Africa Education Review

ISSN: 1814-6627 (Print) 1753-5921 (Online) Journal homepage: www.tandfonline.com/journals/raer20

Tracking Learning Outcomes in Ethiopia: How Large-Scale Educational Reform Shaped Student Attainment from 2012/13 to 2018/19

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

To cite this article: Mesele Araya, Caine Rolleston, Pauline Rose, Ricardo Sabates, Dawit Tiruneh & Tassew Woldehanna (17 Jun 2025): Tracking Learning Outcomes in Ethiopia: How Large-Scale Educational Reform Shaped Student Attainment from 2012/13 to 2018/19, Africa Education Review, DOI: 10.1080/18146627.2025.2504043

To link to this article: https://doi.org/10.1080/18146627.2025.2504043

Published online: 17 Jun 2025.

Article

Tracking Learning Outcomes in Ethiopia: How Large-Scale Educational Reform Shaped Student Attainment from 2012/13 to 2018/19

Mesele Araya
http://orcid.org/0000-0003-3411-7461
University of Cambridge, UK
Addis Ababa University, Ethiopia
ma990@cam.ac.uk

Caine Rolleston
https://orcid.org/0000-0002-4564-4331
University College London, UK

Pauline Rose
https://orcid.org/0000-0002-6701-6774
University of Cambridge, UK

Ricardo Sabates
http://orcid.org/0000-0002-1433-5667
University of Cambridge, UK

Dawit Tiruneh
http://orcid.org/0000-0003-4126-5809
University of Cambridge, UK

Tassew Woldehanna
http://orcid.org/0000-0002-0658-4188
Addis Ababa University, Ethiopia

Abstract

This study examines the impact of Ethiopia's General Education Quality Improvement Programme Phase II (GEQIP-II) by comparing the learning outcomes of two grade 4 cohorts, one before and one after its implementation. Using student learning data collected in 2012/13 (pre-reform cohort) and 2018/19 (post-reform cohort), we matched the data at the school level while controlling for baseline characteristics. The results indicate that the post-reform cohort began grade 4 with a lower average test score than the pre-reform cohort. However, over the academic year, the post-reform cohort showed a modestly higher learning gain, especially in rural schools. This suggests that the reform contributed to higher learning progress over the school year, albeit from a lower initial learning level that might be attributed to the unprecedented expansion of education and the influx of first-generation learners into the system driven by the reform itself. These findings highlight the need for policies that sustain learning progress while simultaneously improving access.

UNISA university of south africa PRESS | Routledge Taylor & Francis Group

Africa Education Review
www.tandfonline.com/raer

https://doi.org/10.1080/18146627.2025.2504043
ISSN 1753-5921 (Online), ISSN 1814-6627 (Print)

Keywords: GEQIP-II; educational reform; learning progress; value-added learning; Ethiopia

Introduction

Over the past two decades, access to education has significantly expanded in many developing countries, largely because of coordinated global efforts following the Millennium Development Goals (MDGs), Education for All (EFA), and Sustainable Development Goals (SDGs). In sub-Saharan Africa, for instance, the gross enrolment rate for primary education rose from 82.06 per cent in 2000 to 99.91 per cent in 2019, with enrolment rates among girls nearly matching those of boys (UNESCO 2021).

Nevertheless, despite progress in school enrolment, the quality of education remains a major concern. Many children spend years in school without acquiring fundamental skills, which highlights a persistent learning crisis (World Bank et al. 2022). For example, in 2017, approximately 125 million children worldwide failed to attain basic numeracy skills (World Bank 2018). This challenge is particularly pronounced in sub-Saharan Africa, where learning poverty—defined as the inability to read and understand a simple text by the age of 10—affects up to 80 per cent of children (Azevedo 2020). The severity of this crisis is also evident in country-level assessments. In Uganda, Ghana, and Kenya, 60–80 per cent of grade 2 students could not solve basic two-digit subtraction problems (World Bank 2018).

Similar patterns of low learning levels have been observed in Ethiopia, where students' performance in core subjects consistently falls short of their curricular expectations (Oketch et al. 2021). National assessments conducted by the Ethiopian Ministry of Education in 2011 and 2015 showed that the average test scores in mathematics and reading remained well below the minimum proficiency standards set by the government. Notably, fewer than 10 per cent of students achieved advanced proficiency in both mathematics and reading tests (National Educational Assessment & Examinations Agency 2016). Using Young Lives longitudinal data, Woldehanna et al. (2016) also reported a sharp decline in the accuracy of mathematics answers among 12-year-olds.

In response to these challenges, the Ethiopian government, with support from development partners, launched a series of large-scale educational reforms aimed at improving learning outcomes. One such initiative was the Second Phase of the General Education Quality Improvement Programme Phase II (GEQIP-II), implemented between 2013/14 and 2018/19. This reform sought to enhance the learning environment by improving access to textbooks and school grants, establishing facilities such as libraries and pedagogical resource centres, expanding digital resources such as computers and internet services, and strengthening teacher development programmes (World Bank 2013). The underlying assumption was that with the proper implementation of the education reforms, these interventions would lead to better learning outcomes for all primary school students.

Despite the completion of the GEQIP-II, empirical evidence on its impact remains limited. Similar educational reforms across Africa produced mixed results. Some interventions have been successful in boosting learning outcomes. For instance, an evaluation of Tanzania's Education Quality Improvement Programme (EQUIP-T) found that it reduced the proportion of students in the lowest literacy performance band (Binci et al. 2018). Similarly, a study by Mbiti et al. (2019a) showed that different teacher performance incentive systems significantly improved student test scores in Tanzania. Research in Ghana (Blunch 2014), Kenya (Duflo et al. 2015), Senegal (Carneiro et al. 2020), and Rwanda (Leaver et al. 2021) has also reported positive effects from various educational interventions, including structured pedagogy programmes and teacher incentives (pay for performance). However, some initiatives have yielded little or no improvement. Studies in Mali (Dedehouanou and Berthe 2013), Niger (Beasley and Huillery 2017), The Gambia (Blimpo et al. 2015), and Tanzania (Mbiti et al. 2019b) found that school-based management interventions and school grant schemes did not translate into significant learning gains after two years of intervention.

While Ethiopia's Ministry of Education conducted an exit evaluation of the GEQIP-II in 2019, its focus was primarily on limited indicators, such as teaching effectiveness, textbook availability, school inspections, grant disbursement timelines, and dropouts. The evaluation found that while the reform met targets in reducing dropout rates and expanding school inspections, it fell short of achieving goals related to textbook-student ratios, textbook utilisation, and timely grant allocation. However, a critical gap remains: The assessment did not directly measure students' learning outcomes, nor did it consider data from students, households, and caregivers, who play a crucial role in shaping attainment. This implies that the extent to which the GEQIP-II has shaped learning outcomes in Ethiopia remains less understood, particularly for marginalised groups such as girls and rural students. To address this gap, this study tracked the learning progress of two cohorts of grade 4 students—one assessed before the implementation of the GEQIP-II using data from Young Lives (2012/13), and another assessed after the reform using data from the RISE Ethiopia school surveys (2018/19). By analysing data from the same 33 public primary schools and controlling for socio-economic and child characteristics, we particularly sought to answer the following research questions:

(1) How do grade 4 mathematics learning levels compare before and after the implementation of the GEQIP-II reforms?

(2) What are the differences in mathematics learning gains over the school year between the pre- and post-reform cohorts?

(3) How do the learning levels and gains of marginalised students, particularly girls and children from rural areas, differ between the two cohorts?

The rest of the article is organised as follows. The following section provides an overview of the GEQIP-II reforms, after which the data type, sample, and methods used

in the analysis are described. The main findings of the study are then presented followed by a discussion section, limitations of the study, and concluding remarks.

General Education Quality Improvement Programme (GEQIP): An Overview

Ethiopia, despite being among the world's least-income countries, has allocated significant funding for its education sector over the last two decades, buoyed by international support (World Bank 2020). Notably, in 2016/17, education spending constituted 27 per cent of total government expenditure, surpassing the internationally agreed-upon target of 20 per cent for national budget allocation to education (Iyer et al. 2017).

International development agencies have strengthened their support for education projects in Ethiopia, recognising the crucial role of education in national development (Ministry of Education 2015). One of the most significant initiatives aimed at enhancing learning outcomes in the country was the introduction of the General Education Quality Improvement Program (GEQIP) in 2008/09. This reform was later extended into a second phase, GEQIP-II (2013/14–2018/19), and further expanded into the ongoing GEQIP for Equity (GEQIP-E: 2018/19–2023/24).

With a budget of approximately USD 500 million (World Bank 2020), GEQIP-II focused on improving teaching and learning conditions in all public schools by addressing critical educational needs. Key investments of the reform included increasing the supply of qualified primary school teachers, providing continuous in-service training to enhance teachers' subject knowledge and pedagogical skills, ensuring the availability of textbooks for each subject, and funding school improvement plans. These plans were supported through per capita school grants allocated based on student enrolment (Kelil et al. 2014).

Table 1 provides a summary of the Theory of Change (TOC) underlying the GEQIP-II reform, which aimed to drive systemic improvements in students' learning conditions. The TOC outlines how the reform sought to enhance learning by strengthening key areas, such as curriculum, teacher development, school management, and institutional capacity. As mentioned, key interventions included the provision of teaching and learning materials, pre-service and in-service teacher training, school grants, and improved education management systems. These measures were designed to expand access to quality resources, improve teacher effectiveness, and enhance accountability through data-driven decision making. Additionally, investments in ICT infrastructure and school inspection mechanisms were intended to modernise education delivery and monitoring. Together, these reforms were expected to create a more inclusive, resource-equipped, and well-managed education system, ultimately leading to improved student learning outcomes (World Bank 2020).

Table 1: Theory of Change of the second Ethiopia General Education Quality Improvement Project (GEQIP-II). Source: World Bank (2020)

Components	Project Activities	Intermediate Project outcomes	Outcomes
Component 1: Curriculum, Textbooks, Assessment, Examination, and Inspection	Supply of teaching and learning materials Provision of e-Braille display readers Support assessments and examinations Rollout 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 Programme	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
Component 3: School Improvement Plan	The support of school improvement plan Provision of school grants	The improved availability of operational and learning resources in schools Continuously participatory school improvement	Strengthened institutions at a different educational administration
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 The increased capacity for planning and management (at central, regional, woreda, and school l levels)	
Component 5: Use of Information and Communications Technology	The provision of ICT infrastructure to target the educational targets The development of an integrated M and 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 and E for selected ICT interventions	
Component 6: Programme Coordination, Monitoring and Evaluation, and Communication	Programme Coordination Monitoring and Evaluation Communication	More effective project management, implementation, and communication Timely monitoring of project progress, results, and impact toward institutional strengthening and improving learning conditions	

Under the GEQIP-II, significant efforts have been made to enhance teaching quality by strengthening pre-service teacher education, expanding in-service training, and implementing career development initiatives alongside teacher performance monitoring. According to Vivekanandan and Sonnenberg (2022), between 2013 and 2018, nearly 300 000 teachers received in-service training, while over 100 000 completed pre-service training, far exceeding the reform's initial target. Consequently, the proportion of qualified teachers with appropriate credentials for grades 1–4 increased sharply from 44 per cent in 2012/2013 to 89 per cent in 2018/2019 (Hoddinott et al. 2024).

A key driver of this progress was the equitable allocation of school grants to all public schools, with disadvantaged schools and those with the lowest education indicators receiving larger grants (World Bank 2020). These grants enabled schools to implement School Improvement Plans (SIPs) that prioritised infrastructure upgrades, procurement of learning materials, teacher professional development, and improved accessibility for children with disabilities (Yorke et al. 2022). Additionally, local communities play a critical role in both planning and monitoring grant utilisation, ensuring accountability and alignment with local needs (Chimier and Harang 2018).

Data and Method

Data

This study used two datasets from 33 public primary schools, longitudinally surveyed by Young Lives (YL) in the 2012/13 school year and the Research on Improving Systems of Education (RISE) surveys in 2018/19. We pooled these data to examine how the GEQIP-II reforms shaped student attainment, measured by mathematics learning and value-added scores over the school year.

Sample

The YL survey was conducted twice in school year 2012/13 to assess students' learning outcomes in six regions of Ethiopia. RISE Ethiopia adopted a similar longitudinal design to understand the impacts of the GEQIP reforms on equitable access to quality primary education in seven regions. The number of schools in each region is approximately proportional to the regional population. It included (a) 33 schools from the YL School Survey, (b) schools targeted in the first phase of the GEQIP-E reforms, and (c) a random selection of additional schools.

While the assessment with YL was a census, only 28 grade 4 students were randomly selected from up to two classes in the RISE survey. In both surveys, learning assessments were conducted in two rounds: round 1 at the start of the school year and round 2 towards the end. In round 1, both school and household surveys were conducted. In round 2, the students completed the second set of learning assessments in mathematics.

However, as stated earlier, it is essential to note that despite an equal number of schools in both surveys, the number of students differed as a result of sampling variations within each school. For instance, all grade 4 students were included at each school in 2012/13, whereas only 28 students were randomly selected in 2018/19. As shown in Table 2, we had 2190 sample students from the pre-reform cohort and 689 from the post-reform cohort. To account for sampling differences, we applied school sample weights and accounted for attrition by constructing dropout-adjusted weights at the end of the school year (see Araya et al. 2022a for sample attrition modelling in panel data).

Table 2: Sample students from 33 YL and RISE school surveys. Source: YL 2012/13 and RISE Ethiopia 2018/19

		YL 2012–13	RISE 2018–19	Combined
Sample size (n)		**2190**	**689**	**2879**
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
Sex	Female	1129	351	1480
	Male	1035	338	1373

Note: SNNP refers to the former region of "Southern Nations, Nationalities and Peoples."

Instruments

Comparable mathematics tests were administered at the beginning and end of the school year in both the YL and RISE surveys to measure the learning outcomes. The YL survey included 25 multiple-choice items in rounds 1 and 2, with 19 common items and six unique items in round 2. Meanwhile, the RISE survey adapted mathematics test items from YL, containing 15 common items and 10 unique items in round 2. By combining both surveys, 13 common items were identified across all the rounds. A two-parameter logistic item response theory model was used to assess differences in learning gain between the two cohorts, providing parameter estimates on a common interval scale. This model concurrently estimates item parameters using pooled data from all rounds. Responses to unique items were treated as missing for those who did not receive them. Anchor items were linked to tests, whereas unique items improved the estimate precision for individual tests. This method has proven effective in accurately estimating item parameters for all test takers, particularly when linking scores across periods like ours (see also Araya et al. 2022b). Finally, the estimated results were transformed into scale test scores with a mean of 500 and standard deviation (SD) of 100 for ease of reference.

Analytical Strategy

When a randomised controlled trial (RCT) is not feasible, propensity score matching (PSM) provides an alternative approach for estimating the impact of educational interventions by matching students based on observable characteristics. This method assumes that after matching, the major systematic difference between the groups is exposure to the intervention, thereby reducing selection bias. The post-reform cohort from 33 public schools was treated as the intervention group, whereas the pre-reform cohort from the same schools served as the comparison group. To ensure comparability between the two cohorts, we first estimated the probability of treatment using selected covariates and then applied a matching algorithm to compare the learning outcomes before and after the reform. The probability of treatment is modelled as follows:

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

where t_i is 1 if the student is from the post-reform cohort and 0 otherwise. F(.) is a binary function, x_i is a vector of covariates including child and household characteristics, α is a parameter, and ε_i is an error term. Once covariate balance at the school level was achieved, we compared students with similar socio-economic backgrounds within the same schools before and after the reform using the following average treatment effect on the treated (ATT) method:

$$ATT \equiv E(ML_{i1} — ML_{0l} | t_i = 1) \qquad (2)$$

where ML_{i0} represents the expected mathematics learning outcome for a pre-reform student, and ML_{i1} represents a post-reform student.

Covariate Selection

In terms of covariate selection, we limited our specification to those strongly correlated with learning outcomes rather than the reform itself (see Table 3 for selected covariates). As schools are the same for both cohorts, we did not include school facility variables in the matching model. In addition, these school facilities are part of the GEQIP-II reform packages and are more likely to be affected by the reform (see earlier for the GEQIP-II's TOC). A propensity score that includes covariates affected by the reform can bias results (Imbens 2004).

In our data, there are some covariates which might be affected by the reform, such as preschool participation, school distance, any record of dropout, and time spent on a typical day at home and school, including time spent on domestic chores, farming, working for pay, or studying or doing homework. We conducted several matching analyses with and without these variables, assuming that the reform might have involved these covariates in the first instance. However, their exclusion did not significantly change the results; we then decided to include them in the analyses because they are more likely to be associated with child–home experience than the reform itself. For

example, the time a child uses at home is more likely to be affected by their family background than by the reform. Furthermore, we undertake matching analysis at the school level to improve matching quality using school identifiers, including all subsamples (sex and locality).

Table 3: Covariate balance across the two cohorts before and after kernel matching at the school level

	Before Matching				After Matching				
	Mean value				Mean value				
	RISE cohort	YL cohort	SMD	p>t	RISE cohort	YL cohort	SMD	% reduce bias	p>t
	(n=689)	(n=2190)	%		(n=666)	(n=2133)	%		
Sex (boy=1)	0.494	0.478	3.1	0.486	0.495	0.507	-2.5	19.9	0.653
Age in years	11.052	11.04	0.6	0.883	11.053	11.144	-5.3	-721.7	0.346
SES_2nd	0.281	0.278	0.9	0.845	0.285	0.289	-0.9	-0.8	0.875
SES_3rd	0.216	0.377	-35.8	0.000	0.219	0.227	-1.8	94.9	0.718
PRESCH	0.494	0.416	15.7	0.000	0.491	0.499	-1.6	89.6	0.768
SCHSTOP	0.103	0.188	-24.2	0.000	0.105	0.121	-4.5	81.4	0.361
PCGLITS	0.408	0.493	-17.2	0.000	0.412	0.4147	-0.4	97.9	0.946
SCHDIST	21.895	18.107	22.5	0.000	21.85	22.165	-1.8	91.9	0.758
CHDCHOR	0.2596	0.386	-27.3	0.000	0.262	0.266	-0.9	96.7	0.864
CHWFARM	0.047	0.126	-28.3	0.000	0.048	0.047	0	99.9	0.996
CHWPAY	0.010	0.105	-41.7	0.000	0.011	0.014	-1.7	95.9	0.523
CHSTUDY	0.600	0.453	29.9	0.000	0.597	0.604	-1.3	95.7	0.811

Note: SES_2nd = 2nd wealth tercile, SES_3rd = 3rd wealth quantile, PRESCH = preschool attendance, SCHSTOP = child ever dropped from school, PCGLITS = primary caregiver is literate, SCHDIST = distance to school, CHDCHORES = child ever participated in chores activities, CHWFARM = child ever participated in farm activities, CHWPAY = child ever participated in paid work, CHSTUDY = child studies after school/at home, SMD = standardised mean difference

Balancing the Data

For the matching analysis to work well, a significant number of 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 depicts the kernel density plot of learning before and after kernel matching. The distribution is better balanced after kernel matching than before the balancing effort. The figure shows a sufficient overlap between the observable covariates of the two cohorts after matching. All subsample (sex and rurality) density plots of propensity scores before and after kernel matching within the same schools are available on request.

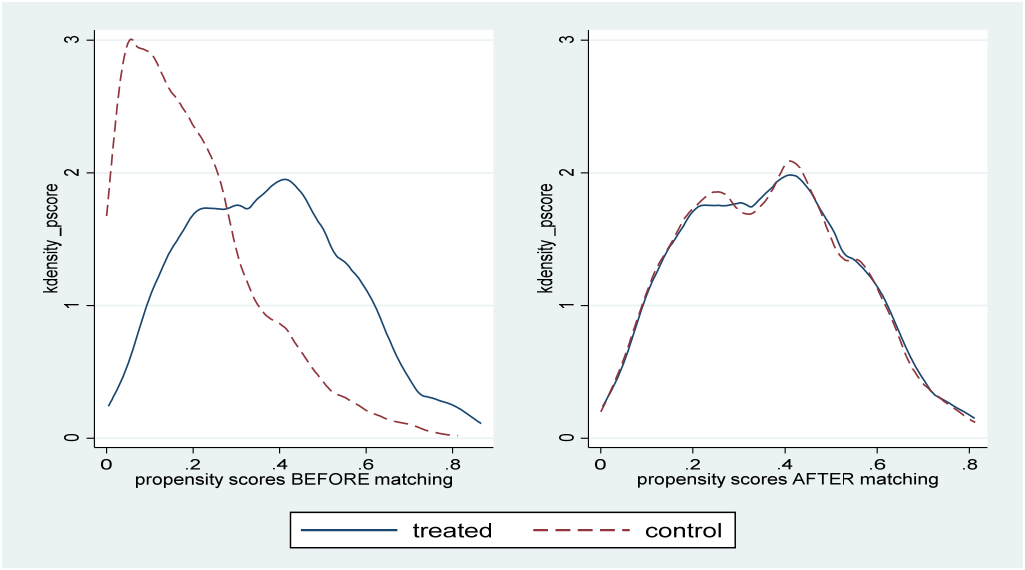

Figure 1: Density plots of propensity scores before and after kernel matching at the school level

The properties of covariate balance can also be substantiated by comparing the standardised difference in the mean values of the covariates before and after the matching analysis. The standardised mean differences in each covariate before and after kernel matching analysis are reported in Table 3. 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 in which children from the pre-reform cohort appeared relatively advantageous and were from higher socioeconomic background families: one-third (37.8%) of the pre-reform cohort were in the third tercile groups of household asset possession compared to only one-fifth (21.6%) for the post-reform cohort. The same was true for primary caregiver status, where a higher proportion of the pre-reform cohort (49.3%) was from literate parents than the post-reform cohort (40.8%). This implies that the post-reform cohort was predominantly first-generation learners. After balancing the baseline covariates, the difference disappeared and became statistically insignificant for the covariates, and a covariate balance between the two cohorts was achieved. Of the 2879 sample students from both cohorts, we found 57 students out of common support and excluded them from the study.

Furthermore, to improve the validity of the standard errors in our estimation, we applied bootstrapping with 100 replicates. This approach accounts for the variance introduced by propensity score estimation and determination of common support, which, if ignored, could lead to misestimated standard errors (Austin and Small 2014; Khandker et al. 2010). By resampling the data, bootstrapping improves the accuracy of the standard error estimation and ensures more reliable results (Imbens 2004).

Main Results

Mathematics Learning Levels and Value-Added Scores for the Whole Sample

Table 4 reports the estimated average treatment effect on the treated (ATT) on mathematics learning levels and value-added scores over a school year from kernel matching estimations. The first two columns provide the ATT on levels at the start and end of grade 4. The average mathematics learning levels of the post-reform cohort are lower than those of the pre-reform cohort both at the beginning and end of grade 4. At the start of grade 4, the post-reform cohort had 35.78 lower scale score points (SD = -0.36; $p < .01$) than the pre-reform cohort. Similarly, by the end of the school year, this was 28.39 lower scale score points (SD = -0.29; $p < .01$) for the post-reform cohort, entailing the mean test score difference between the two cohorts declining by the end of the year.

Next, we ran matching analyses on value-added learning over the school year, which is the difference between the baseline and end-line mathematics scores. The results are presented in the third column. Despite the lower test scores for the post-reform cohort both at the start and end of the school year, the value-added score over the school year is higher for the post-reform cohort. Compared to the post-reform cohort with similar socio-economic backgrounds and matched within the same schools, the post-reform cohort added 7.4 larger score points (SD = 0.074; $p < .05$) higher than their initial mean test score by the end of the school year. This relatively higher value-added score for the post-reform cohort over the school year may be attributed to the GEQIP-II reform, despite the lower initial mean test scores. This may further suggest that the GEQIP-II supported students with lower learning levels during the school year.

Average Learning Levels and Value Added by Rurality

Table 4 also presents ATT estimates for both rural and urban schools. For rural schools, the post-reform cohorts had lower scores than the pre-reform cohort during the baseline survey by 47.81 score points (SD = -0.48; $p < .01$). This suggests that the rural post-reform cohort had significantly lower learning levels at the beginning of the school year than the pre-reform cohort. It is not easy to justify what drove the decline in the initial learning level, but this might be related to the high enrolment induced by the educational reform itself. It is also important to mention that a large proportion of the post-reform cohort were first-generation learners, as seen in the significant difference in parental literacy (Table 3). However, by the end of the school year, the difference in the learning gap appeared to decline significantly to 35.36 score points (SD = -0.35; $p < .01$). This means that when the rural post-reform cohort entered grade 4 in 2018/19, they already had relatively lower skills than the rural pre-reform cohort who joined grade 4 in 2012/13, and much was expected to catch up over the school year.

For value-added scores, we conducted a similar matching analysis at the school level. We found a sizeable improvement in value-added scores for the post-reform cohort. The value-added learning of the rural post-reform cohort is almost double the gain obtained

for the entire sample (rural and urban). This can be seen in column 3 of Table 4, where the reform is statistically significantly associated with the rural value-added learning score over the school year. The rural post-reform cohort achieved 12.45 larger scale score points (SD = 0.125; $p < .01$) over the school year than the rural pre-reform cohort. This is empirical evidence that the reform benefited rural students, at least in terms of progress over the school year.

Similar to the rural cohorts, we also analysed the urban cohorts separately within the same schools to assess the reform's impact on learning levels and value-added scores in urban settings. The results presented in the lower section of Table 4 indicate that the urban post-reform cohort exhibited lower learning levels than their pre-reform counterparts. At the start of grade 4, they scored 24.22 points lower (SD = -0.242; $p < .01$), and by the end of the school year, the gap remained at 20.12 points (SD = -0.201; $p < .01$). However, unlike rural schools, there was no statistically significant difference in value-added scores between the two urban cohorts when controlling for baseline covariates at the school level (SD = 0.041; $p > 0.1$).

Table 4: 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 treatment from kernel matching; standard errors in parentheses (bootstrapping); *** $p < .01$, ** $p < .05$, * $p < .1$.

Learning Levels and Value-Added Scores by Sex and Locality

We further disaggregate our analysis by sex and locality, as this might have important implications for those who have benefited from the GEQIP-II reform. We estimated the ATT in four categories: rural boys, rural girls, urban boys, and urban girls.

Table 5 presents the ATT of the learning levels and value-added scores of these groups. Similar to the whole sample, both boys and girls from the post-reform cohort had lower learning levels at the start and end of their mathematics scores. However, the decline in learning level was higher among girls than boys. At the beginning of grade 4, girls in the post-reform cohorts had 43.45 score points lower than their peers in the pre-reform cohort (SD = -0.43; $p < .01$). Similarly, by the end of the school year, girls in the post-reform cohort experienced 34.84 lower scale score points than their peers in the pre-reform cohort (SD = -0.35; $p < .01$). Although the magnitude of the difference in the mean test score was somewhat smaller than the difference observed among girls, significantly lower test scores were also observed among boys, lower by 38 score points (SD = -0.38; $p < .01$) at the baseline and by 26 scale score points (SD = -0.26; $p < .01$) at the end of the school year. In terms of value-added learning over the school year, we found a positive, statistically significant value-added score for boys (SD = 0.12; $p < .01$) in the post-reform cohort but not for girls (SD = 0.08; $p > .1$), suggesting that a large part of the value-added score was driven by the learning gain made by boys.

Furthermore, rural girls and boys from the post-reform cohort performed significantly less at the baseline and end-line surveys. Rural girls from the post-reform cohorts had 55.41 lower score points (SD = -0.55; $p < .01$) at baseline and 45.46 lower score points (SD = -0.45; $p < .01$) at the end of the survey. Boys from the rural post-reform cohort also showed lower learning levels both at the baseline test (less by 60.66 score points; $p < .01$) and end-line year test (less by 42.31 score points; $p < .01$) than the average score of their peers in the rural pre-reform cohort.

In terms of value-added learning over the school year, the patterns of learning gains from the reform are similar to the analysis we made for the whole rural sample, where boys from the rural post-reform cohort benefited statistically significantly in terms of learning gain (18.35 score points; $p < .05$) over the school year. While the learning gain for rural post-reform girls was favourable and relatively large in magnitude, it was not statistically significant (9.051 scale points; $p \geq .1$). In urban areas, we did not find clear patterns of value-added scores for either boys or girls. Similarly, we did not find statistically significant differences in the baseline and end-line test scores of the urban boys. Urban girls had substantial differences in the baseline and end-line test scores, while the difference in value added between the two cohorts disappeared by the end of the school year (Table 5).

Table 5: The impact of GEQIP-II reforms on learning and value-added scores using kernel matching by locality and sex

	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 treatment from kernel matching; standard errors in parentheses (bootstrapping); *** p < .01, ** p < .05, * p < .1.

Discussion

Ethiopia has recently undertaken a major overhaul of its educational system with the implementation of large-scale reforms such as the GEQIP-II. This comprehensive reform was designed to address long-standing systemic challenges and to improve student learning outcomes across regions. Our study offers new insights into the impact of these reforms by comparing the learning outcomes of grade 4 cohorts before and after the implementation of the GEQIP-II. While the post-reform cohort began grade 4 with lower baseline mathematics test scores, they achieved relatively higher value-added learning gains over the school years, particularly in rural schools. This suggests that despite starting at a disadvantage, the GEQIP-II initiative has positively influenced student progress within the academic year.

Nonetheless, the overall lower average mathematics score in the post-reform cohort warrants further examination. One key factor may be the rapid reform-induced enrolment. The GEQIP-II period saw Ethiopia's net enrolment rate soar from 85.7 per cent in 2012/13 to 95.3 per cent in 2019/20 (see Figure 2). This increase is equivalent to nearly three million additional primary school students into the education system within just a few years (Ministry of Education 2020). This rapid expansion likely strained resources and infrastructure, potentially compromising the quality of instruction, and consequently, learning outcomes (Hoddinott et al. 2024).

Another critical challenge is textbook availability and utilisation, which are direct indicators of reform implementation. Our data on textbooks from the sample schools show that while 75.44 per cent of students in the pre-reform cohort had access to textbooks, this figure declined to 61.83 per cent in the post-reform cohort. Similarly, only 31 per cent of students brought textbooks to class in the academic year of 2017/18 nationally—well below GEQIP-II's targets of 90 per cent for mathematics and 70 per cent for science and social science (World Bank 2020). This shortage of essential learning materials probably exacerbated the challenges posed by rapid enrolment, further affecting overall learning levels. In fact, the distribution of available textbooks from districts to schools has also faced challenges, primarily owing to delayed deliveries by contractors—especially to rural schools affected by poor road infrastructure—and inaccurate data on enrolment, grade levels, and subjects at District Education Offices (Woldetsadik and Raysarkar 2017).

These findings align with the broader regional trends. Similar declines in average learning levels have been observed elsewhere. For instance, Pankhurst et al. (2017) noted stagnant learning improvements in Ethiopia between 2009 and 2016, while Evans and Mendez Acosta (2021) documented falling test scores across several African nations. In Tanzania, Binci et al. (2018) found that the EQUIP-T program did not improve mathematics learning for the lowest-performing pupils. In Indonesia, Beatty et al. (2021) reported a substantial decline in mathematics learning over a 14-year period. A common explanation across these contexts is that expanding access to education, especially to previously underserved populations, can initially lower average test scores as children with less preparation enter the system (Evans and Mendez Acosta 2021).

Despite these challenges, the positive impact on value-added scores observed in our study is promising. This suggests that the GEQIP-II reform may effectively support incremental improvements in learning, particularly for students who start at a disadvantage or who are most likely first-generation learners (Iyer et al. 2020). This evidence, in conjunction with similar findings from other contexts, underscores the importance of complementary policies that not only broaden access, but also strengthen instructional quality and resource allocation (Kelil et al. 2014). In summary, while GEQIP-II has contributed to modest value-added gains over the school year, the dual challenges of rapid enrolment and resource constraints have tempered the overall average learning levels. Our discussion highlights the need for targeted interventions

that maintain the momentum of learning gains while ensuring that the expansion of access does not come at the expense of educational quality.

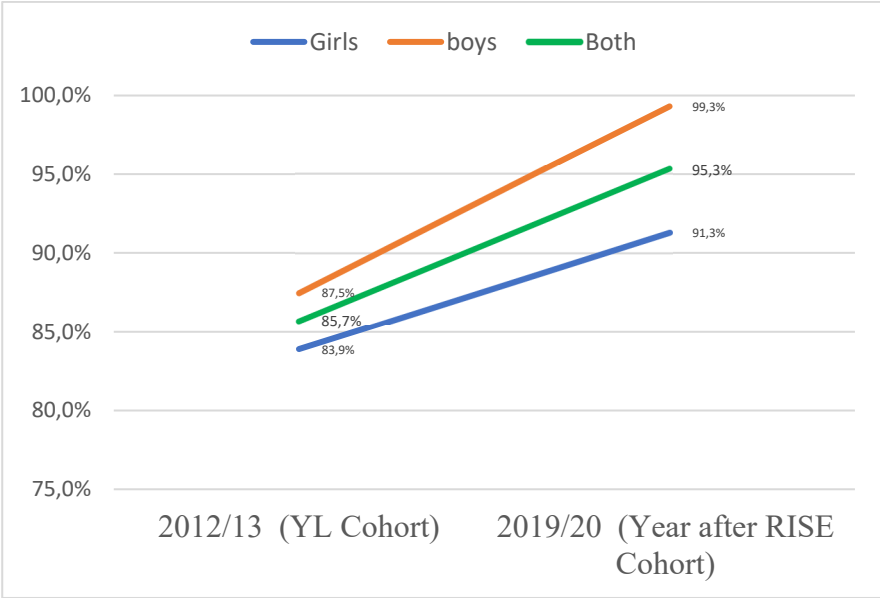

Figure 2: Official net enrolment rate (NER) for primary school (grades 1–8) in Ethiopia. Source: Ministry of Education 2020.

Strengths and Limitations

This study has several strengths and limitations. This is the first study that attempts to provide insights into how GEQIP-II reform in Ethiopia may have impacted learning outcomes in mathematics among grade 4 students, using two unique datasets. It uses rich datasets of two grade 4 cohorts that are five years apart. Given that the GEQIP-II reform is a non-random assignment, we applied a matching method to ensure a baseline covariate balance between the two cohorts. As the sampling strategy within the schools differed slightly between the two cohorts, matching 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).

Despite these strengths, several caveats must be borne in mind when using these findings. First, as an observational study, matching analysis did not capture any unobserved confounding factors that might be related to students' learning levels. It is also true that this method matches control units to treatment units and not vice versa. Therefore, it is essential to consider the findings of this study as a *reflection* of the contribution of the reforms but not as a *causal effect*. Furthermore, our analysis assumed that all planned GEQIP-II educational reforms (as a bundle or result chain) were successfully implemented across 33 common schools. However, this assumption may

be limited because we do not have complete information on the reform's implementation fidelity, which implies a need to interpret the findings with caution.

Concluding Remarks

The Ethiopian education system has undergone significant changes, particularly with interventions such as the GEQIP-II, aimed at improving student learning outcomes. Despite these changes, empirical studies on the impact of such programmes are lacking. This study addressed this gap by analysing data for the 2012/13 and 2018/19 school years. Controlling for several observable covariates and matching at the school level, we compared mathematics learning outcomes in 33 common schools, before and after the GEQIP-II reform. The results show that post-reform students initially had lower mathematics learning levels but made relatively higher progress over the school year compared with pre-reform students. Notably, rural post-reform students showed double learning gains, indicating benefits for rural children in the reform. The increase in primary school enrolment also highlights the challenges in maintaining learning quality amid expanded access. As the Ethiopian education system continues to evolve, it is vital to develop targeted strategies that support first-generation learners and ensure that increased access translates into meaningful educational attainment.

Funding

This study was funded by RISE Ethiopia, supported by funding from the Foreign, Commonwealth & Development Office (FCDO) (RG86520 A29986) and the Bill & Melinda Gates Foundation (INV-023515). The funders had no role in the study design, data collection, analysis, writing, or submission for publication.

References

Araya, M., John Hoddinott, and Dawit Tiruneh. 2022a. "Modelling Sample Attrition in 'Value-Added' Models of Learning." RISE [Blog]. Accessed November 25, 2023. https://riseprogramme.org/blog/modelling-sample-attrition-Ethiopia.html

Araya, Mesele, Pauline Rose, Ricardo Sabates, Dawit T. Tiruneh, and Tassew Woldehanna. 2022b. "Learning Losses during COVID-19 Pandemic in Ethiopia: Comparing Student Achievement in Early Primary Grades before School Closures, and after They Reopened." RISE Insight Series 44. https://doi.org/10.35489/BSG-RISE-RI_2022/049

Austin, Peter C., and Dylan S. Small. 2014. "The Use of Bootstrapping When Using Propensity-Score Matching without Replacement: A Simulation Study." *Statistics in Medicine* 33 (24): 4306–4319. https://doi.org/10.1002/sim.6276

Azevedo, João Pedro. 2020. *Learning Poverty: Measures and Simulations.* The World Bank. https://doi.org/10.1596/1813-9450-9446

Beasley, Elizabeth, and Elise Huillery. 2017. “Willing but Unable? Short-Term Experimental Evidence on Parent Empowerment and School Quality.” *The World Bank Economic Review 31* (2): 531–552. https://doi.org/10.1093/wber/lhv064

Beatty, Amanda, Emilie Berkhout, Luhur Bima, Menno Pradhan, and Daniel Suryadarma. 2021. “Schooling Progress, Learning Reversal: Indonesia’s Learning Profiles Between 2000 and 2014.” *International Journal of Educational Development* 85: 102436. https://doi.org/10.1016/j.ijedudev.2021.102436

Binci, Michele, Madhumitha Hebbar, Paul Jasper, and Georgina Rawle. 2018. “Matching, Differencing on Repeat.” Oxford Policy Management Working Paper. https://www.opml.co.uk/sites/default/files/migrated_bolt_files/wp-matching-differencing-repeat.pdf

Blimpo, Moussa Pouguinimpo, David Evans, and Nathalie Lahire. 2015. *Parental Human Capital and Effective School Management: Evidence from The Gambia.* World Bank Policy Research Working Paper No. 7238. World Bank. https://ssrn.com/abstract=2594083

Blunch, Niels-Hugo. 2014 “Literacy and Numeracy Skills and Education Sector Reform: Evidence from Ghana.” *Education Economics* 22 (2): 209–235. https://doi.org/10.1080/09645292.2011.597954

Chimier, Chloé, and Camille Harang. 2018. “Designing and Implementing a School Grant Policy.” UNESCO IIEP and the Global Partnership for Education. Accessed September 2023. https://policycommons.net/artifacts/8218645/designing-and-implementing-a-school-grant-policy/9132384/

Carneiro, Pedro, Oswald Koussihouèdé, Nathalie Lahire, Costas Meghir, and Corina Mommaerts. 2020. “School Grants and Education Quality: Experimental Evidence from Senegal.” *Economica* 87 (345): 28–51. https://doi.org/10.1111/ecca.12302

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. “School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools.” *Journal of Public Economics* 123: 92–110. https://doi.org/10.1016/j.jpubeco.2014.11.008

Evans, David K., and Amina Mendez Acosta. 2021. “Education in Africa: What Are We Learning?” *Journal of African Economies* 30 (1): 13–54. https://doi.org/10.1093/jae/ejaa009

Dedehouanou, S. F., and A. Berthe. 2013. “Institutional Arrangements and Education Service Delivery in Primary Schools in Mali.” *Journal of African Development* 15 (1): 189–220.

Hoddinott, John, Mesele Araya, Ricardo Sabates Aysa, Tassew Woldehanna, Dawit Tibebu Tiruneh, and Nurullah Eryilmaz. 2024. “Student Composition, Equity, and Mathematics Learning Outcomes During a Time of Educational Reforms in Ethiopia.” *Oxford Review of Education*, September 1–19. https://doi.org/10.1080/03054985.2024.2388047

Imbens, Guido W. 2004. “Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review.” *Review of Economics and Statistics* 86 (1): 4–29. https://doi.org/10.1162/003465304323023651

Iyer, Padmini, Caine Rolleston, Pauline Rose, and Tassew Woldehanna. 2020. “A Rising Tide of Access: What Consequences for Equitable Learning in Ethiopia?” *Oxford Review of Education* 46 (5): 601–618. https://doi.org/10.1080/03054985.2020.1741343.

Khandker, R. S., G. B. Koolwal, and Hussain A. Samad. 2020. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. World Bank Training Series. The International Bank for Reconstruction and Development and The World Bank.

Kelil, K., D. Jebena, D. Dufera, K. Abdi, and S. Wolde. 2014. *The Use and Usefulness of School Grants: Lessons from Ethiopia*. UNESCO and IEEP. Accessed October 2023. https://unesdoc.unesco.org/ark:/48223/pf0000228290

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. 2021. “Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools.” *American Economic Review* 111 (7): 2213–2246. https://doi.org/10.1257/aer.20191972

Mbiti, Isaac, Mauricio Romero, and Youdi Schipper. 2019a. “Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania.” NBER Working Paper No. w25903. National Bureau of Economic Research. https://doi.org/10.3386/w25903

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019b. “Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania.” *The Quarterly Journal of Economics* 134 (3): 1627–1673. https://doi.org/10.1093/qje/qjz010

Ministry of Education. 2015. *Education Sector Development Programme V (ESDP V): 2015/16–2019/20*. *Program Action Plan*. Addis Ababa, Federal Democratic Republic of Ethiopia.

Ministry of Education. 2020. *Education Statistics Annual Abstract (2019/20).* Education Management Information System (EMIS) and ICT Directorate. Federal Democratic Republic of Ethiopia.

National Educational Assessment & Examinations Agency. 2016. *Ethiopian Fifth National Learning Assessment of Grades 4 and 8 Pupils*. NEAEA.

Oketch, Moses, Caine Rolleston, and Jack Rossiter. 2021. “Diagnosing the Learning Crisis: What Can Value-Added Analysis Contribute?” *International Journal of Educational Development* 87: 102507. https://doi.org/10.1016/j.ijedudev.2021.102507

Pankhurst, Alula, Tassew Woldehanna, and Mesele Araya. 2017. *Education and Learning: Preliminary Findings from the 2016 Young Lives Survey (Round 5).* Young Lives.

UNESCO. 2021. *Gross Enrolment Ratio, Primary, Both Sexes (%)*. UNESCO Institute for Statistics (UIS). Accessed May 10, 2025. https://databrowser.uis.unesco.org/browser/EDUCATION/UIS-EducationOPRI/enrol-att/enrolment/IG-GER?highlightGroupId=IG-GER&highlightId=GER.1

Vivekanandan, Ramya, and Krystyna Sonnenberg. 2022. *Supporting Teachers for Quality Teaching: A Hallmark of GPE*. Global Partnership for Education, October 5, 2022. https://www.globalpartnership.org/blog/supporting-teachers-quality-teaching-hallmark-gpe#:~:text=GPE%202025%20prioritizes%20quality%20teaching&text=We%20work%20with%20countries%20to,address%20obstacles%20to%20effective%20teaching

Woldehanna, Tassew, Mesele Araya, and Aregawi Gebremedhin. 2016. "Assessing Children's Learning Outcomes: A Comparison of Two Cohorts from Young Lives Ethiopia." *The Ethiopian Journal of Education* 36 (1): 149–187.

Woldetsadik, Girma, and Chandrani Raysarkar. 2017. *Textbook Provision for All in Ethiopia: Lessons Learned from the General Education Quality Improvement Project (GEQIP).* Working Paper Report No. 116323, Vol. 1. World Bank Group. http://documents.worldbank.org/curated/en/332851497587372108

World Bank. 2020. *Implementation Completion and Results Report on the Ethiopia General Education Quality Improvement Project II*. Report No. ICR00004418. World Bank. Accessed May 10, 2025. https://documents1.worldbank.org/curated/en/141401597014133031/pdf/Ethiopia-Second-General-Education-Quality-Improvement-Project.pdf

World Bank. 2013. *General Education Quality Improvement Project II (GEQIP-II). Project Appraisal Document*. Report No: PAD476. World Bank. Available at https://documents1.worldbank.org/curated/en/882171468257057062/pdf/PAD4760PAD0REP0E0Box379867B00OUO090.pdf

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. The World Bank.

World Bank, UNESCO, UNICEF, USAID, FCDO, and Bill & Melinda Gates Foundation. 2022. *The State of Global Learning Poverty: Update*. The World Bank.

Yorke, Louise, Belay Hagos Hailu, Pauline Rose, and Tirussew Tefera. 2022. *The Pathway to Inclusive Education in Ethiopia: Perspectives of System- and School-Level Stakeholders*. RISE Working Paper Series. 22/115. https://doi.org/10.35489/BSG-RISEWP_2022/115