 51 different indicators, not to include their potential combinations) and how would this estimate help to advance knowledge of what specifically to do in Ethiopia to improve learning.[2]

An alternative way to proceed could involve selecting a certain number of woredas for GEQIP-E implementation, and a certain number of 'control' woredas where GEQIP-E is not implemented, and where schools continue to operate according to a 'business-as-usual' model. The challenge here is that, as mentioned above, GEQIP-E is the third phase of the overall GEQIP reforms; this means that 'control' schools would effectively be 'GEQIP II' schools, which in turn would require the education system to effectively implement two different sets of reforms simultaneously. Such an approach would evidently be a significant strain on a low-resource, low-capacity system. Avoiding partial implementation or lack of blinding to the treatment (Deaton & Cartwright, 2018) would therefore be a significant challenge, particularly on a national scale,which in turn would affect the validity of RCT results.

Another alternative would be to restrict the impact analysis to a single component of the GEQIP-E reform (or a small number of its components) in a single region. This would allow tighter control over the intervention and cleaner identification of impact. But such an approach has its own limitations. There is no guarantee that an intervention in one part of the country would work in another where circumstances (administrative capacity, wealth, infrastructure, language of instruction) differ. It would not tell us how well this component would work (and why it might/might not work) under less controlled conditions, nor what impediments might arise if it were implemented widely. Finally, such an approach would not tell us whether other components of the GEQIP-E reforms complement or detract from the effectiveness of this component. A systems approach to research requires us to not only understand whether specific reforms or components have impact, but crucially to also understand the processes that lead to this impact.

The key point here is that the quantitative research design selected should not just be constrained to estimating causal relations which could lead to generic conclusions, but importantly understanding the reasons behind these results using a systems framework (see Pritchett 2015 for the RISE system framework). For instance, Glewwe, Kremer and Moulin (2007) experiment of providing additional textbooks in Kenya found that there are no effects, however, further analysis found that effects were obtained for children with initial advanced learning skills, i.e. those who were able to benefit from understanding the material of the textbook. Pritchett and Beatty (2012) took this argument forward to

[2] We thank Lant Pritchett for pointing this out.

demonstrate that the overambitious curricula would lead to significantly different results of pedagogical interventions. For children in earlier grades of primary where the curricula is less ambitious and children's learning more aligned to the curricula, pedagogical interventions are more effective, whereas for children in higher grades of primary, who are lagging behind what is taught, pedagogical interventions are less effective or ineffective. The RISE programme, with a common framework for understanding results, is useful to further explain or articular this issue via delegation (what is asked from the teachers by the system). In proposing our design, we keep these issues in mind.

An alternative design to evaluate the GEQIP-E Reform

Given the considerations described above, an alternative design to evaluate GEQIP-E is needed. Here we describe a longitudinal study which aims to track the learning progress of children in selected primary schools over time. For this selection of schools, simple random sampling is unlikely to overcome the challenge of estimating associations between learning outcomes and different indicators of reform implementation. It is therefore important to capture different sources of variation which could enable us to estimate changes in learning outcomes – specifically in maths and literacy – across time and space in order to provide robust estimates of the impact of the reforms. Additionally, we use a number of research instruments – including school, teacher and household-level questionnaires – to measure the process of implementation and therefore the factors associated with the possible impact of the reform. We based our evidence on the common RISE Framework and on the knowledge gathered via our research on the political economy of Ethiopia which helps to address more deeply issues of reform implementation and nature of the interventions (Iyer and Rossiter, 2018). This allows us to conduct both a process and an impact evaluation of GEQIP-E.

Sample design

Our sample design needs to reflect three features of education reforms in Ethiopia: the fact that GEQIP-E builds on earlier reforms; that some components of GEQIP-E will be rolled out in a phased manner; and finally, the importance of including regions where equity concerns are especially salient.

We exploit an earlier study to incorporate the fact that GEQIP-E builds on earlier reforms. "Young Lives" is a longitudinal study of childhood poverty which has followed a total of 3,000 children in Ethiopia, divided into two age cohorts over the course of 15 years since its inception in 2001; an 'Older Cohort' born in 1994-95, and a 'Younger Cohort' born in 2001-02. In addition to five rounds of household surveys (conducted every 3-4 years from 2002 – 2016) focused on these children, Young Lives conducted two school surveys in Ethiopia: a lower primary education in 2012-13 with Grade 4 and 5 students and an upper primary education in 2016-17 with Grade 7 and 8 students. These surveys included value-added measures of learning in maths and literacy (in the regional language of instruction) in 2012-13, and in maths, English

and Amharic in 2016-17.[3] Crucially, data collection for the 2012-13 school survey was conducted at the end of the GEQIP-I reforms, and just before the GEQIP-II reforms were implemented. By including schools from the Young Lives lower primary school survey, we are therefore able to use information from the 2012-13 academic year as baseline information for children in Grade 4, and to capture changes over time by collecting new information during the 2018-19 academic year (the first year of GEQIP-E implementation) for children in the same grade. Additionally, since there is an overlap of 61 schools included in both the 2012-13 and 2016-17 surveys, Young Lives allows us to capture change over time as children moved between Grades 4-5 and 7-8. In RISE Ethiopia, we include 38 of these Young Lives schools in our study, thus making it possible to evaluate school effectiveness at four points in time (facilitated through the use of linked cognitive tests): at the end of GEQIP-I implementation; in the penultimate year of GEQIP-II implementation; in the first year of GEQIP-E implementation; and in the final year of GEQIP-E implementation.

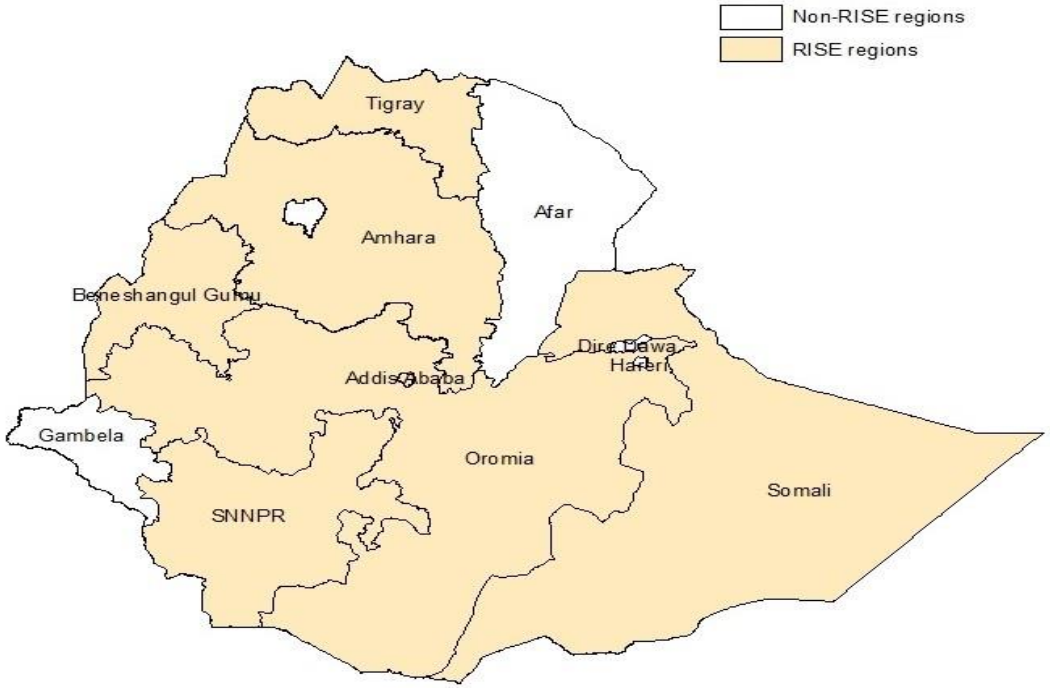
Next, we note that with support from the World Bank under its Payment for Results (P4R) Framework, GEQIP-E will enhance early grade provision and teacher training in primary schools using a phased approach, starting in 2000 schools. Therefore, our sample design needed to cognizant of the inclusion criteria used to select these schools. These included the following: they had to be located in woredas where there at least 10 primary schools; they had to offer O-class as well as all primary school grades (ie to grades 7 and 8); and they had to have participated in baseline testing using the EGRA grade 2 assessment tool which meant that they had to be situated in localities where the test instrument had been adapted to local mother tongues.

Finally, given that we are adopting a systems approach to examining complex, nationwide reforms, our sample should capture the contextual diversity within which the GEQIP-E reforms are being implemented. Moreover, regional variation is a non-negotiable requirement set by the Government of Ethiopia for our impact evaluation. Accordingly, we purposively select the following regions: Addis Ababa; Amhara; Benishangul Gumuz, Oromia; SNNP; Somali; and Tigray (see Figure 1)[4]. The selection of these woredas is partially guided by the Young Lives sample as described above, while we replace Afar with Benishangul Gumuz so as to include one emerging region in the east of the country (Somali), and one emerging region from the west (Benishangul Gumuz). We also aim to capture socio-economic variation in our sample, and therefore aim to select randomly students within schools within regions. Finally, we only include schools offering Grades 1-8, as we are aiming to assess the impact of GEQIP-E across the full primary cycle.

[3] See Aurino et al 2014 and Rossiter et al 2017 respectively for more details on survey and cognitive test design.

[4] Ethiopia is divided into nine regional states – Afar, Amhara,Benishangul Gumuz, Gambela Harrari, Oromia, SNNP, Somali and Tigray – and two city administrations – Addis Ababa and Dire Dawa. These regions and city adminstrations have their own characteristics. Some are dominated by urban populations (Dire Dawa, Harari and Addis Ababa); 'emerging regions' are less economically developed (Afar, Benishangul Gumuz, Gambela and Somali), with the emerging regions in eastern Ethiopia dominated by pastoralist populations (Afar and Somali). The remaining 'core' regions, while relatively developed, are dominated by rural populations (Amhara, Oromia, Tigray and SNNP).

Figure 1: Ethiopian RISE study regions



Balancing these considerations with the resources available for our study, we construct our sample as follows. We aim for a sample of 168 schools. We include a minimum number of schools for the four least populous regions: Addis Ababa; Benishangul Gumuz; Somali; and Tigray. We set this at 20 schools for each of these regions. This leaves us with 88 schools for the larger regions (168 – (20 x4) = 88). If schools are allocated approximately proportionate to regional size, we would include 25 schools in Amhara, 41 in Oromia and 22 in SNNP. Given that, across the seven regions in which we are working, 20% of the population lives in urban areas, 20% of our schools should be located in urban areas. Applying the urban/rural population shares across regions in Ethiopia, Table 2 shows the proposed distribution of schools for the research.

Table 2: Distribution of schools by region, and localities and school types

RISE regions	Total to be surveyed	Locality		School type		
		Rural schools	Urban schools	GEQIP E Phase One schools	Young Lives schools	Additional, randomly selected schools
Addis Ababa	20	0	20	0	7	13
Amhara	25	21	4	9	6	10
Benishangul Gumuz	20	16	4	0	0	20
Oromia	41	35	6	14	6	21
SNNP	22	18	4	8	7	7
Somali	20	17	3	0	5	15
Tigray	20	15	5	7	7	6
TOTAL	168	122	46	38	38	92

In selecting our school sample, it is important to note several additional complications. Woredas are of different sizes, and therefore have different numbers of schools. If we draw a simple random sample of woredas, we artificially increase the likelihood of selecting schools from less populous woredas and decrease the likelihood of selecting schools from more populous woredas. Additionally, there are logistical constraints to be considered; conducting data collection in 168 different woredas (even with 46 woredas located in urban areas) would have significant cost implications. Accordingly, we do the following. First, we populate our sample with schools from the Young Lives sample (specifically, from the sample of 55 schools in both Young Lives school surveys and which are found in the RISE survey regions) in order to ensure the time variation for identification is captured. Next, we add to the sample schools that are included in 'Phase One' of the phased component of GEQIP-E, to allow for the potential geographical/spatial variation in the identification. We do so in proportion to their allocation within regions. So, for example in Tigray, 25% of rural woredas have been selected for the phased approach of GEQIP-E. We therefore include a sufficient number of Phase One schools in Tigray so that 25% of our rural schools are phase one schools. We do so using PPS sampling where proportions are based on woreda populations. We fill out the remaining schools with a random selection from additional woredas.

At the core of our identification strategy is the collection of longitudinal data from two cohorts of children (Cohort A and Cohort B) at four points in time: at the beginning and the end of the 2018-2019 school year (surveys R1 – B and R1 – E); and at the beginning and the end of the 2021-2022 school year (R2-B and R2-E) (see Table 3). This design, replicated from the two Young Lives school surveys, allows us to evaluate school effectiveness using a 'value-added' approach, and to compare school effectiveness over time. At the time of baseline data collection (R1 –B), Cohort A are at the beginning of Grade 1 and Cohort B are at the beginning of Grade 4. Data obtained from Cohort B children attending Young Lives schools will be comparable to data obtained from children surveyed during the 2012-13 Young Lives school survey. We will also be able to compare children from both Cohorts A and B attending GEQIP Phase One schools

with children attending the Young Lives schools, and with those attending schools randomly selected for our study.

Table 3: Survey dates and cohorts

Survey date	Survey name	Cohort A	Cohort B
September- October 2018	R1 – B	Beginning Grade 1	Beginning Grade 4
May-June 2019	R1 – E	End Grade 1	End Grade 4
September- October 2021	R2 – B	Beginning Grade 4	Beginning Grade 7
May-June 2022	R2 – E	End Grade 4	End Grade 7

Over the study period, Cohort A will be exposed to the bundled set of GEQIP-E interventions from Grades 1 – 4. Cohort B will not be exposed to these reforms prior to their attendance in Grade 4, but will be exposed to these interventions from Grades 4 – 7. Children who participated in the Young Lives school surveys between 2012-13 and 2016-17 will not have been exposed to the GEQIP-E reforms. Additionally, all children attending Phase One schools will be exposed to an additional element of the reforms – enhanced teacher training – to which children attending other schools will not be exposed. The identification of sources of variation and our sampling strategy will therefore enable us to conduct both an impact and a process evaluation, described below.

Analytical Strategy for Evaluating the GEQIP-E Reforms

Our analytical strategy formalises how we aim to estimate the impact of GEQIP-E on learning outcomes, and how we estimate the factors associated with why, or why not, we observe changes in learning outcomes. Learning outcomes are defined in terms of changes in achievement in maths and literacy over the school year, and we utilise the type of school attended as well as the year attended to capture both temporal and spatial variations. More formally, define Y_{iHSt} as a learning outcome for child i living in household H, attending school S at time t. C_i is a dummy variable equalling one if the child belongs to Cohort A (see Table 3), zero otherwise (i.e. it equals zero for children in Cohort B defined in Table 3). Over the course of the study, Cohort A will be exposed to the bundled package of interventions (GEQIP-E) between Grades 1 and 4. Depending on the type of school, Cohort A and Cohort B will be exposed to the teacher training component of GEQIP-E implemented under the phased approach.

Define CC_{iHSt} as a vector (or set) of child characteristics for child i living in household H, attending school S at time t. Some of these characteristics are fixed (child sex) while others will vary over time (child age). HC_{iHt} as a vector (or set) of characteristics of the child's household. Some of these characteristics are fixed (parental schooling) while others will vary over time (household wealth). S_S as a set of dummy (0/1) variables, one for each school, but this approach can be relaxed to take into account differences at school level, as indicated below. Z_{iHSt} as a vector (or set) of other factors that might affect learning outcomes; for

example, road construction that makes it easier for children to walk to school; a drought shock that causes household incomes to fall and, as a response, causes parents to pull children out of school. β as a vector of parameters to be estimated and ε_{iHSt} is the error term, capturing random measurement error and all other factors that affect learning and are not captured in our model.

Recall that we will measure learning outcomes just after the start and at the end of each school year (see Table 3). We use data collected from Cohort B in September-October 2018 and May-June 2019 together with data collected from Cohort A in September-October 2021 and May-June 2022. Define the change in learning outcomes over the school year as ΔY_{iHSt}. With this data, we will estimate:

$$\Delta Y_{iHSt} = \beta_C \bullet C_i + \beta_{CC} \bullet CC_{iHSt} + \beta_{HC} \bullet HC_{iHt} + \beta_S \bullet S_S + \beta_Z \bullet Z_{iHSt} + \varepsilon_{iHSt} \quad (1)$$

Parameter β_C is the coefficient of interest. It describes how much learning differs for a child in Cohort A relative to a child in Cohort B, conditional on child and household characteristics (both fixed and time-varying), fixed and time-varying "other factors" relevant to schooling outcomes (Z_{iHSt}) and school fixed effects. Given we condition on all these different factors, what distinguishes a child in Cohort A from a child in Cohort B is that – at the Grade 4 level – a child in Cohort A is fully exposed to the GEQIP-E reforms), while the child in Cohort B was not exposed in Grade 4. Formally, this can be thought of as an intent-to-treat model.

As the implementation of GEQIP E is non-randomized, the estimated parameter is likely to be biased even with the inclusion of factors at the school, household and child level which may affect learning. The key problem we face is omitted-variable bias. There are additional problems in the estimation, such as the common or parallel trends assumption, and (possibly) changes in sample composition. We can mitigate concerns regarding omitted variable bias, the parallel trends assumption and changes in sample composition by carefully thinking through our data collection strategy and collecting information on many of key factors which may impact learning beyond GEQIP-E. Therefore, we ensure that the baseline survey includes a wide array of time-varying and time-invariant child (including, for example, child height as a proxy for pre-school nutritional status), household, locality and school characteristics (including those not directly impacted by GEQIP E). This allows us to saturate our models with controls, thus reducing (though not eliminating) omitted variable bias.

As noted above, key to our sampling design is the inclusion of Young Lives schools. In the Young Lives sample, we have data on learning outcomes at the start and end of Grade 4 for children observed in 2012-13. Define a second cohort variable, CYL_i, which equals one if a child is in the Young Lives cohort, zero otherwise. Next, re-write equation (1) as:

$$\Delta Y_{iHSt} = \beta_C \bullet C_i + \beta_{CYL} \bullet CYL_i + \beta_{CC} \bullet CC_{iHSt} + \beta_{HC} \bullet HC_{iHt} + \beta_S \bullet S_S + \beta_Z \bullet Z_{iHSt} + \varepsilon_{iHSt} \quad (1')$$

Equation (1') allows us to compare over three cohorts: children exposed to GEQIP-E; children not exposed to GEQIP-E but exposed to prior educational reforms; and children exposed to neither. A similar approach is undertaken if we were to compare children attending Phase One schools and those who are attending other, non-Phase One schools.

Next, recall that we are constructing a longitudinal sample of schools. We have 168 schools surveyed twice, and 38 of these schools surveyed three times. Define ΔY_{St} = as school (*S*) level learning outcomes at time *t* (i.e. learning expressed as the value-added measure). T_{St} as a dummy variable equalling one if the school is observed in 2021-2022; =0 if observed in 2018-2019 (and note that we would have an additional dummy variable if we use add in the Young Lives data). CC_{St} as a vector (or set) of mean characteristics of children attending school *S* at time *t*. HC_{St} as a vector (or set) of mean characteristics of households of children attending school *S* at time *t*. S_S as a set of dummy (0/1) variables, one for each school. SC_{St} as a vector (or set) of school characteristics at time *t*. Z_{St} as a vector (or set) of other factors that might affect learning outcomes averaged at the school level at time *t*. B_S as a vector of parameters to be estimated; and ε_{St} is the error term.

Formally, we consider:

$$\Delta Y_{St} = \beta_S \bullet T_{St} + \beta S_{CC} \bullet CC_{St} + \beta_{SC} \bullet HC_{St} + \beta_{SSC} \bullet SC_{St} + \beta_S \bullet S_S + \beta_Z \bullet Z_{St} + \varepsilon_{St} \quad (2)$$

Equation (2) is a school level analogue of equation (1). At the school level, we can consider outcomes, and control variables, in terms of both mean levels and also in terms of distributional attributes. Therefore, equation (2) allows us to estimate the associations between changes in school characteristics and mean (or distributional) learning outcomes conditioning on child, household and other characteristics as well as school fixed effects.

Process Evaluation: Why, or why not, learning outcomes change

An evaluation of the impact of complex reforms requires also an investigation for understanding why, or why not, learning outcomes change, what explains equity differentials and whether the impact of reforms differ across primary school grades.

The first part of the process evaluation focuses on understanding the characteristics of teachers and/or schools which make it more, or less, likely to take-up the GEQIP E bundle of interventions. One example may be the timely distribution of school grants. Define G_{St} as school grants which equals one if the school

receives its grant, zero otherwise and school characteristics as SC. The correlates of receipt can be modelled as:

$$G_{st} = \beta_{SC} \bullet SC_S + \varepsilon_{St} \quad (3)$$

Relevant school characteristics could include geography (region, accessibility), wealth/poverty of the school's catchment area, resources already available at the school (school building quality; availability of textbooks, etc.) as well as characteristics of the school principal and teaching staff. With the caveat that we are capturing associations, not causal relations, parameter estimates from equation (3) would tell us if timely receipt of grants differs by region.[5] Similar analysis can be performed for teacher characteristics. If we define time subscripts more specifically – that is, we look at G_{st} where t equals some period of time post-estimation and SC_S is measured at baseline (ie 2018/19), we can use the results of equation (3) to assess whether program allocation was uncorrelated with baseline school characteristics.

The second part of the process evaluation focuses on estimating which components (or sub-components) of the GEQIP-E reform are associated with greater changes in learning outcomes. In other words, this part of the process evaluation focuses on establishing whether the 'most effective' schools (in terms of value-added measures in maths and literacy) are also the schools which are the 'best implementers' of the GEQIP-E reforms. Using the treatment-on-the-treated model we assess whether there was a differential effect on learning of different components of GEQIP E. For example, given an indicator variable for school grants, G_{st}, the model is specified as:

$$\Delta Y_{iHSt} = \beta_C \bullet C_i + \beta_{CG} \bullet C_i \bullet G_{St} + \beta_{CC} \bullet CC_{iHSt} + \beta_{HC} \bullet HC_{iHt} + \beta_S \bullet S_S + \beta_Z \bullet Z_{iHSt} + \varepsilon_{iHSt} \quad (4)$$

Where the coefficient of interest, β_{CG}, tells us whether timely receipt of school grants is associated with improved learning, conditional on being exposed to the GEQIP-E reforms. Estimating equation (4) across a range of GEQIP-E components thus will give some insight into which components are more likely to have contributed most to changes in learning outcomes. Additionally, we can construct an index of implementation (e.g. "high"; "medium" and "low" take-up of GEQIP-E interventions) to assess whether schools which are the 'best implementers' of the GEQIP-E reforms are also those that create the greatest value-added in learning.

The third part of the process evaluation is to estimate whether the reforms have differential effects on learning by grade (specifically, comparing learning outcomes across Grades 1 and 4 and comparing learning

[5] A relatively large number of schools in order to be able to estimate equation (3) with precision.

outcomes across Grades 4 and 7). We can specify the following equations, equivalent to a child fixed effects regression for both cohorts separately:

$$\Delta Y_{iHSLA} = \gamma_{TA} \bullet T_{Ai} + \gamma_{CC} \bullet \Delta CC_{iHS2022,\,2018} + \gamma_{HC} \bullet \Delta HC_{iHS2022,\,2018} + \gamma_{Z} \bullet \Delta Z_{iHSt} + \gamma \bullet CHILD_i + \Delta\varepsilon_{iHSt} \quad (5)$$

$$\Delta Y_{iHStB} = \gamma_{TB} \bullet T_{Bi} + \gamma_{CC} \bullet \Delta CC_{iHS2022,\,2018} + \gamma_{HC} \bullet \Delta HC_{iHS2022,\,2018} + \gamma_{S} \bullet \Delta S_{S} + \gamma_{Z} \bullet \Delta Z_{iHSt} + \gamma \bullet CHILD_i + \Delta\varepsilon_{iHSt} \quad (6)$$

The coefficient, $\gamma_{T,}$ tells us, conditional on the assumptions noted in equation (1) whether these reforms had a bigger effect on learning in Grade 4 relative to learning in Grade 1.

Where, $\Delta CC_{iHS2022,2018} = CC_{iHS2022} - CC_{iHS2018}$;
$\Delta HC_{iHS2022,2018} = HC_{iHS2022} - HC_{iHS2018}$

Moreover, it is possible to interact TAi and TBi with the gender of children to examine whether the effect on learning by grades was different for girls and boys, and similarly according to other dimensions of equity.

Discussion

Understanding whether wide-ranging education reforms effectively improve learning outcomes is of first-order importance in many developing countries where, despite gains in school enrolments, learning outcomes remain poor. Ethiopia is one such country, where there is high-level political and financial commitment to such reforms, but also an unwillingness to embark on such changes on a small-scale, incremental basis. In such a context, what is the appropriate impact evaluation strategy? While randomized control trials are frequently used in education research, we argue that an RCT design will limit our ability to understand more fully GEQIP-E given the complexity of the intended reforms and the near-impossibility of ensuring nationwide implementation fidelity in a resource-constrained system. We propose an alternative research design, adopting a longitudinal approach which incorporates variations across time and space, which we are likely to observe in the implementation of the GEQIP-E reforms. Our design also allows us to understand processes of implementation, of interest both in their own right and as a contribution to our assessment of the effectiveness (or otherwise) of GEQIP-E reforms. We will investigate this under a common RISE framework (Pritchett, 2015) which will enable us to understand the reasons behind the potential variation in learning outcomes using a common framework.

We think that RCTs can be a powerful means of understanding whether certain types of education interventions work under certain conditions. But as Pritchett and others have noted that focusing solely on RCTs is not adequate if we aim to accelerate learning through effective reforms and interventions. Of course, no single approach can cover the complexity of GEQIP-E. Although in this paper we have focused exclusively on the use of quantitative methods for impact and process evaluation, these methods are used in RISE Ethiopia as part of a mixed-methods research design to understand and explore the design and implementation of the reforms at national, regional, woreda and school level; the motivation and capacity of key stakeholders to enact the reforms; and the ways in which schools operate under the reform. Systems research is by definition complex, requiring multidisciplinary approaches and multiple research methods to understand the impact of education reforms, as well as the factors behind this impact.

References

Andrews, M., Pritchett, L. and Woolcock, M. (2017) Building State Capability: Evidence, Analysis, Action. Oxford, UK: Oxford University Press

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A.and Sanderful, J. (2013). Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education. *Working Paper 321*. Washington DC: Centre for Global Development.

Cartwright, N. (2012). Presidential address: Will this policy work for you? Predicting effectiveness better: How philosophy helps. Philosophy of Science, 79(5), 973–989.

Cartwright, N. (2016). Where's the Rigor When You Need It? Foundations and Trends® in Accounting, 10(2–4), 106–124.

Deaton, A. (2010). Instruments, randomization, and learning about development. Journal of Economic Literature, 48(2), 424–55.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science & Medicine, 210, 2–21.

Evans, D. amd Popova, A. (2015). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. Policy Research Working Paper WPS7203. Washington DC: The World Bank.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). Impact evaluation in practice. The World Bank.

Glewwe, P. W., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2014). School resources and educational outcomes in developing countries: a review of the literature from 1990 to 2010. in Education Policy in Developing Countries, ed. Glewwe, P. University of Chicago Press: Chicago and London

Glewwe, P., Kremer, M. and Moulin, S. (2007). Many Children Left Behind? Textbooks and Test Scores in Kenya, NBER Working Paper No. 13300. Cambridge MA, USA: National Bureau of Economic Research.

Glewwe, P. & Muralidharan, K. (2015). Improving school education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. RISE Working Paper 15/001.

Gropello, E. D. (2004). Education decentralization and accountability relationships in Latin America. The World Bank.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. Handbook of Econometrics, 6, 4779–4874.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). Journal of Economic Literature, 48(2), 399–423.

Iyer, P. and Rossiter, J. (2018) RISE, Coherent for Equitable Learning? Understanding the Ethiopian Education System [online].

Kingdon, G. G., Little, A., Aslam, M., Rawal, S., Moe, T., Patrinos, H., … Sharma, S. K. (2014). A rigorous review of the political economy of education systems in developing countries. Final Report. Education Rigorous Literature Review. Department for International Development (UK).

Kremer, M., Brannen, C., & Glennerster, R. (2013). "The challenge of education and learning in the developing world." Science, 340(6130), 297-300

Manski, C. F. (2013). Public policy in an uncertain world: analysis and decisions. Harvard University Press.

McEwan, P. (2012). "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." Review of Educational Research 20 (10): 1 –42

Popova, A., Evans D. and Arancibia, V. (2016). Inside in-service teacher training: What works and how do we measure it? Mimeo: retrieved from RISE Programme on 23-3-19 at https://www.riseprogramme.org/sites/www.riseprogramme.org/files/inline-files/Evans_Inside_In_Service_Teacher_Training_CLEAN_v2016-06-22.pdf

Pritchett, L. (2015). Creating education systems coherent for learning outcomes. RISE Working Papers 15/005. Oxford: RISE Programme.

Pritchett, L. (2017). "The evidence" about "what works" in education: Graphs to illustrate external and construct validity. RISE Insight Note. https://www.riseprogramme.org/publications/evidence-about-what-works-education-graphs-illustrate-external-validity-and-construct

Pritchett, L. and Beatty, A. (2012). The negative consequences of overambitious curricula in developing countries. Working Paper 293. Washington DC: Centre for Global Development.

Ravallion, M. (2018). Should the Randomistas (Continue to) Rule? CGD Working Paper 492. Washington DC: Centre for Global Development. https://www.cgdev.org/ publication/should-randomistas-continue-rule.

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt and Emmanuel Jimenez (2016). The impact of education programmes on learning and school participation in low- and middle-income countries. 3ie Systematic Review Summary 7. London: International Review for Impact Evaluation (3ie).

World Bank. 2008. Ethiopia - General Education Quality Improvement Program Project. Washington, DC: World Bank.

World Bank. 2017. Ethiopia - Ethiopia General Education Quality Improvement Program for Equity. Washington, DC: World Bank.