RISE Working Paper 23/125
January 2023

Understanding the Impact of Large-Scale Educational Reform on Students' Learning Outcomes in Ethiopia: The GEQIP-II Case

Mesele Araya, Caine Rolleston, Pauline Rose, Ricardo Sabates, Dawit Tiruneh and Tassew Woldehanna

Abstract

The Ethiopian education system has been very dynamic over recent years, with a series of large-scale education program interventions, such as the Second Phase of General Education Quality Improvement Project (GEQIP-II) that aimed to improve student learning outcomes. Despite the large-scale programs, empirical studies assessing how such interventions have worked and who benefited from the reforms are limited. This study aims to understand the impact of the reform on Grade 4 students' maths learning outcomes over a school year using two comparable Grade 4 cohort students from 33 common schools in the Young Lives (YL, 2012-13) and RISE (2018-19) surveys. We employ matching techniques to estimate the effects of the reform by accounting for baseline observable characteristics of the two cohorts matched within the same schools. Results show that the RISE cohort started the school year with a lower average test score than the YL cohort. At the start of Grade 4, the Average Treatment Effect on the Treated (ATT) is lower by 0.36 SD ($p<0.01$). In terms of learning gain over the school year, however, the RISE cohort has shown a modestly higher value-added than the YL cohort, with ATT of 0.074 SD ($p<0.05$). The learning gain particularly is higher for students in rural schools (0.125 SD & $p<0.05$), which is also stronger among rural boys (0.184 SD & $p<0.05$) than among rural girls. We consider the implications of our results from a system dynamic perspective; in that the GEQIP-II reform induced unprecedented access to primary education, where the national Net Enrolment Rate (NER) rose from 85.7 percent in 2012-13 to 95.3 percent in 2019-20, which is equivalent to nearly 3 million additional learners to the primary education at a national level. This shows that learning levels have not increased in tandem with enrolment, and the unprecedented access for nearly all children might create pressure on the school system. Current policy efforts should therefore focus on sustaining learning gains for all children while creating better access.

Keywords: GEQIP-II, Reform, education, learning, value-added, IPW, Ethiopia

Understanding the Impact of Large-Scale Educational Reform on Students' Learning Outcomes in Ethiopia: The GEQIP-II Case

Mesele Araya
Research for Equitable Access and Learning (REAL) Centre, University of Cambridge

Caine Rolleston
Institute of Education, University College London

Pauline Rose
Research for Equitable Access and Learning (REAL) Centre, University of Cambridge

Ricardo Sabates
Research for Equitable Access and Learning (REAL) Centre, University of Cambridge

Dawit Tiruneh
Research for Equitable Access and Learning (REAL) Centre, University of Cambridge

Tassew Woldehanna
Department of Economics, Addis Ababa University

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 Foreign, Commonwealth and Development Office (FCDO), 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:
Araya, M., Rolleston, C., Rose, P., Sabates, R., Tiruneh, D. and Woldehanna, T. 2023. Understanding the Impact of Large-Scale Educational Reform on Students' Learning Outcomes in Ethiopia: The GEQIP-II Case. RISE Working Paper Series. 23/125.
https://doi.org/10.35489/BSG-RISE-WP_2023/125

1. Introduction

Access to education has substantially improved in developing countries over recent years, mainly due to coordinated global actions following the declaration of Millennium Development Goals (MDGs), Education For All (EFA), and Sustainable Development Goals (SDGs). In Sub-Saharan Africa, Gross Enrolment Rate (GER) for primary school rose from 82.06% in 2000 to 99.91% in 2019, with a slight difference between girls (97.64%) and boys (102.12%) (UNESCO, 2021). Nevertheless, despite expansions in school enrolment, learning levels are low in many contexts. For example, about 125 million children could not acquire basic numeracy worldwide in 2017, even after spending years in school (World Bank, 2018). This is particularly pronounced for children in sub-Saharan countries, where learning poverty reaches up to 80% (Azevedo, 2020). Data from the 2018 World Bank Report also indicate that Grade 2 students who could not perform two-digit subtraction account for about 80% in Uganda, 70% in Ghana, and 60% in Kenya. Low learning levels are also similar in many other African countries, including Ethiopia.

In Ethiopia, learning levels are considerably below curricular expectations (Oketch, Rolleston & Rossiter, 2021). The Ethiopian Ministry of Education conducted national-level maths and reading assessments with Grade 4 and 8 children in 2011 and 2015. Those studies indicated that average test scores in maths and reading were below the minimum expected standards set by the Ministry of Education. In most cases, the proportion of children performing at advanced levels was deficient, below 10% in both subjects (NEAEA, 2016). Besides being below curricular expectations, average numeracy learning levels, as measured by students' test scores, have declined (EAEA, 2016; USAID, 2019). Woldehanna, Araya & Gebremedhin (2016) used cohort comparisons on learning outcomes and indicated that the percentage of correct maths scores for children aged 12 declined from 54.42% in 2006 to 37.17% in 2013.

As a response to the low quality of education and weak learning outcomes, the Ethiopian government, supported by development partners, introduced a series of large-scale education reforms, such as the Second Phase of the General Education Quality Improvement Project (GEQIP-II), aimed at improving learning outcomes of students in the country between 2013/14 and 2017/18. The reform focused on providing sufficient school textbooks and grants, establishing facilities such as a library, pedagogical resource centres, computers, and internet services, and undertaking a Teacher Development Program (TDP) (World Bank, 2013). It was expected that if those and other related inputs or resources were provided to schools, the learning levels of all primary school students would improve.

This study aims to understand whether the GEQIP-II educational reforms raised students' learning outcomes. Until now, empirical evidence on the contributions of the reform is scant and similar reforms in other African countries have mixed results, with some being effective in boosting students' learning. For example, Binci, Hebbar, Jasper, and Rawle (2018) evaluated the impact of the Education Quality Improvement Programme in Tanzania (EQUIP-T) on primary education and indicated that EQUIP-T had reduced the proportion of pupils in the bottom performance band for literacy in the program schools. Mbiti, Romero, and Schipper (2019) also compared the effectiveness of two teacher performance pay systems on students' learning and concluded that both systems improved student test scores in the Tanzanian education system. Niels-Hugo Blunch (2014) estimated education outcomes before and after the 1987 Ghanaian Education Reform and showed that numeracy skills levels increased following the reform. Similarly, Duflo, Dupas, and Kremer (2015) assessed Extra Teacher Program (ETP) in Kenya and found a 0.20 SD learning effect on students assigned to contract teachers. Piper, Ralaingita, Akach & King (2016) evaluated the impact of the Primary Math and Reading (PRIMR) Initiative on pupil numeracy achievement and showed that the intervention was able to have modest statistically significant results.

School grants also improved students' test scores in Senegal (Carneiro et al., 2020), while performance pay on recruitment and effort significantly increased student test scores in Rwanda (Leaver et al., 2019). Using a meta-analysis of impact evaluations, Conn (2017) calculated that pedagogical interventions in sub-Saharan Africa have an average effect of 0.54 SD (see Hill, Bloom, Black, & Lipsey, 2008). More broadly, Bashir, Lockheed, Ninan, & Tan (2018) reviewed interventions related to structured pedagogy programs in low-and-middle-income countries and synthesised that structured pedagogy programs raised learning levels by 0.23 SD on average. Conversely, some reforms ended with no significant boost in learning levels. For example, Dedehouanou & Berthe (2013) found no strong relation between school-based management and performance in Mali. A study from Duflo, Dupas, & Kremer (2015) on programs related to reduced class size in Kenya could not significantly increase students' test scores. Interventions pertaining to school grant schemes did not improve learning levels in Niger (Beasley and Huillery, 2017), Gambia (Blimpo et al., 2015), and Tanzania (Mbiti et al., 2019a).

Although the implementation of GEQIP-II reform was already completed a few years ago, large-scale cohort studies that evaluate the impacts of the reform on improving primary school students' learning outcomes over time are missing. The reform's effects on the marginalised students, including girls and children from rural localities, have not been fully understood. To

the best of our knowledge, the only assessment conducted so far is by the Ministry of Education (2019) on the exit evaluation of the reform using data collected from 130 sample woredas (districts) of the country. The exit evaluation mainly focuses on the standard measurement indicators of the reform, such as dropout rate reduction, teaching effectiveness, utilisation of textbooks, school inspection, and on-time arrival of school grants. The results obtained from the exit evaluation indicate that GEQIP-II has been effective in meeting targets like dropout rate reduction and percentage coverage of school inspection but did not meet targets related to textbook-student ratio, textbook utilisation, and on-time arrival of the school grants. Although such a national exit evaluation is useful, the assessment does not include students' learning outcomes directly collected from students and their households and primary caregivers, who could considerably impact learning levels and progress over a given school year.

The assessment also provides limited information on the program's potential impact on students' learning outcomes using benchmark data before implementing GEQIP-II. This study aims to fill this research gap by comparing the learning levels of two cohorts of Grade 4 students with pre-and-post-GEQIP-II learning data. We specifically seek to estimate the differences in maths learning levels and progress in maths over a school year between the two cohorts who attended Grade 4 before and after the GEQIP-II reforms. We explain how improvements in maths scores over a school year are attributed to GEQIP-II after controlling for observable socioeconomic status (SES) and child characteristics, matching and balancing the data at the school level.

The rest of the paper is organised as follows: Section 2 provides an overview of GEQIP-II reforms, and section 3 describes the data type, sample, and methods used in the analysis. Section 4 presents the main findings. Finally, sections 5-7 provide discussion, strength-limitations, and concluding remarks.

2. GEQIP Reforms in Ethiopia: an overview

Ethiopia is among the least income countries globally but has significantly increased funding for the education sector over the past two decades with the support of donations (World Bank, 2020). For example, in 2016-17, education spending accounted for 27 per cent of total government expenditure, significantly higher than the government's commitment to internationally agreed targets of 20 per cent of the national budget for education (UNICEF, 2017). International development agencies have also been calling for more significant resources to be devoted to education and have increased their levels of assistance for education projects in Ethiopia (Ministry of Education, 2015; World Bank, 2017).

One of Ethiopia's main education investment areas has been increasing primary school enrolment rates and improving learning outcomes equitably. Accordingly, primary education enrolment has rapidly expanded from three million learners in the early 1990s to over 20 million in 2018/19 (Ministry of Education, 2019). However, despite the tremendous progress in expanding access to primary education, learning levels have remained low (Ministry of Education, 2010, 2015; World Bank, 2017). Many children complete their primary education lacking basic literacy and numeracy skills (e.g., NEAEA, 2016; USAID, 2019). Some findings indicate that average numeracy learning has deteriorated over recent years (NEAEA, 2016; Woldehanna et al., 2016). Recognising the inadequacy of the primary education system to equip children with the required knowledge and skills sufficiently, significant efforts began in 2008 to address issues of raising learning outcomes by introducing government-and donor-supported extensive educational reforms.

One of the most prominent donor-supported reforms focusing on increasing equitable learning outcomes is the General Education Quality Improvement Program (GEQIP) (World Bank, 2008, 2013). The GEQIP reforms have been implemented in Ethiopia since 2008 in three consecutive phases: GEQIP-I (2008-2012), GEQIP-II (2014-2018), and currently, GEQIP-for Equity (GEQIP-E: 2018-2022). The reforms have been comprehensive and nationwide. The overall aim is to enhance students' learning outcomes equitably by improving teaching and learning conditions in schools and strengthening educational institutions and service delivery at federal and regional levels (World Bank, 2008).

With a total budget of about US$ 500 Million (World Bank, 2020), the GEQIP-II reform focused explicitly on providing essential inputs to all public schools for improving teaching and learning conditions, such as increasing the supply of qualified primary school teachers, providing continuous in-service training for teachers to enhance their content knowledge and pedagogical content knowledge, providing students with textbooks for each subject, and funding school improvement plans through per capita school grants provided based on enrolment (World Bank, 2008). The Theory of Change of the GEQIP-II reform that aims to prompt changes within the education system to improve students' learning conditions is summarized below.

Theory of Change of the second Ethiopia General Education Quality Improvement Project (GEQIP-II)

Components	Project Activities	Intermediate Project outcomes	Outcomes
Component 1: Curriculum, Textbooks, Assessment, examination and inspection	• Supplying of teaching and learning materials • Provision of e-Braille display readers • Support assessments and examinations • Roll-out of a school inspection system	• Increased access to teaching and learning materials, including children with disabilities • Better evidence on progress and determinants of student learning and performance of schools • Improved quality assurance and accountability	• Improved learning conditions in primary schools
Component 2: Teacher Development Program	• Pre-service teacher training • In-service teacher training • Licensing and relicensing of teachers and school leaders	• Improved content knowledge and pedagogical skills of teachers for the delivery of student-centred teaching and learning • Improved quality school leaders	• Improved learning conditions in secondary schools • Strengthened institutions at a different level of educational administration
Component 3: School improvement plan	• Support of school improvement plan • Provision of school grants	• Improved availability of operational and learning resources in schools • Continuously participatory school improvement	
Component 4: Management and capacity building, including EMIS	• Support of EMIS capacity building and data collection • Capacity building for education planning and management • Capacity building for education planning and management	• Improved timeliness and quality of data for education panning and management • Increased capacity for planning and management (at central, regional, woreda, and school l levels)	
Component 5: Use of information and communications technology	• Provision of ICT infrastructure to target educational target • Development of integrated M&E and learning system • Support strengthening of the national policy and institution for ICT	• Improved learning conditions in specific secondary schools and universities ICT (E-Cloud) • Increased capacity for ICT in general education • Improved M&E for selected ICT interventions	
Component 6: Program Coordination, Monitoring and Evaluation, and Communication	• Program Coordination • Monitoring and Evaluation • Communication	• More effective project management, implementation and communication • Timely monitoring of project progress, results and impact towards institutional strengthening and improving learning conditions	

Source: World Bank (2020). Project Implementation Completion and Result Report for GEQIP II

3. Data and Method

3.1. Data sources

This study uses pooled datasets collected from 33 primary schools in the Young Lives (YL) and Research on Improving Systems of Education (RISE) Ethiopia surveys in 2018-19. We constructed the pooled data to explore how GEQIP-II reform contributed to students' learning outcomes, as measured by maths learning levels and value-added scores over a school year. Through matching analysis that compares students with similar socio-economic backgrounds within the same schools, we find sufficient comparability between the two cohorts, mainly in terms of baseline observable control variables and test items. Below details of the two school surveys are provided.

3.1.1. Young Lives 2012-13 School Survey

Young Lives (YL) is a longitudinal study of childhood poverty conducted in Ethiopia starting in 2002, tracing the lives of children through household and school surveys. The 2012-13 School Survey included nearly 12,000 students studying in all Grades 4 and 5 classes in 30 purposely selected sites located across seven regions in Ethiopia: Addis Ababa, Afar, Amhara, Oromia, Southern Nations, Nationalities, and People's (SNNP), Somali, and Tigray. The YL school survey is based on a census of all students studying in the selected schools in Grades 4 and 5 at the time of the study. The school survey offers a unique perspective on regional and site differences in child, teacher, and school characteristics and the factors influencing progress in maths over a single school year. The survey was conducted in two Rounds: Round 1 at the start of the 2012-13 school year in October and Round 2 towards the end of the school year in May. In Round 1, the survey included a student questionnaire and assessment of maths and questions related to household items. In Round 2, students completed the second set of learning assessments in maths. A total of 10,068 students in 94 schools and 280 Grades 4 and 5 classes were surveyed in Rounds 1 and 2 (for details, see Aurino et al., 2014). A total of 5,100 Grade 4 students from 142 classes took the numeracy tests in Rounds 1 and 2 (see Table 1). But our sample in this study is restricted only to the 33 common YL and RISE schools to maintain comparability of school resources and capacity over time (see Table 1A for the detailed description and sites of the common schools).

3.1.2. RISE Ethiopia 2018-19 School Surveys

RISE Ethiopia adopts a similar longitudinal design to YL to understand the impacts of GEQIP reforms on equitable access to quality primary education for all children. The target population are Grades 1 and 4 school children, parents or primary caregivers, school principals, and maths teachers in seven regions: Addis Ababa, Amhara, Benishangul Gumuz, Oromia, SNNP, Somali, and Tigray. The number of schools in each region is approximately proportionate to the population in each region and includes (a) schools from the YL School Survey (2012-13); (b) schools targeted in the first phase of the GEQIP-E reforms, and (c) a random selection of additional schools to represent the urban and rural populations in each region. 28 Grade 4 pupils were randomly selected from up to two classes (Araya et al., 2022).

Similar to the YL school survey, the RISE survey was conducted in two Rounds: Round 1 at the start of the 2018-19 school year in October/November and Round 2 towards the end of the school year in May/June. In Round 1, both the school and household surveys were conducted. The school survey focuses on students' assessment of maths, while the household survey

provides information on the household socioeconomic backgrounds of the students. In Round 2, students completed the second set of learning assessments in maths. A total of 3353 Grade 4 students in 166 schools were surveyed in Rounds 1 and 2 (Table 1).

Table 1. Young Lives 2012-13 and RISE 2018-19 surveys school and student samples

		YL 2012-13 (Pre-GEQIP-II Cohort)		RISE 2018-19 (Post-GEQIP-Cohort)	
		Number of schools	Number of students	Number of schools	Number of students
	Total*	94	5100	166	3353
Region	Addis Ababa	12	1093	20	464
	Amhara	13	578	25	516
	Benishangul-Gumuz**	-	-	19	371
	Oromia	8	494	41	848
	SNNP	20	1146	22	434
	Somali	19	586	19	279
	Tigray	13	723	20	441
	Afar**	9	480	-	-
Gender	Female		2656		1639
	Male		2393		1714
Location	Rural		1431		2207
	Urban		3669		1146

Source: Young Lives 2012-13 and RISE Ethiopia 2018-19

Notes: *The total numbers indicated for both YL and RISE include only those participants who took both the Round 1 and Round 2 tests; **Benishangul-Gumuz region was not included in the YL 2012-13 school survey, and Afar region was not included in the RISE Ethiopia 2018-19 surveys. We have excluded two of them in our analysis as we don't have common schools.

3.2. Study participants

The analysis of this paper covers only some of the YL and RISE sample schools. As mentioned in the previous sub-section, of the RISE schools surveyed in 2018-19, only 33 were part of the YL school survey. Therefore, it only focuses on the sub-set of schools. However, it is worth noting that despite an equal number of schools across the two school surveys, the total number of participants from the common schools differs because of sampling differences within a school. For example, all Grade 4 students from the selected schools were included in the YL survey, whereas only 28 students were randomly selected from each school in the RISE survey. This means that the participants within a school vary in number between the two surveys due to sampling strategy differences. As reported in Table 2, we have 2,879 sample students from the two cohorts in 33 common schools, 2,190 sample students from the YL (pr-GEQIP-II) and 689 from the RISE (post-GEQIP-II) cohort. To adjust for the sampling difference between the two school surveys, we used Inverse Probability Weighting (IPW) techniques to correct the analysis by reweighting the observations with the probability of being selected for the study (Narduzzi et al., 2014).

Table 2. Sample Students from 33 Common Young Lives and RISE school surveys

		YL 2012-13 **Number of students**	RISE 2018-19 **Number of students**	Total sample **Both**
	Total*	**2,190**	**689**	**2,879**
Region	Addis Ababa	220	79	299
	Amhara	427	141	568
	Oromia	409	122	531
	SNNP	550	143	693
	Somali	137	49	186
	Tigray	447	155	602
Gender	Female	1,129	351	1,480
	Male	1,035	338	1,373

Source: Young Lives 2012-13 and RISE Ethiopia 2018-19

3.3. Instruments

Comparable maths test items were administered at the start and end of the school year to measure students' progress in maths over a school year in both surveys. The YL survey included 25 multiple-choice items administered at the start of the school year (Round 1) and end (Round 2). Both Round 1 and 2 tests had 19 commons (anchor) items, while the remaining six items in Round 2 were unique. The RISE maths tests were adapted from the YL maths test items. There were 25 multiple-choice items administered at the start (Round 1) and end (Round 2). Similar to the YL, both Round 1 and 2 test items included 15 common (anchor) items and ten items in Round 2 were unique. Taking both the YL and RISE school surveys together, there were 13 items common across the 4 Rounds. The complete list of unique items was 41(see Table 2A for the percentage of correct responses for each maths test item across the 4 Rounds).

Our initial item fit analysis indicated that the 41 unique items overall functioned well across the 4 Rounds with acceptable item difficulty and discrimination indices. To estimate differences in learning progress in maths over a school year between the two cohorts, we employed a concurrent calibration approach, helped by many common 'anchor' items across the 4 Rounds. A two-parameter-logistic item response theory model (2PL IRT) was fitted to the item responses. The 2PL IRT model provides parameter estimates on a common interval scale. In concurrent calibration, item parameters are estimated simultaneously using pooled data from all rounds, with responses to the unique items to each group treated as missing for respondents who did not receive them. The anchor items provide the link between tests, while the unique items increase the precision of estimates for individual tests. This approach has proven effective in accurately estimating item parameters for all the test takers, especially when linking scores across periods. We transform the students' latent trait estimates to a scale with a mean of 500 and a standard deviation (SD) of 100 for ease of reference. Once the maths test

scores were transformed, we then conducted a matching analysis on three outcomes: 1) baseline test maths score at the start of Grade 4, 2) end-line test maths score at the end of Grade 4, and 3) value-added scores (progress) over a school year, which is the difference between the two tests. Table 3 summarises the three learning outcome indicators.

Table 3. Learning outcome indicators

Indicators for learning outcomes	Descriptions
Baseline maths Score (IRT)	Maths test scores at the start of a school year transformed to IRT with a mean value of 500 and a standard deviation of 100.
End-line maths Score (IRT)	Maths test scores at the end of a school year transformed to IRT with a mean value of 500 and a standard deviation of 100.
Value-added score or progress	Learning gain on maths scores was obtained by subtracting the baseline maths score from the end-line maths score.

3.4. Estimation Strategies

We employ a Propensity Score Matching (PSM) method, which ensures equivalence between treated and comparison groups in terms of observable baseline covariates. In a situation where Randomized Controlled Trial (RCT) cannot be applied, PSM provides an alternative way to evaluate a reform's impact by matching two groups, controlling for observable baseline characteristics. We largely follow the work by Binci, Hebbar, Jasper & Rawle (2018), who evaluated the impact of the Education Quality Improvement Programme in Tanzania (EQUIP-T) on the Tanzanian primary education subsector. These authors used two repeated cross-sectional data and developed methodological guidance that can be applied in the education sector for repeated cross-sectional school surveys with matching. Matching techniques assume that the only remaining relevant difference between the two cohorts is the GEQIP-II reform. Selection bias is minimised by controlling baseline observable covariate balanced at the school level. This means we compare children in the common support with similar conditions and socioeconomic status at a school level. The matching method is also expected to resolve any bias arising from sampling procedure differences within a school between the two cohorts of the study. Studies show that PSM can effectively reduce bias for analyses with sampling differences by comparing like-with-like across two survey rounds (Stuart, 2010; Howarter, 2015).

As a nationwide education reform, GEQIP-II has multiple packages: textbook availability and utilisation, school inspection standards on teachers' knowledge, lesson planning, teaching practices, and assessment practices (see section 2 for GEQIP-II's Theory of Change). As the

inputs are too many, it is not easy to quantify each indicator of the GEQIP-II reform and use it in the matching analysis at a time. Instead, we take the program as a whole package to estimate its impact on learning achievements. Also, it was implemented in all public schools as an extension of the GEQIP-I and with a total budget cost of about USD 550 million over its lifespan. So, we assume that the post-GEQIP-II cohort of children in the 33 public schools had equal access to the reform by 2018-19 compared to the pre-GEQIP-II cohort in the same schools in 2012-13. In this manner, all the post-GEQIP-II cohort pupils from the 33 common schools are considered a treated group. In contrast, the pre-GEQIP-II cohort of children from the YL school survey is regarded as a comparison group.

Estimating PSM models involves two steps: 1) estimating the probability of being treated based on selected covariates to create a comparable group in common support, and 2) selecting an appropriate matching algorithm to estimate the impact of the education reform. The first stage of the estimation assumes the GEQIP-II reform as a dependent variable and applies the following binary model:

$$Pr(t_i = 1) = F(\alpha x_i) + \varepsilon_i \qquad (1)$$

where *t* is the treatment variable that assumes 1 if the student *i* attended Grade 4 in the 2018-19 academic year–as a post-GEQIP-II cohort and 0 for a student *i* who attended Grade 4 in the 2012-13 academic year–as a pre-GEQIP-II cohort. F(.) is a binary function with X as a vector of observable factors, α as a vector of parameters to be estimated, and ε is an error term.

Equation (1) helps us identify the number of students on the common support comparable before and after the GEQIP-II reform. Once students in the common support are determined based on the propensity score estimated from the binary model outlined above and the covariance balance is satisfied at the school level, the next step is then to estimate the Average Treatment Effect on the Treated (ATT) of the GEQIP-II program as follows:

$$ATT \equiv E(ML_{i1} — ML_{01} | t_i = 1) \qquad (2)$$

where t_i is dummy variable which is 1 if the student is from the post-GEQIP-II cohort and 0; otherwise, ML_{i0} and ML_{i1} are maths learning, with ML_{i0} the score of outcome that would be observed if the student is from the pre-GEQIP-II cohort; ML_{i1} is the maths score observed on the same grade for a student from the post-GEQIP-II cohort. To select an appropriate matching technique, we ran different matching algorithms: nearest neighbours matching; radius

matching; kernel matching, and Mahalanobis matching and finally chose the Kernel matching method by applying bias/variance trade-off in the estimated treatment effect (Binci, Hebbar, Jasper & Rawle, 2018) (see Table 4).

3.4.1. Covariate Selection

We are very selective with the potential covariates and limit them to those strongly correlating with learning outcomes rather than the reform itself. As the schools are the same for both cohorts, we don't include school facilities variables in the PSM model. Also, these school facilities are part of the GEQIP-II reform packages and are more likely to be affected by the reform (see section 2 for GEQIP-II's Theory of Change). A propensity score that includes covariates affected by the GEQIP-II can bias results (Imbens, 2004; Garrido, 2014).

Nevertheless, there are some covariates such as preschool participation, school distance, any record of dropout, and time used on a typical day at home and school, including time spent on domestic chores, farming, working for pay or studying/doing homework might be affected by the reform. We conducted several matching analyses with and without these variables as we assumed the reform might have involved these covariates in the first instance. However, their exclusion doesn't significantly change the results; we then decided to include them in the analyses as they are more likely to be associated with child home experience than the reform itself. For example, the time used by a child at home is more likely to be affected by her family background than by the GEQIP-II reform. All selected baseline observable covariates for the PSM calculations are reported in Appendix 4A.

Furthermore, we undertake the PSM estimates at the school level to improve matching quality using school identifiers, including all sub-samples (gender and locality). Using these potential cofounders and school identifiers, we ran logit models to estimate the propensity score of each observation. The estimated binary models to generate propensity scores on the reform are reported in the Appendix (Table 3A).

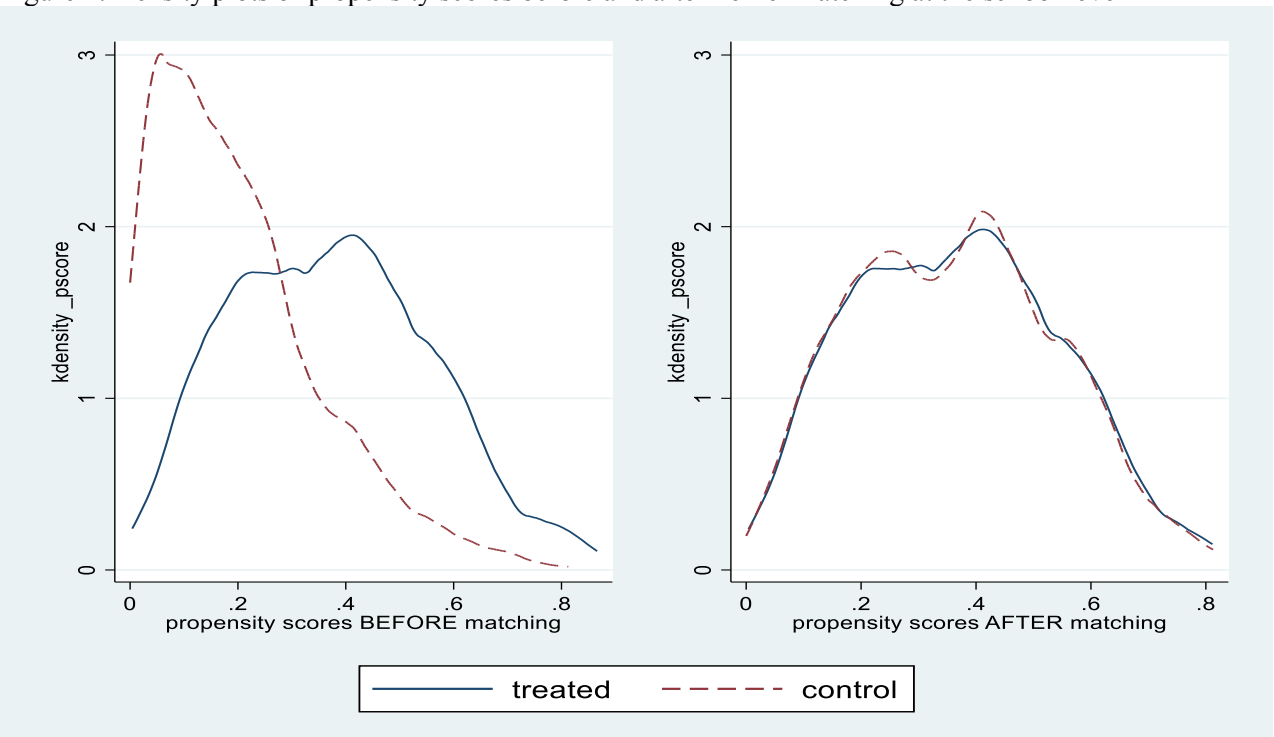
3.4.2. Balancing the Data

Once the propensity score of each sample has been generated using a logit regression, the next important step is to check that the baseline covariates balance property is satisfied and ensure that there is sufficient overlap on the covariates of both cohorts. For the PSM model to work well, a significant number of the students should be within the "common support". When there is sufficient overlap in the range of propensity scores across the two cohorts, the distribution

of propensity scores between the two cohorts will be similar, and the balance property will be satisfied.

Figure 1 portrays a kernel density plot on learning before and after kernel matching. The distribution is better balanced after the kernel matching than before the balancing effort is made. The figure shows a sufficient overlapping between the observable covariates of the two cohorts after matching. All sub-sample density plots of propensity scores before and after kernel matching within the same schools are available on request.

Figure 1. Density plots of propensity scores before and after kernel matching at the school level

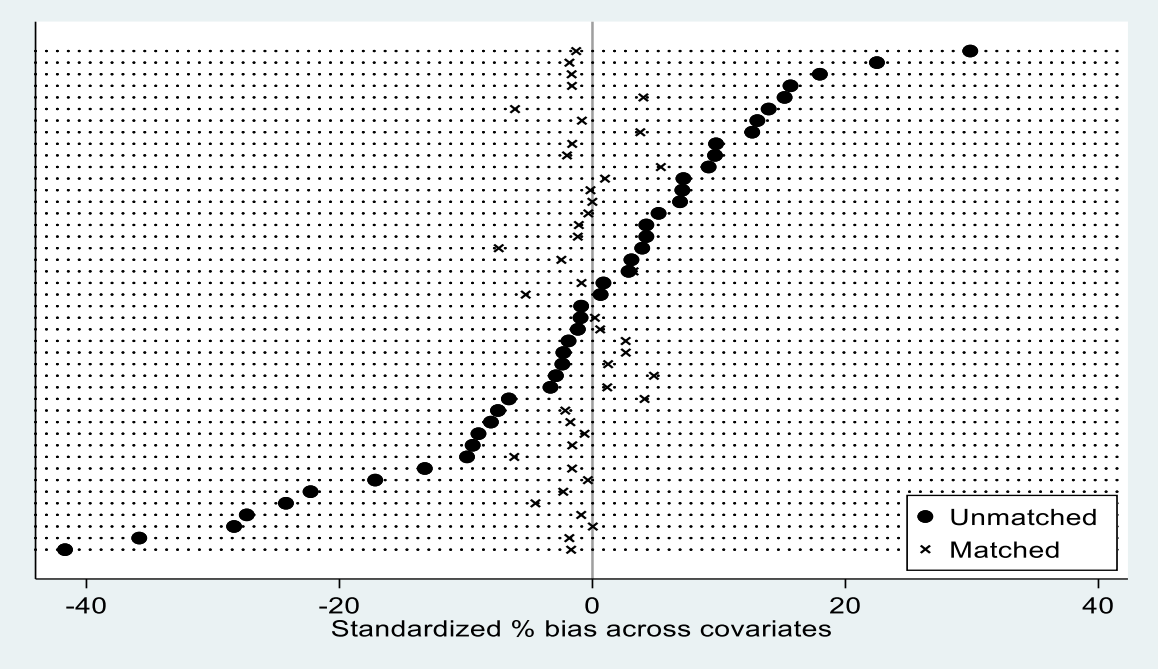


The property of covariate balance can also be substantiated by comparing the standardised difference of the mean values of the covariates before and after matching analysis. After carrying out the matching analysis, the difference in the standardised difference of the average value of each covariate needs to be statistically insignificant to be different for the balance propensity to be satisfied. The standardised difference of each covariate before and after kernel matching analysis is reported in Table 4A. Before matching, we find significant differences in the standardised difference of the baseline covariates' mean values, with most of them statistically significant to be different. That is, significant variations are observed before matching, where children from the pre-GEQIP-II cohort relatively appeared to be advantageous and are from higher socio-economic background families: one-third (37.8%) of the pre-GEQIP-II cohort are in the third tercile groups of household asset possession compared to only one-fifth (21.6%) for the post-GEQIP-II cohort. The same is true with primary caregiver's

status, where a higher proportion of the pre-GEQIP-II cohort are from literate parents (49.3%) than the post-GEQIP-II cohort (40.8%). Also, on a typical day, a higher proportion of children from the post-GEQIP-II cohort spent more time on domestic chores, working on farms, or for pay than the pre-GEQIP-II cohort, which implies that children from the post-GEQIP-II cohort are from more disadvantaged backgrounds.

After balancing the baseline covariates, the difference disappears and becomes statistically insignificant for all the covariates, implying that the covariate balance between the two cohorts is achieved, and children who are not in the common support are dropped from the study. Of the 2,879 sample students from both cohorts, we found 57 students out of common support and, therefore, excluded from the study. A similar balancing pattern is observed from the standardised percentage bias across covariates from kernel matching distribution at the school level. As depicted in Figure 2, the absolute standardised preference across covariates is closer to zero for the matched than for the unmatched samples.

Figure 2. Standardised bias across covariates and school level



Post-estimation overall model indicators also guide whether the matching analysis works well. Rubin's B and R values are some of those post-estimation indicators. For a baseline covariate balance property to be achieved, Rubin's B needs to be below 25, while Rubin's R-value should be within the range of 0.5-2 (Rosenbaum & Rubin,1985). After the matching estimation, we generated Rubin's B and R values (Table 4). The average standardised difference in covariates before matching is very high. However, after balancing the covariates at the school level,

Rubin's B declined to 17.3, which is below the threshold value (25) for unbiased estimates of covariates. Similarly, Rubin's R, a ratio of variances in the propensity score, is within the range of 0.5-2, which is acceptable for matching to be satisfactory in terms of covariate balance. Overall, the results show no systematic differences in the selected baseline observed characteristics after balancing. This enabled us to estimate the impact of the reform on the learning outcomes of the two cohorts with a minimum sample selection bias.

Table 4. Alternative algorithms

	Number of Treated students in a common support	Number of Comparison students in a common support	Mean Standardized Difference in Covariates (%)	Median Standardized Difference in Covariates (%)	Rubin's B	Rubin's R
			11.2	8.5	104.3*	0.63
Kernel matching	666	2,133	2.2	1.6	17.3	1.01
k-Nearest neighbours matching (K=4)	666	2,133	3.9	3.1	34.1*	0.51
Radius matching	664	2,133	3.2	2	31.3*	0.49*
Local linear regression matching	664	2,133	4.6	3.5	41.9*	0.89
Mahalanobis matching	674	2133	3.1	0	65.7*	0.95

3.4.3. Choice of Algorithm

Given the satisfactory balance of covariates, alternative algorithms can do estimations. We run several alternative algorithms to choose one among the several methods that minimise the selection bias: nearest neighbours, radius, kernel, local linear regression, and Mahalanobis matching (Table 4). Of the five matching algorithms, kernel matching demonstrates a significant reduction in the mean standardised difference of the covariates, from 11.2 per cent before matching to 2.2 per cent after matching, and maintains a good number of students from both cohorts. Algorithms like radius matching and Mahalanobis matching also significantly reduced the mean standardised difference to 3.2 % and 3.1%, respectively. However, these are greater than the values in the kernel matching method. We then chose kernel matching to adjust our sample for selection bias and used it to estimate the impact of the reform on learning levels and value-added scores over the school year. Also, one advantage of kernel matching is that it has a non-parametric matching estimator that uses a weighted average of all observations to generate the counterfactual match for each sample (Khandker et al., 2010), which is appropriate for our analysis as there are sampling differences within schools between the two cohorts. As pointed out earlier, interpretations are made in terms of the estimated Average Treatment Effect on the Treated (ATT). ATT, in this case, is the average treatment effect of the GEQIP-II reform

computed as the mean difference in learning outcome across the matched samples. We only compared children with similar propensity scores in both cohorts to obtain the effects of GEQIP-II reform on learning outcomes. Children for which no match is found are dropped from the analysis. Also, we use bootstrapping to improve the validity of standard errors instead of the traditional standard errors. Because in the process of PSM estimation, the estimated variance of the treatment effect needs to include the variance attributable to the derivation of the propensity score and common support determination (Austin & Small, 2014). Ignoring this variation will cause the standard errors to be misestimated (Khandker et al., 2010; Heckman, Ichimura, and Todd, 1998). So, bootstrapping will likely lead to valid standard errors (Imbens,2004; Khandker et al., 2010). Furthermore, as a robustness check, we apply Inverse Probability Weighting (IPW) to adjust the analysis by reweighting the observations (Narduzzi et al., 2014). This is particularly important as our sampling strategies within a school are different between the two cohorts. IPW estimator is called ‘‘double robustness’’ (Glynn & Quinn, 2010).

4. Major Findings

4.1. Maths learning levels and value-added score for all samples

Table 5 reports the estimated Average Treatment Effect on the Treated (ATT) on maths learning levels and value-added score over a school year from kernel matching estimations. The first two columns provide the ATT on learning levels at the start and end of Grade 4. The average maths learning levels of the post-GEQIP-II cohort are lower than that of the pre-GEQIP-II cohort both at the beginning and end of Grade 4. At the start of Grade 4, the post-GEQIP-II cohort had, on average, scored 35.78 lower scale score points (-0.3578 SD; $P<0.01$) than the pre-GEQIP-II cohort. Similarly, by the end of the school year, this was 28.39 lower scale score points (-0.2839 SD; $P<0.01$) for the post-GEQIP-II cohort, entailing the mean test-score difference between the two cohorts declined by the end year. This way, we run matching analyses on the value-added over the school year, which is the difference between the baseline and end-line maths scores. The result is reported in the third column. Despite the lower test scores for the post-GEQIP-II cohort both at the start and end-year of the academic year, the value-added score over the school year is higher for the post-GEQIP-II cohort. Compared to the pre-GEQIP-II cohort with similar socio-economic backgrounds and matched within the same schools, the post-GEQIP-II cohort, on average, added 7.389 larger scale score points (0.074 SD; $P<0.05$) to their initial mean test score by the end of the school year. This higher value-added over the school year for the post-GEQIP-II cohort might be attributed to the

GEQIP-II reform despite the lower initial mean test scores. That means there seems to be a "catch-up effect" of the reform in that lower achievers from the post-GEQIP-II cohort were able to add more value-added learning than higher achievers over the school year. This might further suggest that GEQIP-II supported students with the lowest learning skills over the school year.

4.2. Average learning levels and value-added by locality

The preceding section of the analysis does not distinguish between rural and urban areas and does not tell us how the GEQIP-II reform has been faring in terms of value-added learning for rural and urban students. It thus is essential to examine the learning levels and value-added scores over the school year separately by locality, as such analysis helps us identify who has benefited from the large-scale educational reform. To do this, we further conducted a separate balance test of the data at the sub-samples for the rural and urban localities at the school levels.

Table 5 also presents the ATT estimates for both rural and urban schools. For the rural schools, the post-GEQIP-II cohort scored lower than the pre-GEQIP-II cohort during the baseline survey by -47.81 scale score points (-0.4781 SD; $P<0.01$). This suggests that at the beginning of the school year, the rural post-GEQIP-II cohort has significantly lower learning levels than the pre-GEQIP-II cohort. It is not easy to justify what drives the decline in initial learning level, but this might be related to a high enrollment induced by the educational reform itself. It is also important to mention that a large proportion of the post-GEQIP-II cohort is first-generation learners with less preparation for schooling (Iyer et al., 2020), who might need additional school resources to maintain a smooth educational process. However, by the end of the school year, the difference in the learning gap appeared to decline significantly, with the ATT falling to -35.36 Scale score points (-0.3536 SD; $P<0.01$). This means that when the rural post-GEQIP-II cohort entered Grade 4 in 2018-19, they already had relatively low skills than the rural pre-GEQIP-II cohort who joined Grade 4 in 2012-13, and much was expected from them to catch up over the school year.

To examine the difference in value-added between the two cohorts, we similarly conducted the PSM on the rural cohort at the school level. A good part of the rural PSM estimation on the value-added learning is that there is a sizeable improvement in value-added for the post-GEQIP-II cohort. The rural post-GEQIP-II cohort's value-added learning is almost double the gain obtained for the whole sample (rural and urban). This can be seen in Column 3 of Table 5, where the reform is statistically significant to impact the rural value-added learning score over the school year. The rural post-GEQIP-II cohort achieved 12.45 larger scale score points

(0.1245 SD; P<0.01) over the school year than the rural pre-GEQIP-II cohort. This is a piece of empirical evidence that the large-scale educational reform has benefited the very rural student, at least in terms of progress (catching up) over the school year.

Like the rural cohorts, we also estimated the urban cohorts separately within the same schools. This helps to show how the reform has been working in terms of learning levels and value-added scores in urban schools. The results for the urban cohort are reported in the bottom part of Table 5. The urban post-GEQIP-II cohort also experienced lower learning levels than their counterparts. At the start of Grade 4, the urban post-GEQIP-II cohort scored 20.15 lower scale score points (-0.2015 SD; P<0.01) than the urban pre-GEQIP-II cohort. Equally, by the end of the school year, the post-GEQIP-II cohort achieved 17.46 lower scale score points than the urban post-GEQIP-II cohort (-17.46 SD; P<0.01). However, we observe no statistically significant difference in value-added learning between the two urban cohorts. Unlike in the rural schools, when controlled for baseline covariates matched at the school level, there is no strong evidence of the value-added difference (4.09 SD; P>0.1) between the urban pre-and-post-GEQIP-II cohorts.

Table 5. GEQIP-II reforms on learning levels and value-added scores using kernel matching by locality

	(1)	(2)	(3)
	Start Grade 4 (score)	End Grade 4 (score)	Value-added score
Rural and urban			
ATT	-35.78***	-28.39***	7.389**
	(5.16)	(4.48)	(3.59)
N	2799	2799	2799
Rural area			
ATT	-47.81***	-35.36***	12.45**
	(7.60)	(6.80)	(5.75)
N	1201	1201	1201
Urban area			
ATT	-24.22***	-20.12***	4.099
	(6.13)	(6.71)	(4.78)
N	1598	1598	1598

Note: ATT=Average Treatment Effect of the Treated from kernel matching; standard errors in parentheses (bootstrapping); *** p<0.01, ** p<0.05, * p<0.1

4.3. Learning levels and value-added scores by gender and locality

We further disaggregate our analysis by gender and locality as this might have important implications for those who benefited from the GEQIP-II reform. We estimated the ATT in four

categories: rural boys, rural girls, urban boys, and urban girls. Table 6 presents the ATT on these groups' learning levels and value-added scores. Like the whole sample, both boys and girls from the post-GEQIP-II cohort have significantly lower learning levels at the start and end-year maths scores. However, the decline in learning level is higher among the post-GEQIP-II girls than boys. At the beginning of Grade 4, girls in the post-GEQIP-II cohort scored 43.45 scale score points lower than their peers in the pre-GEQIP-II cohort (-0.4345 SD; P<0.01). Similarly, by the end of the school year, girls in the post-GEQIP-II cohort scored 34.84 lower scale score points than their peers in the post-GEQIP-II cohort (-0.3484 SD; P<0.01). Though the magnitude of difference in the mean test score is somewhat smaller than the difference observed among girls, significantly lower test scores are also observed among boys, lower by 38.07 scale score points (-0.3808 SD; P<0.01) at the baseline and by 26.46 scale score points (-0.2646 SD; P<0.01) at end-line surveys. In terms of value-added learning over the school year, we find a positive, statistically significant value-added score for boys (0.1161 SD; P<0.01) in the post-GEQIP-II cohort but not for girls (0.086 SD; P>0.1), suggesting that a large part of the value-added score has been driven by learning achievement made by boys.

Furthermore, rural girls and boys from the post-GEQIP-II cohort performed less at baseline and end-line surveys. Rural girls from the post-GEQIP-II cohort scored 55.41 lower scale score points (-0.55 SD; p<0.01) at the baseline and 45.46 lower scale score points (-0.45 SD; P<0.01) at the end-year survey. Boys from the rural post-GEQIP-II cohort have also shown lower learning levels both at the bassline test (less by 60.66 scale score points or -0.6 SD; P<0.01) and end-line year test (less by 42.31 scale score points or -0.42 SD; P<0.01) than the average score of their peers in the rural pre-GEQIP-II cohort. In terms of value-added learning over the school year, the patterns of the benefits from the GEQIP-II reform are similar to the analysis we made for the whole rural sample, where boys from the rural post-GEQIP-II cohort benefited statistically significant from value-added learning (18.35 scale score points; or 0.18 SD; p<0.05) over the school year. While the learning gain for rural post-GEQIP-II girls is favourable and relatively large in magnitude, it is not statistically significant (9.051 scale points; $p \geq 0.1$). In urban areas, we don't find clear patterns of the value-added scores for both boys and girls. Similarly, we don't find statistically significant differences in urban boys' baseline and end-line test scores. Urban girls have substantial differences in the baseline and end-line test scores, while the difference in value-added between the two cohorts disappears by the end year (Table 6).

Table 6. The impact of GEQIP-II reform on learning and value-added scores using kernel matching by locality and gender

	Boys			Girls		
	1	2	3	4	5	6
	Start Grade 4	End Grade 4	Value-added score	Start Grade 4	End Grade 4	Value-added score
Rural and urban						
ATT	-38.08***	-26.46***	11.61**	-43.45***	-34.84***	8.608
	(7.50)	(7.96)	(5.78)	(5.83)	(8.71)	(5.99)
N	1252	1252	1252	1268	1268	1268
Rural area						
ATT	-60.66***	-42.31***	18.35**	-55.41***	-45.46***	9.954
	(12.65)	(15.93)	(8.95)	(11.14)	(14.67)	(11.33)
N	593	593	593	526	526	526
Urban area						
ATT	-15.91	-14.82	1.090	-27.99***	-18.94**	9.051
	(11.80)	(18.69)	(11.71)	(10.82)	(9.03)	(7.91)
N	659	659	659	742	742	742

Note. ATT = Average treatment effect of the treated from kernel matching; standard errors in parentheses (bootstrapping); *** p<0.01, ** p<0.05, * p<0.1

4.4. Robustness check

Comparing the results in different ways can help us ensure whether the estimated program effects are invariably consistent with the models employed (Khandker et al., 2010). To check the sensitivity of the estimates obtained from the matching analysis, we repeat the analysis using the Inverse Probability Weighting (IPW) method, an effective approach to address selection bias in observational data studies (Carry et al., 2021; Hernán & Robins, 2006). Avagyan & Vansteelandt (2021) note that IPW has become a prevalent method to adjust statistical analyses for bias due to confounding or selection in observational studies.

Table 7 presents the average treatment effect on the treated (ATET) on the three learning outcomes: baseline, end-year, and value-added scores over the school year using the IPW method. The results obtained from this method are similar to the ones obtained from the kernel matching analysis. The first column of the Table reports the average treatment effect on the treated (ATET) baseline test score. The ATET on the baseline test score of the two cohorts is -35.68 scale score points (-0.36 SD; P<0.01). Similarly, the end-line test observes an ATET of -28.60 scale score points (-0.286 SD; P<0.01). In terms of value-added learning over the

school, we find a similar value of ATET (0.071 SD) and remains statistically significant at 5% (P<0.05). We also estimated for rural and urban students separately to further check the robustness of the results at sub-sample levels. The IPW results are similar to kernel matching analysis applied for the whole sample (rural +urban). The estimated effect of the reform on the learning levels and value-added score from the IPW is similar to the ones we obtained from the kernel matching estimation, implying that the results are less sensitive to alternative impact evaluation methods.

Table 7. The impact of GEQIP-II reform learning levels and gains using the Inverse Probability Weighting (IPW) method

	(1)	(2)	(3)
	Start Grade 4 (score)	End Grade 4 (score)	Value-added (score)
Rural and urban			
ATET	-35.68***	-28.60***	7.076**
	(4.14)	(4.64)	(3.49)
N	2799	2799	2799
Rural Areas			
ATET	-50.07***	-36.27***	13.80***
	(5.98)	(6.72)	(4.98)
N	1201	1201	1201
Urban Areas			
ATET	-20.66***	-19.55***	1.113
	(5.69)	(6.40)	(4.87)
N	1598	1598	1598

Notes. ATET is the average treatment effect on the treated from the IWP method; standard errors in parentheses; * p<0.1; ** p<0.05; *** p<0.01.

We also conducted similar estimations separately for gender by locality using the IPW method. The results are similar to the overall analysis we have seen so far. However, the ATET value-added scores of the reform for rural boys appeared to be much larger and statistically significant (0.21 SD; P<0.01) when estimated using IPW (Table 8). This shows that the Average Estimated Effects on the Treated (ATT) are robust and less sensitive to alternative estimation methods at the sub-sample level.

Table 8. The impact of GEQIP-II reform learning levels and gains using the Inverse Probability Weighting (IPW) method by gender and locality

	Boys			Girls		
	(1)	(2)	(3)	(1)	(2)	(3)
	Start Grade 4 (Score)	End Grade 4 (Score)	Value-added (Score)	Start Grade 4 (Score)	End Grade 4 (Score)	Value-added (Score)
Rural and urban						
ATET	-31.87***	-20.94***	10.93**	-37.94***	-32.75***	5.194
	(6.48)	(7.25)	(5.21)	(5.60)	(6.36)	(4.95)
N	1324	1324	1324	1428	1428	1428
Rural area						
ATET	-49.73***	-28.70***	21.03***	-53.31***	-44.14***	9.172
	(9.23)	(10.15)	(7.33)	(8.99)	(10.02)	(7.32)
	613	613	613	569	569	569
Urban area						
ATET	-15.73*	-17.75*	-2.022	-17.81**	-13.99	3.822
	(8.88)	(10.50)	(7.32)	(7.69)	(8.59)	(7.03)
N	711	711	711	859	859	859

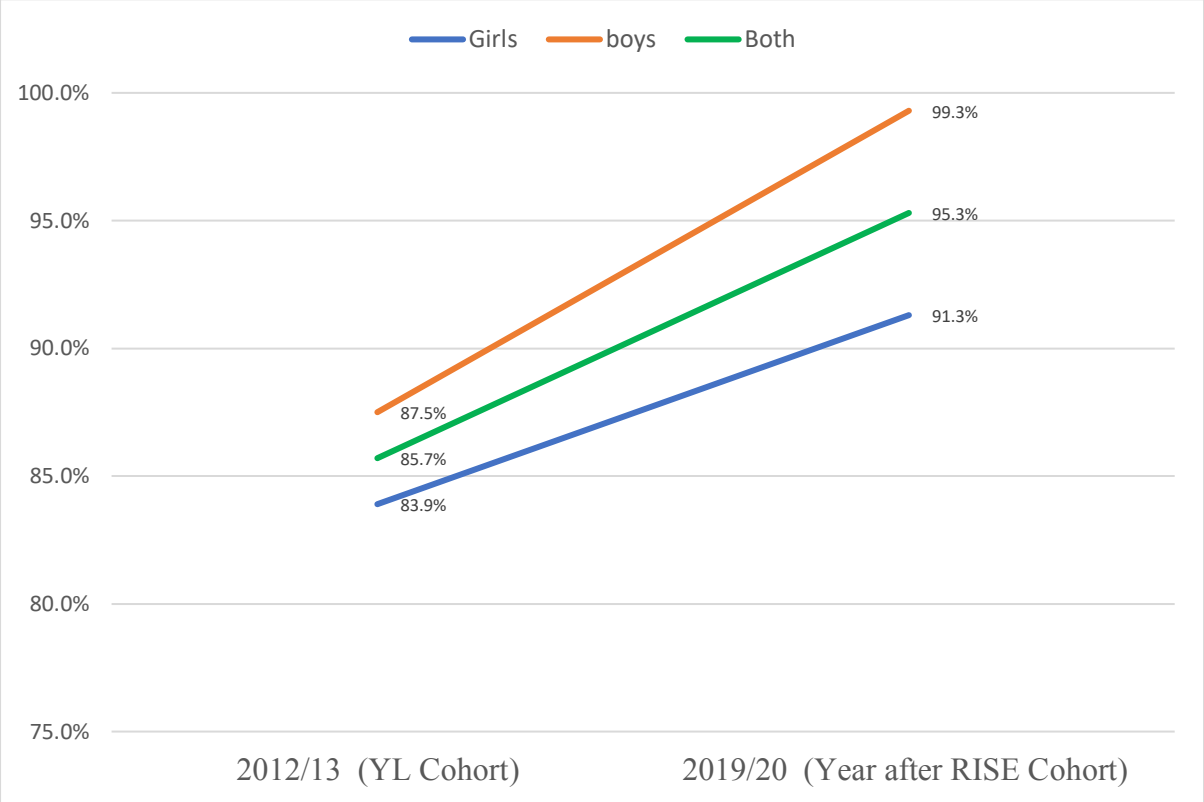
Notes. ATET is the average treatment effect on the treated from the IWP method; standard errors in parentheses; * p<0.1; ** p<0.05; *** p<0.01.

5. Discussion

Although the post-GEQIP-II cohort achieved relatively higher value-added learning over the school year, their average maths test scores are lower than that of the pre-GEQIP-II cohort. It is then important to ask why average maths learning levels are lower for the post-GEQIP-II cohort than the pre-GEQIP-II cohort. Explaining the underlying reasons behind the declining maths learning test scores might be challenging. However, given the dynamic nature of the education sector, the reform might be accompanied by reform-induced rapid enrolment in the country. By reform-induced enrolment, we mean that following the GEQIP-II reform, high enrolment arose from mass mobilisation to send children to schools. We can support this argument with the recent official Net Enrolment Rate (NER). When we look at the NER (Figure 3), we find strong evidence of increasing school enrolment after the reform, where the official NER for Grades 1-8 has shown a steeper increase from 85.7% (with 87.5% for boys and 83.9% for girls) in 2012/2013 to 95.3% (with 99.3% for boys and 91.3% for girls) in 2019/2020 (MoE, 2020). The increase is more than 9.6 percentage points and makes the total number of learners

of appropriate age at the primary school level (Grade 1-8) about 18.6 million[1], up from 15.7 million[2] learners in 2012/13 (MoE, 2020). This implies that an additional nearly 3 million new children joined the Ethiopian primary school between 2012/13 and 2019/20. Indeed, the NER does not show us the actual number of students in the classrooms, as this number is adjusted for ages. When looking at the total number of primary school students regardless of age, we find an even higher influx of students over recent years, where the Gross Enrolment Rate (including overage students) reached 104.9%, accounting for 20.4 million learners at the primary school level in 2019/20 (MoE, 2020).

Figure 3. Net Enrolment Rate (NER) for primary school (Grades1-8) in Ethiopia



Source: Ministry of Education Ethiopia, 2019/20

This might show that the learning level has not increased together with enrolment (Pankhurst, 2017b). The rapid increase in enrolment has not been matched by improvements in learning levels, maybe due to pressure on the school system or resources (Woldehanna & Araya, 2016). If we look at textbook availability and utilisation for the 2017/18 academic year, which is one of the critical indicators of the GEQIP-II reform as measured by the ratio of primary students to newly procured mother tongue textbooks, it was short of reaching its target ratio (1:1) (World

[1] *Actual number is 18,554,222*
[2] *Actual number is 15,708,293*

Bank, 2020). Similarly, when we compare the textbook availability for the two cohorts from our datasets, we find considerable differences in terms of coverage. While three-fourths (75.44%) of the pre-GEQIP-II cohort had their textbook provided per student at school, this was only less than two-thirds (61.83%) for the post-GEQIP-II cohort. Also, additional information from the World Bank (2020) indicates that only 31% of students were nationally bringing their textbooks to classes, while the targets of the GEQIP-II reform were 90% for maths and 70% for science and social science. The shortfall of resources can also be evidenced by the fact that the Implementation Completion Report (ICR) (dated July 26, 2020) for GEQIP-II rated the overall outcome of the reform only as "Moderately Satisfactory" (World Bank, 2020).

Furthermore, the results on lower maths learning levels for the post-GEQIP-II cohort are not surprising given the fact that other studies show similar results in Ethiopia and other developing countries: Using data from Young Lives for comparable 12-year-old children, Woldehanna et al. (2017) found that the percentage of correct maths scores declined by 17.25 percentage point from 2006 to 2013. Similarly, Pankhurst et al. (2017b) compared 15-year-old children who correctly answered similar maths questions in 2009 and 2016. There has been no overall improvement in learning levels over the seven years between the two cohorts. There are similar stories in many other developing countries, particularly in Africa. Using data from the World Bank's Human Capital Project, Evans & Mendez Acosta (2021) reported that harmonised learning test scores (IRT) fell in 18 African countries out of 35 case study countries with data available between 2000 and 2017. The highest decline was for Congo, Dem. Rep (−112 points or -1.12 SD); Madagascar (-83 points or -0.83 SD); Mali (-79 points or -0.79 SD)) and Cameron (-70 points or -0.7 SD) (see also, Angrist, 2019). Binci et al. (2018) also applied an augmented PSM approach to evaluate an education program in Tanzania (EQUIP-T). They indicated that the program did not improve mat learning for pupils in the bottom performance band test. In Indonesia, maths learning declined over 14 years by about a fourth of a standard deviation from 2000 to 2014 (Beatty et al., 2021). There could be several country-specific reasons for declining maths learning levels across countries, but many argue that the learning decline might be accompanied by expanding access to previously inaccessible areas. That is, as children with less preparation gain access to school and participate in tests, average scores could fall even while learning is rising (Evans & Mendez Acosta, 2021). The positive impact of the GEQIP II on the value-added score over the school is similar to Binci et al. (2018), who finds strong evidence that EQUIP-T has improved literacy in Tanzania, particularly for pupils in the bottom performance band of the test or low achievers.

6. Strengths and Limitations

The study has several strengths and some limitations to be acknowledged. This is the first study that attempts to provide insights on how the GEQIP-II reform in Ethiopia might have affected learning outcomes in maths among Grade 4 students using two unique datasets. It uses rich data sets of two Grade 4 cohorts five years apart. Given that the GEQIP-II reform is a non-random assignment, we apply a PSM approach to ensure baseline covariates balance between the two cohorts. As the sampling strategy within the schools slightly differs between the two cohorts, PSM is believed to be an appropriate approach to reduce any bias arising from sampling issues by resampling the dataset when bootstrapped (Khandker et al., 2010).

Nevertheless, despite these strengths, several caveats must be highlighted while using the findings. First, as an observational study, the PSM does not capture any unobserved confounding factors that might have to do with the learning levels of the pupils. It is also true that PSM matches control units to treatment units and not vice-versa. So it is essential to take the findings of this study as a *reflection* of the reforms' impact, but not as a *representative*. What is more, in our analysis, we assume that all the planned GEQIP-II educational reforms (as a bundle or result chain) were successfully implemented across the 33 common schools. But this assumption might be limited as we don't have complete information on the reform's implementation fidelity, which implies a need to take the findings with caution.

7. Concluding Remarks

The Ethiopian education system has been very dynamic over recent years, with a series of large-scale education program interventions, such as the Second Phase of the General Education Quality Improvement Project (GEQIP-II) that aims to improve the learning conditions of students by strengthening institutions in different levels of educational administration (World Bank, 2013). Despite the large-scale education program interventions and considerable investments in the education sector, limited empirical studies assess the program's impact on students' learning outcomes and who benefited from such large-scale educational reforms using data from students, schools, and households. Therefore, this empirical study intends to fill this research gap using two similar repeated cross-sectional data sets collected by YL in the academic year of 2012/13 and RISE Ethiopia in 2018/19. Using these data sets derived from 33 common schools that participated in both the YL and RISE school surveys, we employed Propensity Score Matching (PSM) to estimate the effects of the GEQIP-II reform, which is a quasi-experiment design commonly used in the absence of

randomisation of a program. In matching the YL and RISE Children, we used covariates such as the household socio-economic background of the students, parental literacy status, and children's characteristics, including their time use and how frequently they were absent from school and how long they travelled to school in a typical school day. We matched the two cohorts at the school level. We observed significant differences in the baseline covariates of the two cohorts before matching, but we only kept students who had balanced covariates in the matching analysis. Once balanced, we found a good range of overlapping in common support between the two cohorts, enabling us to estimate the impact of the GEQIP-II reform and the benefits derived from this large-scale education intervention. We also bootstrapped the estimates to obtain valid standard errors and resample the dataset to reduce bias from within-school sampling differences between the two surveys.

Results show that students' average maths learning levels are lower for the post-GEQIP-II cohort when compared to the pre-GEQIP-II cohort. Children surveyed after the GEQIP-II reform have lower maths mean test scores at the start and end of Grade 4. By the time the students joined Grade 4, the average maths test score that the post-GEQIP -II cohorts achieved was lower by 35.78 scale score points (-0.3578 SD; $P<0.01$) than that of a matched group of students in the pre-GEQIP-II cohort. By the end of the year, the average learning difference, however, declined to -28.39 points (-0.2839 SD; $P<0.01$), suggesting that the learning gap between the two cohorts had been narrowed down over the school year and the post-GEQIP-II cohort showed relatively positive learning progress, with ATT of 7.3 scale score points (0.073 SD; $P<0.01$) compared to the pre-GEQIP-II cohort. More importantly, the value-added learning was double for rural post-GEQIP children despite their lower initial average scores, which could be one of the benefits of the GEQIP-II reform for rural children who are less prepared to be in school compared to urban children. All estimated ATT results are also robust to alternative estimation methods.

Entering Grade 4 with a lower learning level for the post-GEQIP-II children than that of the pre-GEQIP-II cohort might imply that improvements in access to education have brought more children with less skill and preparation to school, which is indeed an important positive development by itself. This can be evidenced by the steeper increase in primary school enrolments across the country, where enrolment becomes nearly universal (95.3%) in 2019/20, up from 85.7% in 2012/13, accounting for almost 3 million additional learners nationwide for the primary education sub-sector. This shows that learning levels could be compromised with such unprecedented school access as learning achievement takes time, unlike schooling per se. Moreover, when access to school is mainly created for more new-generation learners (Iyer et

al., 2020), there could be increased pressure on the school system. In such an education system, finding a new cohort of students to fall behind in learning is possible until the system responds appropriately to their learning capacity. Therefore, it is essential to consider how the Ethiopian education system might support new-generation learners attracted to schools through educational reforms.

References

Angrist N., Djankov S., Goldberg P. K., Patrinos H. A. (2019) 'Measuring Human Capital,' The World Bank, 1–46 (no. WPS8742).

Araya, M., Rose, P., Sabates, R., Tiruneh, D.T. and Woldehanna, T. 2022. Learning Losses during the COVID-19 Pandemic in Ethiopia: Comparing Student Achievement in Early Primary Grades before School Closures, and After They Reopened. RISE Insight Series. 2022/049. https://doi.org/10.35489/BSG-RISE-RI_2022/049

Aurino, E., James, Z., & Rolleston, C. (2014). Young Lives Ethiopia School Survey 2012-13. Data Overview Report (Working Paper 134). Oxford: Young Lives.

Austin, P. C., Jembere, N., & Chiu, M. (2016). Propensity score matching and complex surveys. Statistical Methods in Medical Research, 27, 1240–1257.

Austin, P. C. (2011a). An introduction to propensity score methods for reducing confounding effects in observational studies. Multivariate behavioral research, 46(3), 399-424.

Austin, P. C., & Small, D. S. (2014). Using bootstrapping when using propensity-score matching without replacement: a simulation study. Statistics in medicine, 33(24), 4306-4319.

Avagyan, V., & Vansteelandt, S. (2021). Stable inverse probability weighting estimation for longitudinal studies. Scandinavian journal of statistics, 48(3), 1046-1067.

Azevedo, J. P. (2020). Learning Poverty. Measures and Simulations. Policy Research Working Paper 9446

Beasley, E., and Huillery, E. (2017). Willing but Unable? Short-term Experimental Evidence on Parent Empowerment and School Quality. The World Bank Economic Review, 31(2), 531–52. https://doi.org/10.1093/wber/lhv064.

Beatty, A., Berkhout, E., Bima, L., Pradhan, M., & Suryadarma, D. (2021). Schooling progress, learning reversal: Indonesia's learning profiles between 2000 and 2014. International Journal of Educational Development, 85, 102436.

Bashir, S., Lockheed, M., Ninan, E., & Tan, J. P. (2018). Facing forward: Schooling for learning in Africa. World Bank Publications.

Binci, M., Hebbar, M., Jasper, P., & Rawle, G. (2018). Matching, differencing on repeat. Oxford Policy Management, January 2018

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. American journal of epidemiology, 163(12), 1149-1156.

Carry, P. M., Vanderlinden, L. A., Dong, F., Buckner, T., Litkowski, E., Vigers, T., ... & Kechris, K. (2021). Inverse probability weighting is an effective method to address selection bias during high-dimensional data analysis. Genetic Epidemiology, 45(6), 593-603.

Carneiro, P., Koussihouèdé, O., Lahire, N., Meghir, C., and Mommaerts, C. (2020). School Grants and Education Quality: Experimental Evidence from Senegal. Economica, 87(345), 28–51. https://doi.org/10.1111/ecca.12302.

Conn, K. M. (2017). Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of impact evaluations. Review of Educational Research, 87(5), 863-898.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. Journal of Public Economics, 123, 92-110.

Evans, D. K., & Mendez Acosta, A. (2021). Education in Africa: What are we learning?. Journal of African Economies, 30(1), 13-54.

Dedehouanou, S. F., & Berthe, A. (2013). Institutional arrangements and education service delivery in primary schools in Mali. Journal of African Development, 15(1), 189-220.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. Journal of Public Economics, 123, 92-110.

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. Health services research, 49(5), 1701-1720.

Glewwe, P., Krutikova, S., & Rolleston, C. (2017). Do schools reinforce or reduce learning gaps between advantaged and disadvantaged students? Evidence from Vietnam and Peru. Economic Development and Cultural Change, 65(4), 699–739. https://doi.org/10.1086/691993

Goldstein, H. (1997). Methods in School Effectiveness Research. School Effectiveness and School Improvement, 8(4), 369–395. https://doi.org/10.1080/0924345970080401

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. Political analysis, 18(1), 36-56.

Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. Journal of Epidemiology & Community Health, 60(7), 578-586.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. The review of economic studies, 65(2), 261-294.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. Child development perspectives, 2(3), 172-177.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis, 15(3), 199-236.

Hoddinott, J., Iyer, P., Sabates, R., & Woldehanna, T. (2019). Evaluating Large-Scale Education Reforms in Ethiopia. RISE Working Paper (19/034).

Howarter, Stephani. The Efficacy of Propensity Score Matching in Bias Reduction with Limited Sample Sizes. Diss. University of Kansas, 2015.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics, 86(1), 4-29.

Khandker, R. S., Koolwal, G. B., & Samad, H. A. (2010). Handbook on Impact Evaluation: Quantitative Methods and Practices (World Bank Training Series). Washington, DC: The International Bank for Reconstruction and Development, The World Bank. eISBN, 978, 0-8213.

Iyer, P., Rolleston, C., Rose, P., & Woldehanna, T. (2020). A rising tide of access : what consequences for equitable learning in Ethiopia ? Oxford Review of Education. https://doi.org/10.1080/03054985.2020.1741343

Leaver C., Ozier O., Serneels P., Zeitlin A. (2019) 'Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools'. https://www.poverty-action.org/publication/recruitment-effort-and-retentioneffects-performance-contracts-civil-servants

Ministry of Education. (2020). Education Statistics Annual Abstract 2012 E.C. (2019/20). Addis Ababa. Federal Ministry of Education

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., and Rajani, R. (2019a). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. The Quarterly Journal of Economics, 134(3), 1627–73. https://doi.org/10.1093/qje/qjz010

Mbiti I., Romero M., Schipper Y. (2019b) 'Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania', National Bureau of Economic Research, (working paper no. 25903). https://doi.org/10.3386/w25903.

Ministry of Education. (2010). Education Sector Development Programme IV (ESDP IV): 2010/11 - 2014/15. Program Action Plan.

Ministry of Education. (2015). Education Sector Development Programme V (ESDP V): 2015/16 - 2019/20. Program Action Plan. Addis Ababa. Federal Ministry of Education.

Ministry of Education. (2019). Education Statistics Annual Abstract 2011 E.C. (2018/19). Addis Ababa. Federal Ministry of Education.

NEAEA. (2016). Ethiopian Fifth National Learning Assessment of Grades 4 and 8 Pupils. Addis Ababa. National Educational Assessment & Examinations Agency.

Narduzzi, S., Golini, M. N., Porta, D., Stafoggia, M., & Forastiere, F. (2014). Inverse probability weighting (IPW) for evaluating and" correcting" selection bias. Epidemiologia e Prevenzione, 38(5), 335-341.

Niels-Hugo Blunch (2014) Literacy and numeracy skills and education sector reform: evidence from Ghana, Education Economics, 22:2, 209-235, DOI: 10.1080/09645292.2011.597954

Oketch, M., Rolleston, C., & Rossiter, J. (2021). Diagnosing the learning crisis: What can value-added Does analysis contribute?. International Journal of Educational Development, 87, 102507.

Perry, T. (2016). English Value-Added Measures: Examining the Limitations of School Performance Measurement. British Educational Research Journal, 42(6), 1056–1080. https://doi.org/10.1002/berj.3247

Pritchett, L., & Beatty, A. (2012). The negative consequences of over-ambitious curricula in developing countries. Faculty Research Working Paper Series (Working Paper 293). Center for Global Development.

Piper, B., Ralaingita, W., Akach, L., & King, S. (2016). Improving procedural and conceptual mathematics outcomes: A randomised controlled trial in Kenya. Journal of Development Effectiveness, 8(3), 404-422.

Rolleston, C., James, Z., Pasquier-Doumer, L., & Tam, T. N. T. M. (2013). Making Progress. Report of the Young Lives School Survey in Vietnam. Working Paper 100. Oxford: Young Lives.

Rolleston, C., & Moore, R. (2018). Young Lives School Survey, 2016-17: Value-added Analysis in India. Oxford: Young Lives.

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. Annual Review of Statistics and Its Application, 7, 143-176.

Rosenbaum, P. and Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', Biometrika, vol. 70, pp. 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39(1), 33-38.

Scheerens, J., Glas, C., & Thomas, S. . (2003). Educational Evaluation, Assessment, and Monitoring. A Systematic Approach. Taylor & Francis.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1), 1.

Stuart, E. A., Marcus, S. M., Horvitz-Lennon, M. V., Gibbons, R. D., & Normand, S.-L. T. (2009). Using non-experimental data to estimate treatment effects. Psychiatric Annals, 39(7), 41451.

Tiruneh, D. T., Sabates, R., Plavgo, I., Rolleston, C., Kim, J., Bayru, A., Hoddinott, J., Araya, M., & Woldehanna, T. (2021). Understanding Achievement in Numeracy Among Primary School Children in Ethiopia: Evidence from RISE Ethiopia Study (Forthcoming RISE Working Paper).

Tiruneh, D. T., Sabates, R., & Woldehanna, T. (2020). Disadvantaged Schools and Students in Ethiopia: Why is the GEQIP-E reform necessary?

UNICEF. (2017). Education sector budget brief - 2016/17. UNICEF for every child in Ethiopia. https://www.unicef.org/ethiopia/sites/unicef.org.ethiopia/files/2020-01/National_Education_Budget_Brief_2016_17update.pdf

USAID. (2019). Early Grade Reading Assessment (EGRA) 2018 Endline Report. Addis Ababa. USAID.

Woldehanna, T., & Araya, M. (2016). Educational inequalities among children and young people in Ethiopia.

Woldehanna, T., Araya, M. and Pankhurst, A. 2017b. Education and Learning: Preliminary Findings from the 2016 Young Lives Survey (Round 5): Ethiopia. Round 5 Survey Fact Sheet, Oxford: Young Lives.

Woldehanna, T., Araya, M., & Gebremedhin, A. (2016). Assessing Children's Learning Outcomes: A Comparison of two cohorts from Young Lives Ethiopia. The Ethiopian Journal of Education, XXXVI(1), 149–187.

World Bank. (2018). World development report 2018: Learning to realise education's promise. The World Bank. https://doi.org/10.1596/978-1-4648-1096-1

Appendix

Table 1A. YL 2012-13 school survey site descriptions

Region	Cluster ID	Reference in Text	Description of site
Addis Ababa	1	AD1	An overcrowded area in the centre of the capital city, Addis Ababa
Addis Ababa	2	AD2	An industrial area in the southern part of the capital city, Addis Ababa
Addis Ababa	3	AD3	A slum area in the capital city, Addis Ababa
Amhara	4	AM4	A tourist town in the Amhara region, with some extremely poor neighbourhoods
Amhara	5	AM5	A poor rural community in the Amhara region
Amhara	6	AM6	A rural area near Lake Tana in the Amhara region
Amhara	7	AM7	A rural food-insecure area in the Amhara region
Oromia	8	OR8	A rural area near lake Ziway in the Oromia region
Oromia	9	OR9	A drought-prone rural area in the Oromia region
Oromia	10	OR10	A fast-growing town in the Oromia region
Oromia	11	OR11	A relatively rich rural area in the outskirts of Debrezeit town in the Oromia region
SNNP	12	SN12	A densely populated rural area growing enset ('false banana') in the SNNP region
SNNP	13	SN13	A densely populated town in the SNNP region
SNNP	14	SN14	A fast-growing business and tourist town in the SNNP region
SNNP	15	SN15	A coffee-growing rural area in the SNNP region
SNNP	16	SN16	A poor and densely populated rural community in the SNNP region
Tigray	17	TI17	A drought-prone rural area highly dependent on government support in the Tigray region
Tigray	18	TI18	An extremely poor rural area dependent on the Productive Safety Net Scheme and other government support in the Tigray region
Tigray	19	TI19	A small, very poor town in the Tigray region
Tigray	20	TI20	A model rural area in the Tigray region known for its success in soil and water conservation
Somali	21	SO21	A drought-prone area where animal husbandry is the main means of livelihood for the community
Somali	22	SO22	A drought-prone rural area affected by frequent shortages of water and grazing land. Animal husbandry is the main means of livelihood
Somali	23	SO23	An area within the regional capital, Jijiga. Compared with other sites, the economy is stronger, consisting of trade, services, business and government employment
Somali	24	SO24	A drought-prone rural area affected by frequent shortages of water and grazing land. The main means of livelihood for the local community are animal husbandry and farming
Afar	25	AF25	A town about 725km from the capital city, Addis Ababa. A small power station, a health centre and various primary and secondary schools have been constructed since 2005
Afar	26	AF26	A better-off rural area where most households own livestock
Afar	27	AF27	A drought-prone rural area affected by frequent shortages of water and grazing land
Afar	28	AF28	A drought-prone urban area affected by frequent shortages of water and grazing land
Afar	29	AF29	A drought-prone rural area affected by frequent shortages of water and grazing land
Afar	30	AF30	A small urban town densely populated by commercial farm workers and government employees

Source: Aurino et al., 2014.

Table 2A. Percentage corrects for each maths item across the four Rounds.

	Rounds			
Math item number	**Round1 (YL)**	**Round2 (YL)**	**Round 3 (RISE)**	**Round4 (RISE)**
1	91	-	83	-
2	63	66	51	62
3	87	-	75	-
4	89	90	75	-
5	70	73	59	-
6	54	60	40	50
7	67	-	67	-
8	75	77	65	-
9	80	83	59	64
10	71	77	58	62
11	72	79	43	48
12	45	52	35	50
13	33	51	26	41
14	43	-	31	-
15	53	78	51	71
16	52	55	43	43
17	48	55	37	45
18	19	35	25	-
19	18	-	-	-
20	48	45	33	36
21	27	24	-	-
22	18	32	18	
23	33	37	27	32
24	16	-	-	-
25	11	17	-	-
26	-	62	35	42
27	-	83	61	
28	-	55	39	42
29	-	51	35	45
30	-	29	-	-
31	-	18	-	-
32	-	-	-	60
33	-	-	-	35
34	-	-	-	49
35	-	-	-	43
36	-	-	-	57
37	-	-	-	44
38	-	-	-	49
39	-	-	-	38
40	-	-	-	28
41	-	-	-	46

Table 3A. Logit models: estimation of program participation to generate propensity scores

VARIABLES	(1) All Sample	(2) Rural	(3) Urban	(4) Boys	(5) Girls
GENDER	0.00	-0.15	0.09		
	(0.10)	(0.15)	(0.14)		
AGE	0.06*	0.12***	-0.03	0.05	0.09*
	(0.03)	(0.04)	(0.05)	(0.04)	(0.05)
SES_2nd	-0.32**	-0.57***	-0.09	-0.12	-0.47**
	(0.13)	(0.19)	(0.19)	(0.19)	(0.19)
SES_3rd	-1.07***	-1.16***	-0.94***	-0.97***	-1.14***
	(0.15)	(0.30)	(0.19)	(0.22)	(0.22)
PRESCH	0.92***	1.33***	0.53***	1.00***	0.84***
	(0.11)	(0.18)	(0.15)	(0.17)	(0.16)
SCHSTOP	-0.78***	-1.02***	-0.46**	-1.04***	-0.61***
	(0.15)	(0.22)	(0.22)	(0.23)	(0.22)
PCGLITS	-0.14	-0.42***	0.09	-0.17	-0.14
	(0.10)	(0.16)	(0.14)	(0.15)	(0.15)
SCHDIST	0.01***	0.02***	0.00	0.01**	0.02***
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
CHDCHORES	-0.53***	-0.57***	-0.52***	-1.16***	-0.08
	(0.11)	(0.16)	(0.16)	(0.18)	(0.15)
CHWFARM	-1.16***	-0.79***	-2.46***	-0.68***	-1.94***
	(0.21)	(0.25)	(0.60)	(0.26)	(0.42)
CHWPAY	-2.57***	-3.02***	-2.08***	-2.02***	
	(0.40)	(0.61)	(0.53)	(0.42)	
CHSTUDY	0.81***	0.66***	0.97***	0.84***	0.84***
	(0.10)	(0.15)	(0.15)	(0.15)	(0.15)
SCHOOL	**YES**		**YES**	**YES**	**YES**
Constant	-1.93***	-2.51***	-0.96	-1.59**	-2.61***
	(0.44)	(0.75)	(0.61)	(0.64)	(0.66)
Observations	2,807	1,209	1,598	1,354	1,334

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3A. Logit models: estimation o program participation to generate a propensity score- (continued)

VARIABLES	(1) Rural boys	(2) Rural Girls	(3) Urban Boys	(4) Rural boys
AGE	0.11*	0.22***	-0.01	-0.01
	(0.06)	(0.07)	(0.07)	(0.07)
SES_2nd	-0.25	-0.91***	-0.07	-0.07
	(0.26)	(0.30)	(0.29)	(0.29)
SES_3rd	-0.81*	-1.46***	-1.05***	-1.05***
	(0.42)	(0.45)	(0.29)	(0.29)
PRESCH	1.18***	1.61***	0.84***	0.84***
	(0.25)	(0.28)	(0.24)	(0.24)
SCHSTOP	-1.43***	-0.67**	-0.53	-0.53
	(0.33)	(0.34)	(0.32)	(0.32)
PCGLITS	-0.47**	-0.45*	0.09	0.09
	(0.23)	(0.24)	(0.22)	(0.22)
SCHDIST	0.01	0.04***	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
CHDCHORES	-1.23***	-0.00	-1.02***	-1.02***
	(0.25)	(0.24)	(0.27)	(0.27)
CHWFARM	-0.25	-1.88***	-2.14***	-2.14***
	(0.31)	(0.53)	(0.75)	(0.75)
CHWPAY	-2.50***		-1.49***	-1.49***
	(0.64)		(0.56)	(0.56)
CHSTUDY	0.54**	0.92***	1.14***	1.14***
	(0.21)	(0.24)	(0.23)	(0.23)
SCHOOL	**YES**	**YES**	**YES**	**-YES**
Constant	-2.34**	-3.81***	-1.07	-1.07
	(0.95)	(1.11)	(0.94)	(0.94)
Observations	627	530	727	727

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4A. Covariate Balance across Young Lives and RISE cohorts before and after kernel matching at the school level

	Before Matching					After Matching					
	Mean Value					Mean Value					
Variable	Treatment	Comparison	Standardized Difference (%)	t	p>t	Treatment	Comparison	Standardized	% reduce bias	t	p>t
	(n=689)	(N=2,190)				(N=666)	(n=2,133)	Difference (%)			
GENDER	0.49407	0.47867	3.1	0.7	0.486	0.4955	0.50783	-2.5	19.9	-0.45	0.653
AGE	11.052	11.041	0.6	0.15	0.883	11.053	11.144	-5.3	-721.7	-0.94	0.346
SES_2nd	0.2819	0.27801	0.9	0.2	0.845	0.28529	0.2892	-0.9	-0.8	-0.16	0.875
SES_3rd	0.21662	0.37787	-35.8	-7.79	0.000	0.21922	0.22748	-1.8	94.9	-0.36	0.718
PRESCH	0.49407	0.41632	15.7	3.56	0.000	0.49099	0.49909	-1.6	89.6	-0.3	0.768
SCHSTOP	0.10386	0.18894	-24.2	-5.17	0.000	0.10511	0.12097	-4.5	81.4	-0.91	0.361
PCGLITS	0.40801	0.4932	-17.2	-3.87	0.000	0.41291	0.41474	-0.4	97.9	-0.07	0.946
SCHDIST	21.895	18.107	22.5	5.26	0.000	21.857	22.165	-1.8	91.9	-0.31	0.758
CHDCHORES	0.25964	0.38631	-27.3	-6.02	0.000	0.26276	0.26692	-0.9	96.7	-0.17	0.864
CHWFARM	0.04748	0.12658	-28.3	-5.81	0.000	0.04805	0.04799	0	99.9	0	0.996
CHWPAY	0.01039	0.10595	-41.7	-7.92	0.000	0.01051	0.0144	-1.7	95.9	-0.64	0.523
CHSTUDY	0.60089	0.45335	29.9	6.73	0.000	0.5976	0.60401	-1.3	95.7	-0.24	0.811