RISE Working Paper 23/138
March 2023

Which Aspects of Educational Reforms in Ethiopia Have Promoted Equitable Achievements in Mathematics?

John Hoddinott, Mesele Araya, Tassew Woldehanna, Ricardo Sabates, Dawit T. Tiruneh, and Nurullah Eryilmaz

Abstract

This paper assesses the factors underpinning trends in mathematics learning for Grade 4 pupils in Ethiopia based on data collected in 2012-13 (the Young Lives surveys, YL) and 2018-19 (the RISE surveys). It combines comparable data on attainments on tests of students' mathematics knowledge with information on their family background, their teachers, and the schools they attend. The period covered by the study encompasses an education reform, the General Education Quality Improvement Program – Phase II (GEQIP-II). GEQIP-II's goals included increasing access to primary education and the quality of education that was provided.

We find that mathematics teachers' educational qualifications and teacher content knowledge in mathematics improved between 2012 and 2018. Despite this, students in 2018 have learning levels lower than students in 2012. This reflects, in part, changes in the composition of the samples, with the 2018 sample coming from poorer and rural households. However, students in 2018 show greater learning progress over the course of a school year compared to students in 2012. Using a value-added model, we show that learning outcomes in mathematics at the end of the school year are associated with teacher content knowledge, that this association has increased in magnitude over time, and that this content knowledge is especially important for weaker students. The key to reconciling these paradoxical findings is, thus, to recognize that the reforms that underpin these improvements in school quality may well be working in multiple ways, changing the composition of the student body who attend school as well as increasing the amount of learning that takes place in the classroom, particularly for weaker students. This, juxtaposed with the fact that the education reforms implemented over the last decade have sought to both include students from disadvantaged backgrounds and to improve teacher quality, is suggestive of the possibility that the GEQIP initiatives were producing positive learning results.

Keywords: Ethiopia, education, GEQIP-II, mathematics learning, teacher content knowledge

Which Aspects of Educational Reforms in Ethiopia Have Promoted Equitable Achievements in Mathematics?

John Hoddinott
Cornell University

Mesele Araya
REAL Centre, University of Cambridge

Tassew Woldehanna
Addis Ababa University

Ricardo Sabates
REAL Centre, University of Cambridge

Dawit T. Tiruneh
REAL Centre, University of Cambridge

Nurullah Eryilmaz
REAL Centre, University of Cambridge

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:
Hoddinott, J., Araya, M., Woldehanna, T., Sabates, R., Tiruneh, D.T. and Eryilmaz, N. 2023. Which Aspects of Educational Reforms in Ethiopia Have Promoted Equitable Achievements in Mathematics? RISE Working Paper Series. 23/138. https://doi.org/10.35489/BSG-RISE-WP_2023/138

The findings, interpretations, and conclusions expressed in RISE Working Papers are entirely those of the author(s) and do not necessarily represent those of the RISE Programme, our funders, or the authors’ respective organisations.

1. Introduction

While access to education has improved in the last 20 years across many low- and middle-income countries, millions of children continue to complete their schooling without the fundamentals of literacy and numeracy. There are many possible explanations for this phenomenon. For example, first-generation learners may be entering an education system which is not designed for students from disadvantaged backgrounds. Alternatively, the instruction provided by teachers is less relevant for these learners (Glewwe & Kremer, 2006). Still, another explanation are external factors, particularly poverty and marginalization, that result in irregular school attendance and early dropout from school.

Across the world, numerous policies and programmes have been implemented to address these issues, with a particular focus on improving the quality of education provided. Ethiopia is no exception. Despite being a poor country, in Ethiopia financing for the education sector has increased significantly (UNICEF, 2017; World Bank, 2018), supported in part by both bilateral and multilateral donors. Complementing this has been a series of a comprehensive set of government-led, donor-funded changes known as the General Education Quality Improvement Package (GEQIP, 2008-2020) ((GEQIP-I (2008-2012); GEQIP-II (2012-2018); and GEQIP-for Equity (GEQIP-E: 2018-2022)) (World Bank, 2008, 2013, 2017). The most current phase of the reforms (GEQIP-E) concentrated attention on equity issues (earlier GEQIP phases stressed curriculum reform, textbook supplies, and teacher development). This includes an emphasis on the developing regions as well as the difficulties girls, children with impairments, and those from pastoralist households face when it comes to schooling. These increases in funding and reforms have coincided with an expansion of the education sector. Ethiopia's primary education system has grown from three million students in the early 1990s to more than 20 million in 2018–19 (Ministry of Education, 2019). However, a significant portion of children leave elementary school without even the most basic literacy and numeracy abilities (NEAEA, 2016; USAID, 2019) and it is unclear whether the GEQIP reforms have yet had the anticipated impact on system efficiency and educational quality, despite high levels of political backing and donor finance.

The large international research literature on schooling outcomes shows that differences in teacher quality affect students' odds of school completion and attainment; this pattern is consistent across countries (Hanushek et al., 2019). This literature also shows that child, school, and household characteristics are associated with disparities in student test results (Hungi et al., 2017; Iyer et al. 2020). However, to the best of our knowledge, there has been little work in Ethiopia on the school and teacher characteristics that are associated with learning progress over the school year. Moreover, there are no studies to our knowledge that compare the factors that are associated with learning progress over time. Particularly relevant is the examination of teacher and school factors associated with changes in learning outcomes over the school year before and after the General Education Quality Improvement Programme (GEQIP-II) reforms described above.

In this paper, we seek to redress some of these knowledge gaps. We draw on unique data on mathematics knowledge of Ethiopian grade four students. These data were collected in 2012, prior to the implementation of the GEQIP-II reforms and, also in 2018, at the end of the implementation of this phase of the reform. Using a value-added model, we assess what factors – child, household, teacher, and school – are associated with learning. Do these factors, particularly those associated with schools, improve the equity dimension of learning. Lastly, what role did the GEQIP-II reforms play, particularly for low-achieving pupils?

2. Data and methods[1]

Sample

We use data from two sources: (1) The 2012-13 round of the Young Lives (YL) Ethiopia School Survey; and (2) the RISE Ethiopia 2018-19 Household and School Surveys.

The YL data is nested within the broader Young Lives study. Since 2000, Young Lives Ethiopia has traced the lives of children in two age cohorts living in 20 different sites in five regions of Ethiopia. This work has also included surveys of school children living in these localities. The YL 2012-13 School Survey included nearly 12,000 pupils studying in both Grades 4 and 5, in Addis Ababa, Amhara, Oromia, Southern Nations, Nationalities and People's (SNNP), Somali, and Tigray.[2] In addition to the 20 sites included in the ongoing YL study, the 2012-13 School Survey included an additional ten sites. Within each site, all primary schools were included in the School Survey. The survey was conducted in two stages. In October 2012, the survey included a pupil questionnaire; an assessment of mathematics and reading comprehension; an assessment of teacher content knowledge; and a principal questionnaire to gather indicators of school and class quality. In May 2013, students completed a second set of learning assessments in mathematics and reading comprehension.

Our participant flow diagram is shown in Figure 1. We start with the 5,993 fourth grade students in 90 schools that completed the baseline math test at the start of the 2012-13 academic year. It is not uncommon for students in rural Ethiopia to start school at older ages, to repeat grades or to drop out and re-enroll. Less common are children who start school at younger ages. A consequence of all this is that child age in grade 4 can vary markedly. In light of this, we restrict the YL sample to students who started grade 4 in September 2012 and who were aged 9-18 years at that time; this reduces our sample by 166 children. We then drop observations where we are missing child (n=81) or caregiver or household data (n=0), yielding a sample with non-missing child, caregiver, or household data of 5,746 children. Next, we drop children from schools where we are missing data on teachers or on school characteristics. We do not have that data from 20 schools, and so this causes our YL sample to fall to 4,232 students in 70 schools. Of these students, 3,654 took the math test at the end of the school year; the implications of this attrition for our results are discussed below.

RISE Ethiopia used a longitudinal design similar to the YL survey (see Tiruneh, Sabates, et al., 2021). It included Grade 1 and Grade 4 school children from 166 schools, their parents (or primary caregivers), school principals, and Grade 4 mathematics and reading teachers found in seven regions: Addis Ababa, Amhara, Benishangul Gumuz (Be-Gu), Oromia, SNNP, Somali and Tigray.[3] Similar to YL, the 2018-19 RISE Ethiopia surveys were conducted in two phases: Phase 1 at the start of the 2018-19 school year in November, and Phase 2 towards the end of the 2018-19 school year in June. Phase 1 included both the school and household surveys. In Phase 2, pupils completed a second set of learning assessments in mathematics and reading comprehension, and teachers completed a questionnaire and

[1] This section draws heavily on Aurino et al (2014), Oketch, Rolleston & Rossiter (2021); Hoddinott, et.al (2019); and Woldehanna & Pankhurst (2014).

[2] Young Lives is a longitudinal study of childhood poverty conducted in Ethiopia starting in 2002, tracing the lives of children using school surveys (Pankhurst et al., 2018; Aurino et al. 2014).

[3] See Hoddinott, Iyer, Sabates, and Woldehanna (2019) for further details.

an assessment of their mathematics content knowledge. Supplementary Appendix 1 provides more detail on how schools were selected from inclusion into the RISE sample.

We begin with a sample of 4,144 students in 166 schools who completed our math test at the start of the 2018-19 academic year in grade 4 (Figure 2). We drop 93 children who were either younger than nine years of age or older than 18, and we drop 96 observations where we are missing child, caregiver, or household data. We lack teacher or school data for 15 schools, which reduces our sample to 3,635 children in 151 schools. Of these students, 2,977 took the math test at the end of the school year; the implications of this attrition for our results are also discussed below.

Supplementary Appendix 1 describes the sampling process for RISE in more detail. A feature of this process is that, in all regions where both YL and RISE were fielded (Addis Ababa, Amhara, Oromia, SNNP, Somali and Tigray), there are 23 schools that are common to the YL and RISE samples. However, there are marked geographic differences in the distribution of schools in YL and RISE. We highlight two here: a clear majority of the Young Lives sample are located in urban areas (71 percent), compared to 34 percent in RISE; the RISE sample includes a much larger proportion of students (27 percent) in Ethiopia's most populous region (Oromia) than in YL (12 percent). We discuss the implications of this at multiple points below.

Survey instruments

Both YL and RISE included several survey instruments. We describe these here.

The core survey instrument was a mathematics tests administered to grade four students. The 2012-13 YL survey included 25 multiple-choice items in each test phase, i.e., at the start and end of the 2012-13 school year. The endline test included 19 common (anchor) items from the baseline test and six that were unique. The 2018-19 tests were adapted from the 2012-13 versions. The baseline test contained 25 items. The endline test, administered at the end of the 2018-19 academic year, included 15 common (anchor) items from the baseline test and 10 items that were unique. Across the 2012-13 and 2018-19 surveys together, there were 13 items common across the four tests that were administered to grade four students.

School principal and teacher questionnaires were administered in both the YL and RISE surveys. The principal questionnaires in both rounds focused on gathering information on the school principals' levels of education, and their experience, as well as on indicators of school quality: pupils' access to educational resources (library, textbooks, computers, radio, working toilets, access to school grants, etc.). The teacher questionnaire focused on gathering data related to the teacher's age, experience, education levels, and teacher training qualifications. In both YL and in RISE, grade four teachers were administered a mathematics content knowledge test. In YL, the test consisted of 30 items; 20 of them were re-administered to grade 4 teachers in the 2018-19 survey.

Lastly, in both YL and RISE, parents of children who took the math content tests were administered a short household survey. Topics included household demographics, livelihoods, access to public assistance, and assets.

Methods

We begin by describing the variables used in our analysis, starting with our outcome of interest, student scores on the mathematics tests described above. To enable comparable estimations of learning levels over the school year between the 2012-13 and 2018-19 cohorts, we employed a concurrent calibration approach in an Item Response Theory (IRT) modelling framework. A two parameter-logistic IRT model (2PL IRT) was fitted to the item responses, providing parameter estimates on a common interval scale. In concurrent calibration, item parameters are estimated simultaneously using pooled data from all four tests, with responses to the items that were unique to each group treated as missing for respondents that did not receive them. The anchor items provide the link between tests, while the unique items increase the precision of estimates for individual tests. As we did in earlier work (Rolleston et al., 2013; Tiruneh, Hoddinott, et al., 2021), we transformed the pupils' latent trait estimates for the entire pooled sample to a scale with a mean of 500 and a standard deviation (SD) of 100. We note that in both the tests administered to students, items functioned well across all tests with acceptable item difficulty and item discrimination indices.

Student learning may differ by age and sex and so we include these child characteristics in our analysis. We do not have data on school attendance throughout the academic year, but we do know whether students had previously attended pre-school and whether their education has been disrupted in the past, specifically whether they had been absent for more than three months when completing a previous grade. Given the well-known relationship between household economic status and learning outcomes (Baker et.al., 2002; OECD,2010), we control for household wealth. We take responses to questions on ownership of consumer durables that were asked in the household surveys implemented in both Young Lives and RISE. Pooling the Young Lives and RISE data, we use Principal Components to create a household wealth index with mean zero and standard deviation equalling one. Lastly, we account for whether students live in an urban or rural area and whether their caregiver was literate.

Our teacher questionnaire provided data on years of teaching experience, education, and teaching qualifications and whether the teacher's education included specialization in mathematics. Pooling the Young Lives and RISE data on teachers, we use Principal Components to create an index of teacher qualifications with mean zero and standard deviation one. Using the IRT methods described above, we calibrate a common scale metric for teacher knowledge items across both YL and RISE surveys. We take this score, subtract the mean, and divide by the score's standard deviation to create a z-score of teacher content knowledge.

Our principal and school questionnaire provided data on principals' characteristics (years of experience as a principal at the school, education, and teaching qualifications). We pool the Young Lives and RISE school-level data and use Principal Components to create an index of principal qualifications with mean zero and standard deviation one. We create an index of school infrastructure quality using Principal Components, the Young Lives and RISE data and the following school characteristics: whether the school (a) provides grade four students with their own math textbook; (b) has a functioning library; (c) internet access; (d) a working radio used by students; (e) a functional pedagogical centre; (f) is a full-shift school; (g) has separate toilets for boys and girls; and (h) source of drinking water. We also include the following continuous variables: the number of working computers available for students; and the number of working toilets. As part of the school questionnaire, we asked about grade 4 class sizes.

Modelling

We estimate a value-added model of learning (Hoddinott, et al, 2019).

$$Y_{i,E,S} = \alpha + \beta_{i,B,S} \cdot Y_{i,B,S} + \beta_{C,I,B} \cdot C_{i,B,} + \beta_{TE,I,B,S} \cdot TE_{i,B,S} + \beta_{SC,I,B,S} \cdot SC_{i,B,S} + \varepsilon_{i,E,S} \quad (1)$$

where Y is the math score of child i in survey S, where E and B refer to the end-of-school and beginning-of-school math test scores and survey S refers to the Young Lives and RISE surveys respectively. C is a vector of child characteristics (age, sex, whether previously stopped schooling for three months or more, the household wealth index and whether the child lives in an urban area). TE and SC are vectors that capture teacher (TE) characteristics (Teacher math content knowledge, z score; Math teacher qualification index) and school (SC) characteristics (Principal qualification index; School infrastructure index). The β's are parameters to be estimated. Lastly, $\varepsilon_{i,E,S}$ is an iid distriburance term.

We estimate (1) using Ordinary Least Squares with the standard errors clustered at the sampling unit, the school. We present these results for the full samples; we also disaggregate by child sex. As is well known, the parameter estimates that are generated are based around the conditional mean of the outcome (that is, the predicted value of the outcome given the mean value of our regressors). But given our interest in understanding the distributional effects of efforts to improve primary school education in Ethiopia, using quantile regression methods we also estimate equation (1) at a series of conditional quantiles (the 10th, 25th, Median, 75th, and 90th percentiles) of the distribution of the outcome variable. For these estimates, standard errors are based on bootstrap methods with 1000 replications (Davison & Hinkley,1997, Efron and Tibshirani 1993; Horowitz, 2003). Our main variables of interest are teacher math content knowledge, math teacher qualification, school infrastructure, and principal qualifications.

3. Results (1): Descriptive statistics and sample attrition

Descriptive statistics

We begin with student characteristics; then the characteristics of their teachers and schools before turning to the outcome variables, math content test scores.

Table 1 shows student characteristics by survey round. Student average age is 10.9 years in YL and 11.2 years in RISE, a reflection of the fact that there are a larger percentage of older children, that is children 12 or older, in the RISE sample (38.0) compared to YL (28.3). There are slightly more girls than boys in YL with the opposite true in RISE. Children in YL were more likely to have had their schooling interrupted in the past. In YL, most children (71.4 percent) reside in urban areas compared to RISE (34.3 percent). In RISE, we have a greater fraction of children living on Oromia region and unlike YL, we also have children living in the western, more remote region of Benishangul Gumuz. Consequently, relative to YL, there are fewer children in the RISE sample who live in Addis Ababa, Amhara, SNNP or Tigray. Given this different geographic distribution (by urban/rural and by region), it is not surprising that relative to the RISE sample, children in YL are more likely to have a literate caregiver and to live in wealthier households. All these differences in characteristics are statistically significant.

Table 2 describes the characteristics of students' teachers by survey round. Relative to YL, teachers in RISE have considerably less teaching experience (4.6 v 11.0 year), are less likely to hold some form of post-secondary certificate, diploma or degree but are more likely to have a higher teacher training qualification. They are also more likely to have an education that includes a specialization in mathematics (57.1 v 49.2 percent in YL). Consequently, teachers in RISE score higher on our Teacher

Qualification Index than teachers in YL. They also do better on the math content test, the difference between the mean values for YL and RISE teachers is 0.38 standard deviations.[4] We wondered if this reflected the presence of a small number of RISE teachers who scored highly on the math test and/or a small number of YL teachers who scored poorly. For both surveys, we generated kernel density functions for teachers' math content knowledge. These show that the distribution of scores for RISE teachers is right-shifted relative to that of YL, indicating that in general, teachers found in the RISE survey were more knowledgeable about the math content they were teaching than were YL teachers (Figure 3). All differences in teacher characteristics are statistically significant.

Table 3 describes characteristics of students' principals and schools, by survey. When looking at these, it is again important to recognize that the geographic distribution of the YL and RISE samples are different with RISE having many more pupils residing in rural areas and in regions, such as Benishangul Gumuz which historically have been under-resourced, particularly relative to urban schools. Table 3 shows that while differences in the mean Principal Qualification Index are statistically significant, the magnitude of this difference is not large with principals in the YL sample having attained more general education but less teacher training qualifications. Looking at physical school characteristics, there is no obvious pattern in the data, but it is noteworthy that children in the YL survey were more likely to have their own math textbook (81.4 percent) compared to pupils in the RISE sample (45.6 percent). This, and other differences, results in students in the YL sample having access to schools with a higher School Infrastructure Quality Index score.

For our final descriptives, we consider our outcome variable, end-of-school (or endline) math content scores for Grade 4 students in the YL and RISE samples. As Table 4 shows, the mean values for these are 540.8 and 492.5 respectively. However, as Table 4 also shows, the start-school scores for RISE are also lower than those found in YL, 459.5 v 512.4 and a test rejects the null that these baseline means are equal. The lower baseline scores for RISE can also be seen in Figure 4 where we show, by survey, the kernel density functions of the math content scores. Recall that these are calculated so that across the entire pooled sample, they have a mean of 500 and a standard deviation of 100. The mass of the distribution of the RISE baseline score lies to the left of this mean (and a Kolmogorov-Smirnov test rejects the null that the distributions are equal). Again, this is not entirely surprising given that the RISE sample is poorer and more rural.

Attrition

Not all students who took the baseline math tests completed the endline test at the close of the school year. Attrition between baseline and endline samples was 13.6% for YL 2012-13 and 18.1%[5] for the RISE 2018-19 sample. Reasons for attrition included: absenteeism at the time of test administration;

[4] Note that these are calculated based on the sample of students; that is to say, what are the qualifications of the teachers of the average student in YL or RISE. This is not the same as the average characteristic of a teacher because class sizes differ across and within our two surveys. However, if we calculate teacher averages, similar patterns emerge. For example, the average teachers z score for math content knowledge is -0.35 in YL and 0.17 in RISE.

[5] These are calculated as the number of observations with missing endline test scores divided by the number of children who took the baseline test and are not missing child, caregiver, household, teacher or school data; see Figures 1 and 2 for additional details.

dropouts; a change of school because parents relocated to other areas; and a failure to track some pupils due to lack of proper class rosters in the surveyed schools.

Table 5 compares mean pupil, teacher and school characteristics by attrition status and survey round. We highlight the following. In the YL survey, for nine of the 13 characteristics we consider, we reject the null of equality of means between attritors (those who did not completed the endline math test) and completors. In RISE, we reject the null of equality of means for 10 of these 13 characteristics. Second, the mean scores of children who attrit in either survey is less than the mean score of completors; in other words, children who were weaker academically (in terms of their math score) were less likely to complete the endline math test. Third, in both the YL and RISE samples, older children, boys, children from poorer households and those with caregivers who were not literate were more likely to attrit – all characteristics that are associated with lower test scores at baseline. Fourth, apart from class size (larger classes are associated with higher attrition, especially in RISE), there is no clear pattern in the association between attrition and teacher or school characteristics.

4. Results (2): Correlates of endline test scores

Results of estimating equation (1) are reported in Table 6, by survey round. We note the following.

For our YL data, our included variables explain just over half of the variation in endline test scores (R^2=0.54). Endline scores are highly correlated with baseline test scores – every additional point (technically, SD) of baseline math score is associated with an increase of 0.77 points at endline.[6] Conditional on baseline test scores, there is no statistically significant association with most child (age, sex), household (wealth), teacher (math content z score) and school characteristics (class size, school infrastructure, principal qualifications). Among our main interest variables, only math teacher qualifications are statistically significant with endline scores. Urban students have higher scores on the test administered at the end of the school year as did children who had ever attended pre-school; students who had dropped out in the past had lower scores as did children whose caregiver was literate.

The covariates included in equation (1) also explain just over half of the variation in endline scores (R^2=0.541) in the RISE survey (this falls to 0.19 when baseline test scores are excluded). Endline scores are correlated with baseline test scores, but slightly less so than in YL with every additional point of baseline math score associated with an increase of 0.73 points at endline. But unlike YL, the associations between teacher characteristics (math content knowledge and qualifications) and student learning are large and statistically significant. In RISE, increasing both teacher math content knowledge and qualifications by 1 SD is associated with increasing math learning scores by 20.4 points. (Note that the mean difference in scores between all students who sat the baseline test and the students who sat the endline test is 36 points).

We might worry that our associations might be biased because of the non-random attrition described above. In RISE, we can account for this by re-weighting our data; when we do, again our results are largely unchanged (see Supplementary Appendix 2). As we explain in Supplementary Appendix 3, our ability to account for non-random attrition in YL is somewhat constrained by the more limited set of variables available from which we can construct attrition-weighted estimates. With the

[6] Note that if we drop baseline scores from the regression, the R2 falls to 0.15.

caveat that our ability to account for factors that affect attrition (but not outcomes) is more limited in the Young Lives sample, attrition does not appear to be biasing these associations. We (cautiously) note that one reason for this apparent absence of bias may be a consequence of our ability to control for baseline test scores.

Next, we disaggregate our results by child sex, focusing on associations with teacher and school characteristics. The coefficient plots are reported in Figure 5. These show that in YL, across all teacher and school characteristics, there is no statistically significant differences in association by sex. In RISE, teacher characteristics have the same association with endline scores for girls and for boys, but improved school infrastructure is associated with greater learning for girls.

As noted in the methods section, we estimate quantile regressions at the 10th, 25th, median, 75th and 90th percentiles of the distribution of the outcome variable, math content test scores at the end of the school year. As Table 7 shows, in the YL data there are no differences in parameter estimates for any characteristic – pupil, teacher, school – across quantiles. This is not the case for the RISE survey. We highlight two interlinked associations.

First, the association between teacher content knowledge and endline scores falls in magnitude as we move from pupils at the bottom of the distribution of endline math scores (the 10th and 25th percentiles) to pupils at the top of the distribution (the 75th and, most notably, the 90th percentiles). The point estimates fall from 12.0 to 3.8 (10th v 90th percentiles) and from 10.6 to 3.8 (25th v 90th percentiles) and both differences are statistically significant with P =0.01 and P=0.03 respectively. Second, the association between baseline and endline math scores is smaller in magnitude at the 10th and 25th percentiles than it is for children at the median or higher. With the caveat that these are associations, one way of thinking about this is to say that the coefficient on the baseline score captures the persistence of past teaching on current (ie end-school year) learning. (Put colloquially, the coefficient on baseline scores captures the idea that “history is destiny”). A lower coefficient indicates lower persistence of past scores, or put differently, greater scope for children at the lower end of the distribution to catch up to their peers (for example, if they have teachers with better content knowledge).

5. Results (3): Extensions. Quantile regressions and disaggregations

We extend our analysis in several ways. First, using the RISE data, we estimate quantile regressions by child sex (Table 8). For girls, these show that the association between teacher content knowledge and endline scores falls in magnitude as we move from girls at the bottom of the distribution of endline math scores (the 10th and 25th percentiles) to pupils at the top of the distribution, most notably the 90th percentile. The differences in the point estimates between girls at the 10th and 90th percentiles, and at the 25th and 90th percentiles, are statistically significant. For boys, this decline is less marked and the differences across quantiles are not statistically significant. The coefficients on teacher qualifications tend to be larger for boys than for girls. For example, at the 10th percentile, increasing both teacher content knowledge and qualifications by one standard deviation is associated with a 17.6 increase in endline test scores for boys compared to 5.1 points for girls. Unlike teacher content knowledge, for both boys and girls, there are no statistically significant differences in associations between endline test scores and teacher qualifications. For boys at the 10th percentile of test scores, larger class sizes areassociated with reduced test scores.

We next disaggregate by household wealth, dividing each sample between children in the poorest third of the wealth distribution (as measured by the wealth index) and children in the remaining terciles (Figure 6). In RISE, but not YL, teacher content knowledge has a larger impact on test scores for children in the poorest households and the difference in these associations is statistically significant. For other teacher, school infrastructure and principal characteristics, in both YL and RISE, we do not reject the null hypothesis that associations with endline math test scores are equal between children in the poorest and less poor households.

Lastly, in both YL and RISE, disaggregating by location (urban/rural) does not show differential associations between teacher and school characteristics and endline math scores (Figure 7).

6. Results (4): Extensions. Do sampling differences confound?

We begin here by re-iterating our seemingly paradoxical finding that scores on math tests by grade four students are lower in the RISE survey than they are in Young Lives (YL) even though the quality of teachers and schools (as measured in a variety of ways) improved between 2012-13 and 2018-19. However, as noted in section 3, there are differences in the sampling that underpins our YL and RISE data. Children in RISE households were more likely to live in poorer, rural households. Further, we have more children per school in our YL survey than we do in the RISE survey. Given this, a “sampling critique” would argue that no such paradox exists. Only some of the schools included in YL were also included in RISE, schools in RISE are systematically different from those in RISE and, thus, it is this difference in school inclusion that accounts for the drop in scores between 2012-13 and 2018-19.

We begin addressing this concern by taking our data on schools and students and dividing it into four groups:

- Schools (and students) who only appear in YL (Group 1). For these schools, we only have data for 2012/13
- Schools (and students) who were surveyed in YL but, subsequently were also included in RISE (Group 2). For these schools, we have data from 2012/13
- Schools (and students) who were surveyed in both YL and RISE (Group 3). For these schools, we also have data for 2018/19. These are the same schools as found in Group 2
- Schools (and students) who only appear in RISE (Group 4). For these schools, we only have data for 2018/19.

After accounting for schools for which we are missing data, there are 48 schools in Group 1, 22 schools in Groups 2 and 3, and 129 schools in Group 4.

With these grouping, we consider baseline (start of school year) and endline (end of school year) math scores. Recall, and as Table 9 shows, our “headline” result was that test scores had declined between 2012/13 and 2018/19, so much so that the mean endline score in 2018/2019 lay below the mean baseline score in 2012/13. This can be seen here by comparing the endline score for all RISE schools, 492.6 (Row 4, Column D) with the baseline score for all YL schools, 512.0 (Row 1, Column A).

Some of this looks like it is a consequence of the schools that either only appeared in YL or in RISE. ***However***, if we look at the schools that appear in both YL and RISE (Groups 2 and 3), we still see this pattern. At endline, Group 3 schools (surveyed as part of RISE) had a mean score of 494.3 (Row 5, Column D) compared to a mean baseline score of 497.0 in YL (Row 3, Column A). In other words, when we look at the ***same*** schools, but surveyed in ***different*** years, we continue to see this decline, it is just not as marked as when we simply compare all YL schools to all RISE schools.

Next, consider the characteristics of the students, teachers, and schools with a particular focus on Groups 2 and 3 - the same schools but at different points in time (Table 10). The key finding here is that for this common set of schools, teacher quality (as measured content knowledge and qualifications) goes up by a lot (for example, compare Row 3, Column A with Row 5, Column A) though there is a slight decline in school quality and principals' qualifications.

Now look at students' characteristics. Again focusing on Groups 2 and 3, there is no meaningful difference in ages or child sex. However, children in Group 3 schools come from households that are less wealth off than children in Group 2. Children in Group 3 are also less likely to have a caregiver who is literate. This is consistent with the drop in baseline math scores between 2012/13 and 2018/19.

Next, we run our regression models for these three sub-samples (Figure 8). We need to be a bit careful here because we are taking our existing samples and dividing them into smaller groups; consequently, all other things being equal, we should expect to see larger confidence intervals. With that noted, we look at the associations between endline test scores and teacher and school characteristics, controlling (as before) for baseline scores and child and family characteristics.

We start with teacher content knowledge. The magnitude of the association is highest for the samples (Groups 3 and 4) where, on average, children come from the poorest backgrounds. When we restrict ourselves to the same schools but observed in different years (Groups 2 and 3), as the school population becomes more disadvantaged, the magnitude of the association between teacher content knowledge and learning gains increases. This is consistent (and arguably further strengthens) our contention that for pupils who are lagging behind in learning mathematics as a subject, students from the poorest backgrounds and more disadvantaged areas, having a teacher who knows their subject matters.

Associations between math qualifications and learning outcomes appears roughly similar across all four groups. (Note that because our sample sizes are now much smaller, we do not test for differences across groups.) There is a suggestion that larger class sizes reduce learning outcomes (the parameter estimates for Groups 3 and 4 are much larger in magnitude that Groups 1 and 2) but the confidence intervals are so wide that not too much should be read into this. Finally, as before, there are no associations between school infrastructure and learning outcomes and no associations between principal qualifications and learning outcomes.

We note two additional issues that arise from the sampling critique. First, more children were surveyed per school in YL than in RISE. If it also the case that YL schools are simply larger than RISE schools and *if* student performance is affected by school size, not just class size (which we control for), then this will introduce an additional difference between YL and RISE schools. As we do not know school size, we note this concern as a caveat. Second, because we have more students observed in some schools than others, and because we observe more students per school in YL than we do in RISE, we might be concerned that an over-representation of students from schools where more students were surveyed will skew our results. To address this concern, we construct sample weights that account for the differences in the number of students surveyed per school. When we apply these weights, our results change only marginally (available on request).

7. Discussion

In less than 15 years, Ethiopia's educational system went from being relatively privileged to one emphasizing the education of all children. The increases in enrollment, particularly of children from disadvantaged backgrounds, is a noteworthy accomplishment in and of itself. Despite this, how to provide quality education for all, especially for underprivileged and marginalised children, in a system that has experienced such rapid development, is a crucial concern not just for Ethiopia but for many other low- and middle-income countries (Iyer, et al. 2020).

Our findings should be interpreted within the context of rapid increase in educational access. Focusing on math scores for grade four students, Tiruneh et al. (2022) found that students in RISE showed somewhat higher learning *progress* over the course of a school year compared to Young Lives, but learning *levels* were lower. They note that students in 2018–19 were likely to carry an educational disadvantage which started from early grades, in part because during the school year of 2018-19 more children from disadvantaged backgrounds continued to attend school compared to those observed in the 2012-13 school year.

We used the longitudinal school survey from Young Lives and RISE to examine which of the teacher and school factors which are related to the GEQIP -II reform are associated with higher mathematics achievement by the end of the school year. We note the important caveat that not all schools in the YL survey appear in RISE (and vice versa). That said, the key findings highlight the importance of math teacher qualifications, teacher knowledge and location.

In terms of the educational qualifications for mathematics teachers, there has been an improvement over the six years between YL and RISE. It is interesting to note that despite the improvement in math teacher qualifications, students' mathematics levels in 2018 are still at a lower base than those in 2012. Yet, over the course of one academic year, students in 2018 are making more progress than students in 2012 and this progress is positively associated with the educational qualifications of mathematics teachers. In other words, students who have teachers with higher levels of educational qualifications are more likely to have a higher progress over the 2018 academic year relative to 2012.

In relation to the teacher content knowledge in mathematics, there is also an improvement over the six years between YL and RISE. Mathematics teachers in 2018 scored higher in content knowledge compared with mathematics teachers in 2012. And importantly, this content knowledge was found to be positively associated with progress in mathematics over the course of one academic year in 2018 relative to 2012. In other words, students who have teacher who are more knowledgeable about their topic achieve a higher progress over the 2018 academic year relative to 2012.

We used quantile regression analysis to assess whether the magnitude of the associations we observe in these data differ across different centiles of the distribution of math content knowledge at the end of the school year. We find that weaker students, those at the lower end of the test score distribution benefited relatively more from having a teacher who is knowledgeable in the subject. Similar findings were observed when we looked at children from the poorest backgrounds, or students in schools that were only sampled in the RISE survey. This points towards an important equity aspect of the GEQIP reforms. For pupils who are lagging behind in learning mathematics as a subject, students from the poorest backgrounds and more disadvantaged areas, having a teacher who knows their subject matters. Conjecturing, teachers with more content knowledge may be better at diagnosing student errors and correcting them, whereas weaker content knowledge teachers are less able to do so. By contrast, at this grade level, stronger students may be less dependent on the content knowledge of their teachers.[7]

A first glance at our descriptive findings suggests a paradox – lower levels of schooling attainment (as measured by end-of-school-year grade four math scores) coinciding with improvements in teacher quality. The key to reconciling this paradox is to recognize that the reforms that underpin these improvements in school quality may well be working in multiple ways, changing the composition of the student body who attend school as well as increasing the amount of learning that takes place in the classroom, particularly for weaker students. This, juxtaposed with the fact that the education reforms implemented over the last decade have sought to both include students from disadvantaged backgrounds and to improve teacher quality, is suggestive of the possibility that the GEQIP initiatives producing positive learning results.

[7] Our thanks to anonymous reviewer who suggested this possibility.

References

Aurino, E., James, Z., Rolleston, C., 2014. Young Lives Ethiopia School Survey 2012–13. Young Lives, Oxford.

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman-Loxley effect" on mathematics and science achievement. Comparative education review, 46(3), 291-312.

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application (No. 1). Cambridge university press

Efron, Bradley, and Robert J. Tibshirani. (1993). An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall.

Glewwe, P., & Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. Handbook of the Economics of Education, 2, 945-1017.

Horowitz, Joel. 2003. "The Bootstrap in Econometrics." *Statistical Science* 18 (2): 211–18

Hanushek, E. A., Piopiunik, M., & Wiederhold, S. (2019). The value of smarter teachers international evidence on teacher cognitive skills and student performance. *Journal of Human Resources, 54*(4), 857–899

Hoddinott, J., Iyer, P., Sabates. R, and Woldehanna T. (2019) Evaluating large-scale education reforms in Ethiopia. RISE Working Paper No 19/034. RISE Programme: Oxford Policy Management.

Hungi, N., Ngware, M., Mahuro, G., & Muhia, N. (2017). Learning barriers among Grade 6 pupils attending rural schools in Uganda: implications to policy and practice. *Educational Research for Policy and Practice, 16*(2), 129–155. https://doi.org/10.1007/s10671-016-9199-2

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*, *46*(5), 601-618.

Le Cook, B., & Manning, W. G. (2013). Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. *Shanghai archives of psychiatry*, *25*(1), 55.

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.*

OECD (2010), "Learning Outcomes and Socio-Economic Background", in PISA 2009 Results: Overcoming Social Background: Equity in Learning Opportunities and Outcomes (Volume II), OECD Publishing, Paris. DOI: https://doi.org/10.1787/9789264091504-7-en

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

Pankhurst, A., Woldehanna, T., Araya, M., Tafere, Y., Rossiter, J., Tiumelissan, A., & Birhanu, K. (2018). *Young Lives Ethiopia: Lessons from longitudinal research with the children of the millennium*. Young Lives.

Rolleston, C., Hoddinott, J., Dawit, T., Sabates, R., & Woldehanna, T. (2021). Understanding Achievement in Numeracy Among Primary School Children in Ethiopia: Evidence from RISE Ethiopia Study.

Tiruneh, D. T., Sabates, R., Rolleston, C., & Hoddinott, J. (2022). Trends in mathematics learning in Ethiopia: 2012 – 2019. *Bahir Dar Journal of Education*, *21*(1), 26-45. Retrieved from https://journals.bdu.edu.et/index.php/bje/article/view/669

UNICEF. (2017). Education sector budget brief -2016/17. UNICEF for every child 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., & Pankhurst, A. (2014). Young Lives survey design and sampling in Ethiopia. Young Lives, Oxford.

World Bank. (2008). *General Education Quality Improvement Project (GEQIP-I). Project Appraisal Document.* Report No.: 45140-ET.

World Bank. (2013). *General Education Quality Improvement Project II (GEQIP-II). Project Appraisal Document*.Report No: PAD476.

World Bank. (2017). *General Education Quality Improvement Program for Equity (GEQIP-E): Project Appraisal Document: Report No: 121294-ET.* The World Bank.

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

Table 1: Student characteristics, by survey

	Young Lives	RISE	P Values, test of differences in means
	Mean (SD)		
Age, years	10.9 (1.5)	11.2 (1.6)	<0.01
	Percent		
Girl (%)	53.5	48.8	<0.01
Boy (%)	47.5	51.2	<0.01
Attended pre-school (%)	0.52	0.42	<0.01
Dropped out of school for ≥ three months at least once (%)	16.5	9.9	<0.01
Lives in urban area (%)	71.4	34.3	<0.01
Lives in: (%)			
Addis Ababa	20.1	15.3	<0.01
Amhara	15.1	14.8	
SNNP	27.1	11.7	
Somale	5.8	6.9	
Tigray	19.4	14.1	
Oromia	12.5	27.2	
Benishangul Gumuz	-	10.1	
	Mean (SD)		
Caregiver literate (%)	49.0	34.4	<0.01
Household wealth index	0.65(1.7)	-0.39(1.1	<0.01

Note: Sample sizes are: Young Lives, 3,654; RISE, 2,977.

Table 2: Characteristics of students' teachers, by survey

	Young Lives	RISE	P Values, test of differences in means
	Mean (SD)		
Teaching experience, years	11.0 (8.6)	4.6 (4.1)	<0.01
	Percent		
Highest level of general education is grade 12 or less	34.5	50.4	<0.01
Highest level of general education is post-secondary certificate, diploma, or degree	64.5	49.6	
Highest level of teacher training qualification is certificate	39.1	11.1	<0.01
Highest level of teacher training qualification is diploma or degree	60.9	88.9	
Education included specialization in mathematics	49.2	57.1	<0.01
	Mean (SD)		
Teacher qualification index	-0.37(1.4)	0.43 (0.8)	<0.01
Teacher math content z score	-0.15 (0.9)	0.23 (0.9)	<0.01

Note: Sample sizes are: Young Lives, 3,654; RISE, 2,977.

Table 3: Characteristics of students' principals and schools, by survey

	Young Lives	RISE	P Values, test of differences in means
	Principals		
	Mean (SD)		
Years as principal, at school	3.4(2.4)	3.9(3.4)	<0.01
	Percent		
Highest level of general education is post-grad diploma or less (%)	51.8	67.9	<0.01
Highest level of general education is degree (%)	48.2	32.1	
Highest level of teacher training qualification is certificate (%)	63.0	34.7	<0.01
Highest level of teacher training qualification is diploma or degree (%)	37.0	65.3	
	Mean (SD)		
Principal qualification index	0.08(1.27)	-0.01(1.51)	<0.01
	Schools		
	Mean (SD)		
Number of classrooms	14.1(9.3)	13.4(8.0)	<0.01
	Percent		
Grade 4 students have their own math textbook	81.4	45.6	<0.01
School has:			
Functioning library (%)	75.3	76.2	0.45
Number working computers available for students	1.1(2.6)	1.6(4.3)	<0.01
Internet access (%)	13.3	11.7	0.05
Working radio used by students (%)	72.3	76.7	<0.01
Functional pedagogical centre (%)	52.1	79.5	<0.01
Full shift school (%)	28.8	21.8	<0.01
Number working toilets	9.3(5.1)	6.7(6.1)	<0.01
Separate toilets for girls and boys (%)	94.6	88.8	<0.01
	Mean (SD)		
Class size, grade 4	53.1(15.5)	54.5(18.8)	<0.01
School Infrastructure Quality Index	0.33(1.54)	-0.17(1.84)	<0.01

Note: Sample sizes are: Young Lives, 3,654; RISE, 2,977.

Table 4: Student scores on math content test, by survey

	Young Lives	RISE	P Values, test of differences in means
	Mean (SD)		
Baseline, October	512.4(86.8)	459.5(93.0)	<0.01
Endline, May	540.8(99.8)	492.5(104.3)	<0.01

Note: Sample sizes are: Young Lives, 3,654; RISE, 2,977.

Table 5: Comparison of student, teacher, and school characteristics, by survey round and attrition status

Young Lives

	All pupils	Attritors	Completors	Difference in means	P values for T test on equality of means
Student score, math content test	510.3	496.8	512.4	15.6	<0.01
Age, years	11.0	11.6	10.9	-0.7	<0.01
Boy	0.49	0.55	0.47	-0.08	<0.01
Attended pre-school	0.52	0.50	0.52	0.02	0.22
Dropped out of school for ≥ three months at least once	0.18	0.24	0.15	-0.08	<0.01
Household wealth index	0.60	0.32	0.65	0.33	<0.01
Lives in urban area	0.69	0.58	0.72	0.14	<0.01
Caregiver literate	0.49	0.46	0.49	0.03	0.11
Teacher math content z score	-0.15	-0.21	-0.15	0.06	0.17
Class size	53.3	54.7	53.2	-1.5	0.03
Teacher qualification index	-0.37	-0.39	-0.37	0.02	0.74
School Infrastructure Quality Index	0.32	0.15	0.34	0.19	<0.01
Principal qualification index	0.06	-0.07	0.08	0.15	<0.01
Sample size	4232	578	3654		

RISE

	All pupils	Attritors	Completors	Difference in means	P values for T test on equality of means
Student score, math content test	454.5	432.1	459.5	27.4	<0.01
Age, years	11.3	11.6	11.2	-0.4	< 0.01
Boy	0.51	0.54	0.51	-0.03	0.12
Attended pre-school	0.41	0.36	0.42	0.16	<0.01
Dropped out of school for ≥ three months at least once	0.10	0.11	0.10	-0.01	0.39
Household wealth index	-0.44	-0.66	-0.39	-0.44	<0.01
Caregiver literate	0.33	0.28	0.34	0.06	<0.01
Lives in urban area	0.32	0.22	0.34	0.12	<0.01
Teacher math content z score	0.23	0.23	0.23	<0.01	0.89
Teacher qualification index	0.46	0.55	0.44	-0.11	<0.01
Class size	55.8	61.6	54.5	-7.1	<0.01
School Infrastructure Quality Index	-0.24	-0.63	-0.16	0.47	<0.01
Principal qualification index	<0.00	0.14	-0.03	-0.17	<0.01
Sample size	3635	658	2977		

Table 6: Correlates of endline test scores, by survey

	Young Lives	RISE
Teacher and school characteristics		
Teacher math content knowledge, z score	1.590	10.070***
	(2.378)	(2.803)
Math teacher qualification index	4.944***	10.335***
	(1.540)	(3.433)
Log class size	-5.782	-12.670
	(8.967)	(8.378)
School infrastructure index	-1.705	1.442
	(1.876)	(2.020)
Principal qualification index	-0.617	1.415
	(1.823)	(1.849)
Child characteristics		
Baseline math score	0.773***	0.731***
	(0.022)	(0.025)
Age	0.317	4.801***
	(0.896)	(0.984)
Sex (=1 if boy)	0.881	5.001
	(2.294)	(2.705)
Ever attended pre-school	8.609**	-4.760
	(3.448)	(3.805)
Stopped schooling for three months	-13.496***	-3.041
	(3.970)	(4.246)
Household wealth index	0.876	-2.830
	(1.042)	(1.861)
Location (=1 if urban)	18.222***	25.765***
	(6.876)	(6.383)
Caregiver is literate	-4.342**	1.939
	(2.065)	(3.311)
Log time (days) elapsed, baseline to endline	57.891	-28.930
	(51.988)	(38.316)
Intercept	-144.652	287.920
	(274.610)	(197.081)
Observations	3,654	2,977
R-squared	0.540	0.541

Notes: Cluster robust standard errors in parentheses. *** p<.01, ** p<.05.

Table 7: Quantile Regression results, by survey and selected quantiles

Young Lives

	Quantile				
	10th	25th	Median	75th	90th
Teacher and school characteristics					
Teacher math content knowledge, z score	3.87	2.469	0.45	0.23	0.64
	(2.24)	(1.74)	(1.68)	(1.72)	(1.88)
Math teacher qualification index	4.84***	4.14***	4.88***	5.67***	5.06***
	(1.56)	(0.99)	(0.93)	(1.10)	(1.31)
Log class size	-8.69	-2.89	-7.95	-7.86	-10.86
	(8.38)	(5.57)	(5.15)	(6.31)	(6.79)
School infrastructure index	-1.36	-2.42**	-1.49	-0.41	-1.15
	(1.77)	(1.18)	(1.71)	(1.40)	(1.44)
Principal qualification index	-0.80	-1.24	-0.09	-0.29	-1.63
	(1.87)	(1.33)	(1.15)	(1.49)	(1.70)
Child characteristics					
Baseline math score	0.77***	0.80***	0.81***	0.80***	0.72***
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)

Notes: *** p<.01, ** p<.05. Sample size is 3654. Other variables included but not reported are child age and sex, whether attended pre-school, stopped schooling for three months or more, household wealth and location, caregiver is literate.

Quantile Regression results: All children, RISE

	Quantile				
	10th	25th	Median	75th	90th
Teacher and school characteristics					
Teacher math content knowledge, z score	12.08***	10.56***	11.06***	7.04***	3.84
	(2.03)	(1.80)	(1.90)	(2.03)	(3.13)
Math teacher qualification index	8.75***	11.48***	9.58***	9.19***	12.18***
	(2.32)	(2.11)	(1.98)	(2.338)	(3.25)
Log class size	-21.92***	-14.33**	-7.72	-4.96	-15.21
	(5.86)	(7.00)	(5.49)	(6.31)	(9.54)
School infrastructure index	0.06	0.16	2.73	3.53**	-0.34
	(1.59)	(1.40)	(1.40)	(1.50)	(2.09)
Principal qualification index	1.02	1.47	2.10	1.34	1.69
	(1.49)	(1.30)	(1.21)	(1.38)	(1.89)
Child characteristics					
Baseline math score	0.59***	0.67***	0.78***	0.81***	0.80***
	(0.02)	(0.02)	(0.02)	(0.023)	(0.03)

Notes: *** p<.01, ** p<.05. Sample size is 2977. Other variables included but not reported are child age and sex, whether attended pre-school, stopped schooling for three months or more, household wealth and location, caregiver is literate.

Table 8: Quantile Regression results, RISE, by sex

Girls

	Quantile				
	10th	25th	Median	75th	90th
Teacher and school characteristics					
Teacher math content knowledge, z score	12.65***	9.55***	8.99***	6.69***	-1.73
	(2.31)	(2.54)	(2.52)	(2.97)	(4.09)
Math teacher qualification index	5.16***	10.17***	8.74***	6.89	9.22
	(3.04)	(3.08)	(2.42)	(3.54)	(5.00)
Log class size	-19.79***	-25.86***	-12.12	-7.31	-16.71
	(8.79)	(9.55)	(8.55)	(8.49)	(11.37)
School infrastructure index	2.09	1.83	6.56***	7.95***	2.98
	(2.13)	(2.32)	(1.83)	(2.03)	(2.92)
Principal qualification index	1.36	0.46	3.35	1.46	1.14
	(1.83)	(2.04)	(1.79)	(1.95)	(2.59)
Child characteristics					
Baseline math score	0.54***	0.58***	0.71***	0.75***	0.75***
	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)

Notes: *** p<.01, ** p<.05. Sample size is 1457. P values for F tests of null hypotheses that associations between teacher and school characteristics and endline test scores are equal across selected quantiles.

Quantile comparison	Characteristic	P value
10th and 90th	Teacher content knowledge	<0.01
25th and 90th	Teacher content knowledge	0.01
10th and 90th	Teacher qualifications	0.46
10th and 90th	Log class size	0.82
10th and 90th	School quality index	0.80
10th and 90th	Principal qualifications	0.94

Boys

	Quantile				
	10th	25th	Median	75th	90th
Teacher and school characteristics					
Teacher math content knowledge, z score	13.61***	10.29***	10.51***	10.82***	8.25
	(2.94)	(2.36)	(2.29)	(3.13)	(4.31)
Math teacher qualification index	17.63***	14.67***	12.04***	9.44**	12.65***
	(2.95)	(3.05)	(2.94)	(3.87)	(4.21)
Log class size	-27.44***	-6.80***	1.63	-1.55	9.47
	(9.62)	(8.14)	(7.74)	(8.96)	(15.12)
School infrastructure index	-1.60	-1.97	-1.74	-1.76	-2.59
	(2.41)	(1.97)	(2.06)	(2.21)	(2.94)
Principal qualification index	-1.44	1.06	0.05	0.95	0.70
	(1.38)	(1.65)	(1.59)	(2.10)	(2.59)
Child characteristics					
Baseline math score	0.67***	0.72***	0.83***	0.85***	0.84***
	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)

Notes: *** p<.01, ** p<.05. Sample size is 1520. P values for F tests of null hypotheses that associations between teacher and school characteristics and endline test scores are equal across selected quantiles.

Quantile comparison	Characteristic	P value
10th and 90th	Teacher content knowledge	0.28
25th and 90th	Teacher content knowledge	0.65
10th and 90th	Teacher qualifications	0.35
10th and 90th	Log class size	0.03
10th and 90th	School quality index	0.78
10th and 90th	Principal qualifications	0.52

Table 9: Test scores by school type

		Young Lives		RISE	
		Baseline	Endline	Baseline	Endline
		(A)	(B)	(C)	(D)
(1)	All YL schools	512.0	540.9	-	-
(2)	Group 1	522.5	552.6	-	-
(3)	Group 2	497.0	524.0	-	-
(4)	All RISE schools	-	-	459.6	492.6
(5)	Group 3	-	-	457.6	494.3
(6)	Group 4	-	-	460.1	492.2

Note: Sample sizes are as follow. Group 1: 2,278. Group 2: 1,577. Group 3: 562. Group 4: 2,449.

Table 10: Child and school characteristics, by school type

		Teacher content knowledge (Z score)	Teacher qualifications (Z score)	School quality (Principal Components Index)	Principal qualifications (Principal Components Index)
		(A)	(B)	(C)	(D)
(1)	All YL schools	-0.19	-0.38	0.38	0.06
(2)	Group 1	-0.13	0.13	0.11	0.02
(3)	Group 2	-0.28	-1.12	0.79	0.12
(4)	All RISE schools	0.23	0.44	-0.17	-0.02
(5)	Group 3	0.39	0.55	0.49	-0.42
(6)	Group 4	0.20	0.41	-0.32	0.08

		Child age	Proportion of children who are boys	Household wealth (Principal Components Index)	Proportion of caregivers who are literate
		(A)	(B)	(C)	(D)
(1)	All YL schools	10.9	0.47	0.68	0.49
(2)	Group 1	10.9	0.47	0.94	0.51
(3)	Group 2	10.9	0.47	0.31	0.47
(4)	All RISE schools	11.2	0.51	-0.39	0.34
(5)	Group 3	11.0	0.49	-0.33	0.39
(6)	Group 4	11.3	0.52	-0.41	0.33

Figure 1: Participant flow diagram, Young Lives

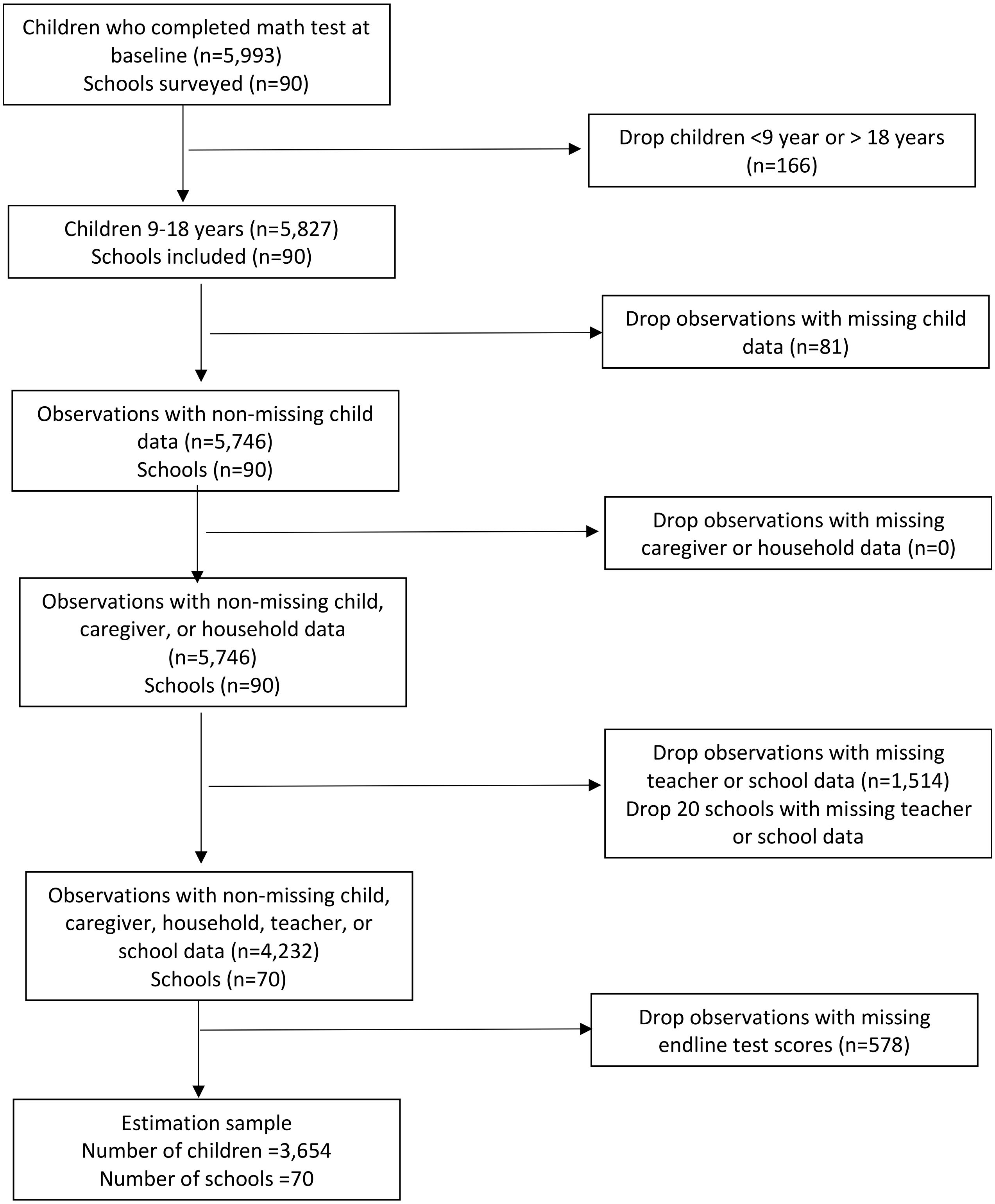

Children who completed math test at baseline (n=5,993)
Schools surveyed (n=90)
Drop children <9 year or > 18 years (n=166)
Children 9-18 years (n=5,827)
Schools included (n=90)
Drop observations with missing child data (n=81)
Observations with non-missing child data (n=5,746)
Schools (n=90)
Drop observations with missing caregiver or household data (n=0)
Observations with non-missing child, caregiver, or household data (n=5,746)
Schools (n=90)
Drop observations with missing teacher or school data (n=1,514)
Drop 20 schools with missing teacher or school data
Observations with non-missing child, caregiver, household, teacher, or school data (n=4,232)
Schools (n=70)
Drop observations with missing endline test scores (n=578)
Estimation sample
Number of children =3,654
Number of schools =70

Figure 2: Participant flow diagram, RISE

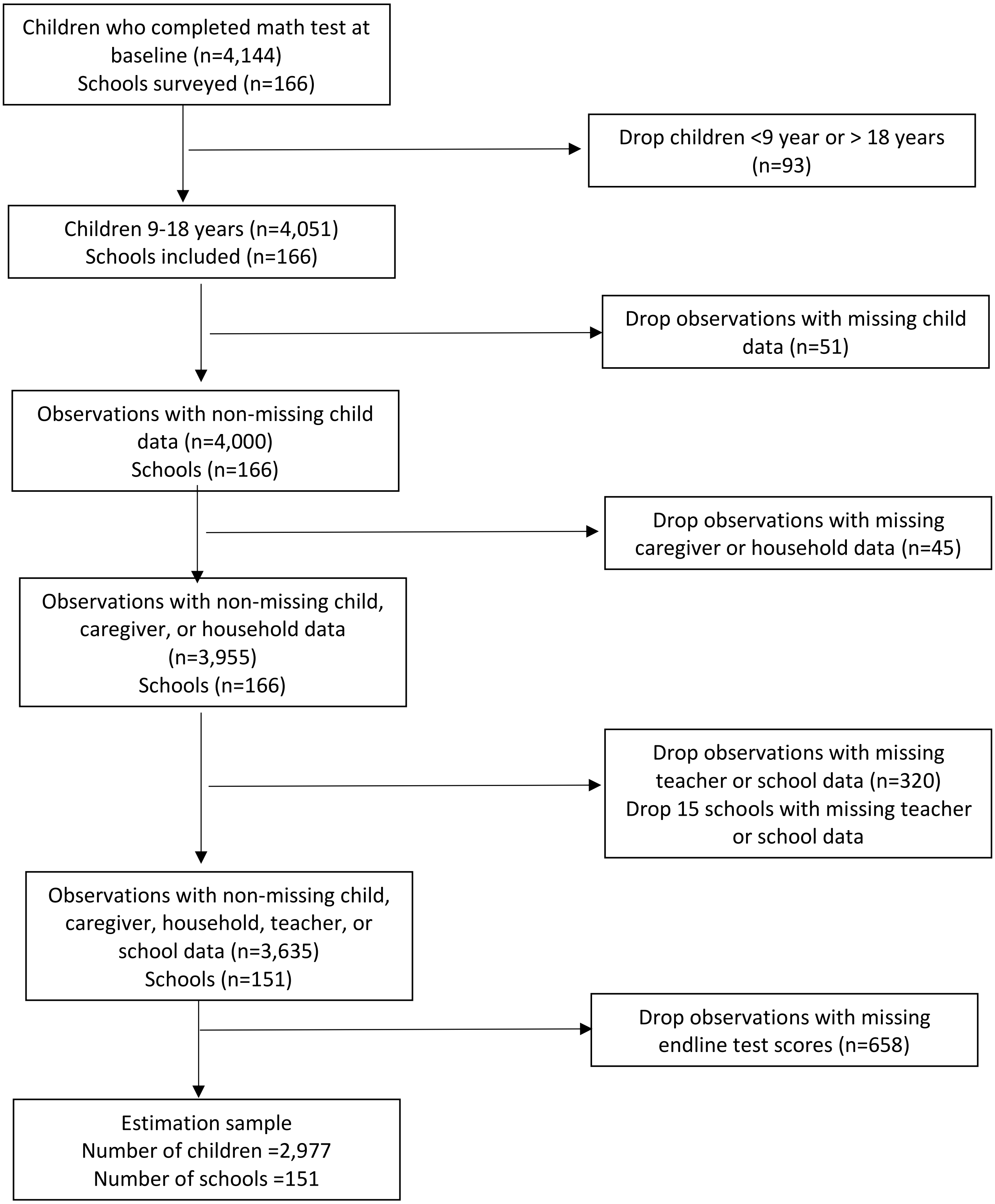

Children who completed math test at baseline (n=4,144)
Schools surveyed (n=166)
Drop children <9 year or > 18 years (n=93)
Children 9-18 years (n=4,051)
Schools included (n=166)
Drop observations with missing child data (n=51)
Observations with non-missing child data (n=4,000)
Schools (n=166)
Drop observations with missing caregiver or household data (n=45)
Observations with non-missing child, caregiver, or household data (n=3,955)
Schools (n=166)
Drop observations with missing teacher or school data (n=320)
Drop 15 schools with missing teacher or school data
Observations with non-missing child, caregiver, household, teacher, or school data (n=3,635)
Schools (n=151)
Drop observations with missing endline test scores (n=658)
Estimation sample
Number of children =2,977
Number of schools =151

Figure 3: Kernel density functions for teacher content knowledge, by survey

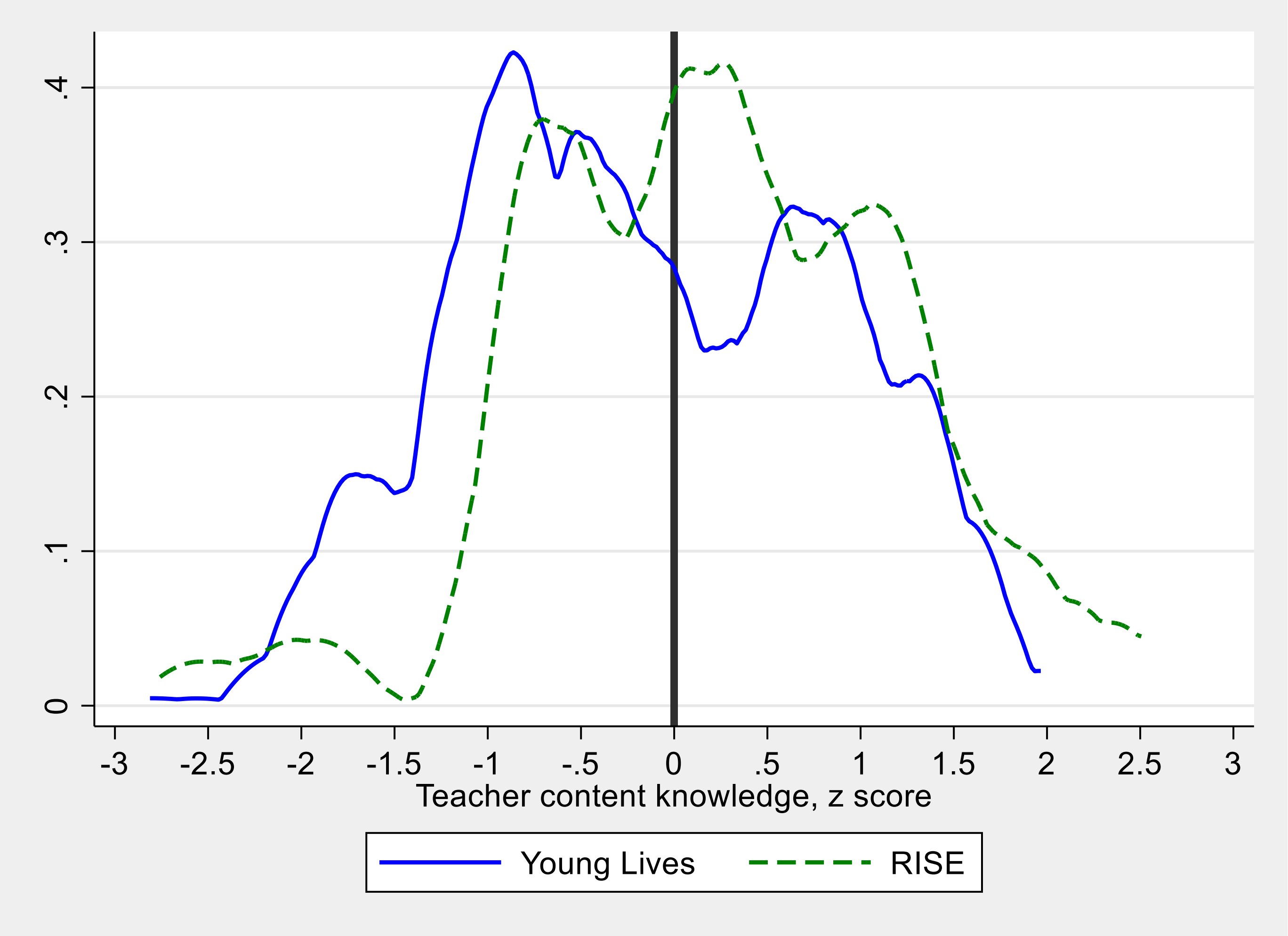
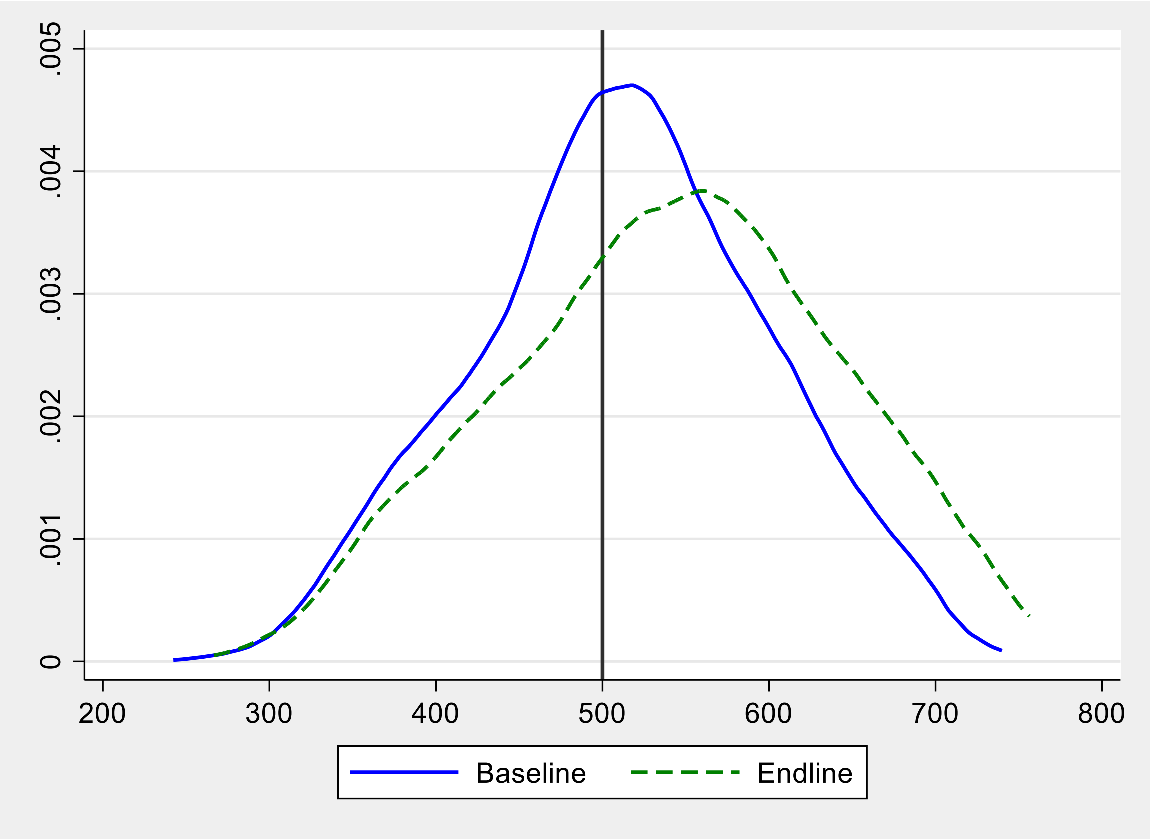


Figure 4: Kernel density functions for student content knowledge, by survey

Young Lives



RISE

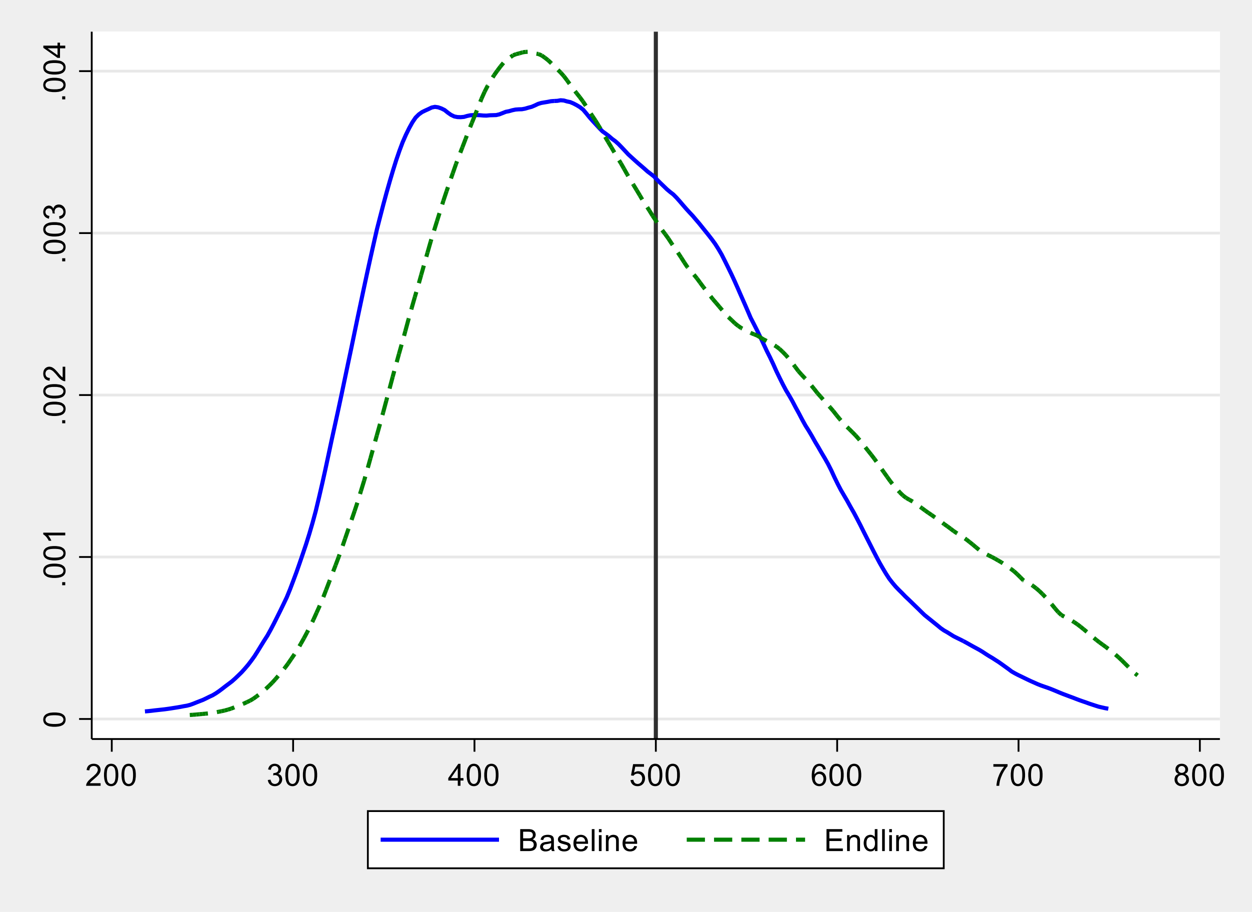
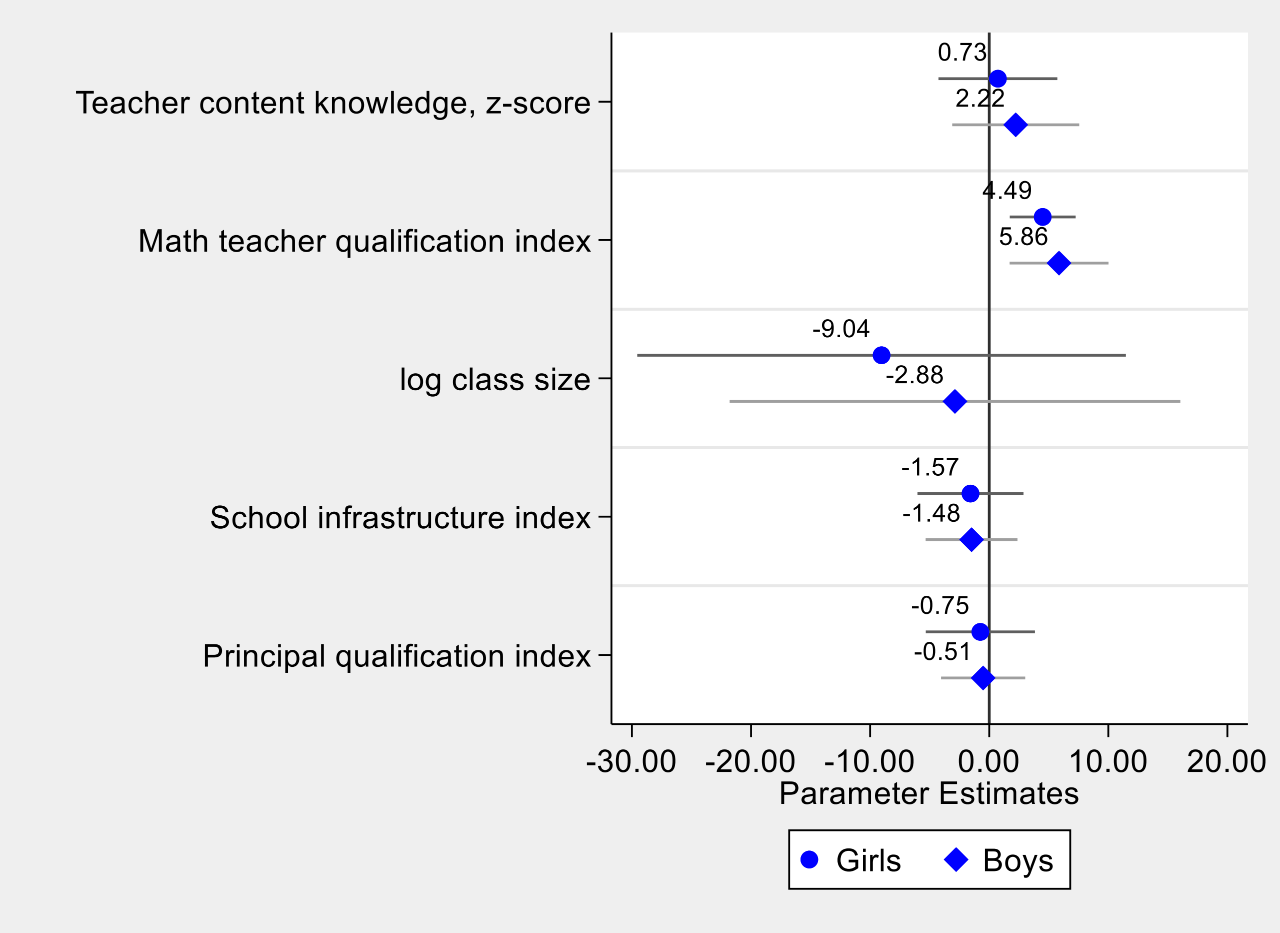


Figure 5: Associations between teacher and school characteristics and endline test scores, by survey and child sex

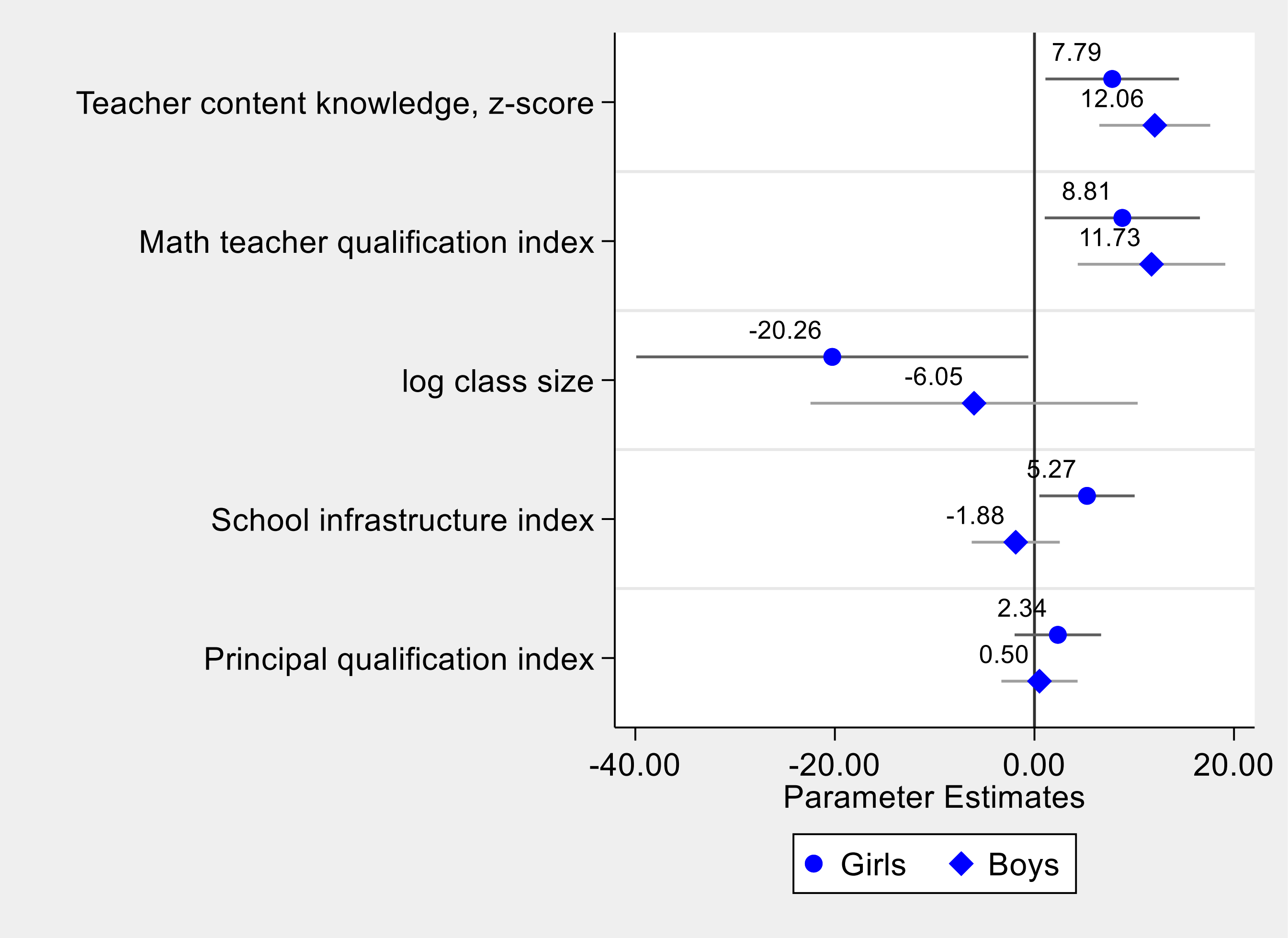
Young Lives



Notes: Sample size is 1919 for girls and 1735 for boys. P values for chi-squared test of null hypotheses that associations between teacher and school characteristics and endline test scores are equal between girls and boys.

Characteristic	P value
Teacher content knowledge	0.43
Math teacher qualifications	0.42
Log class size	0.46
School quality index	0.96
Principal qualifications	0.91

RISE

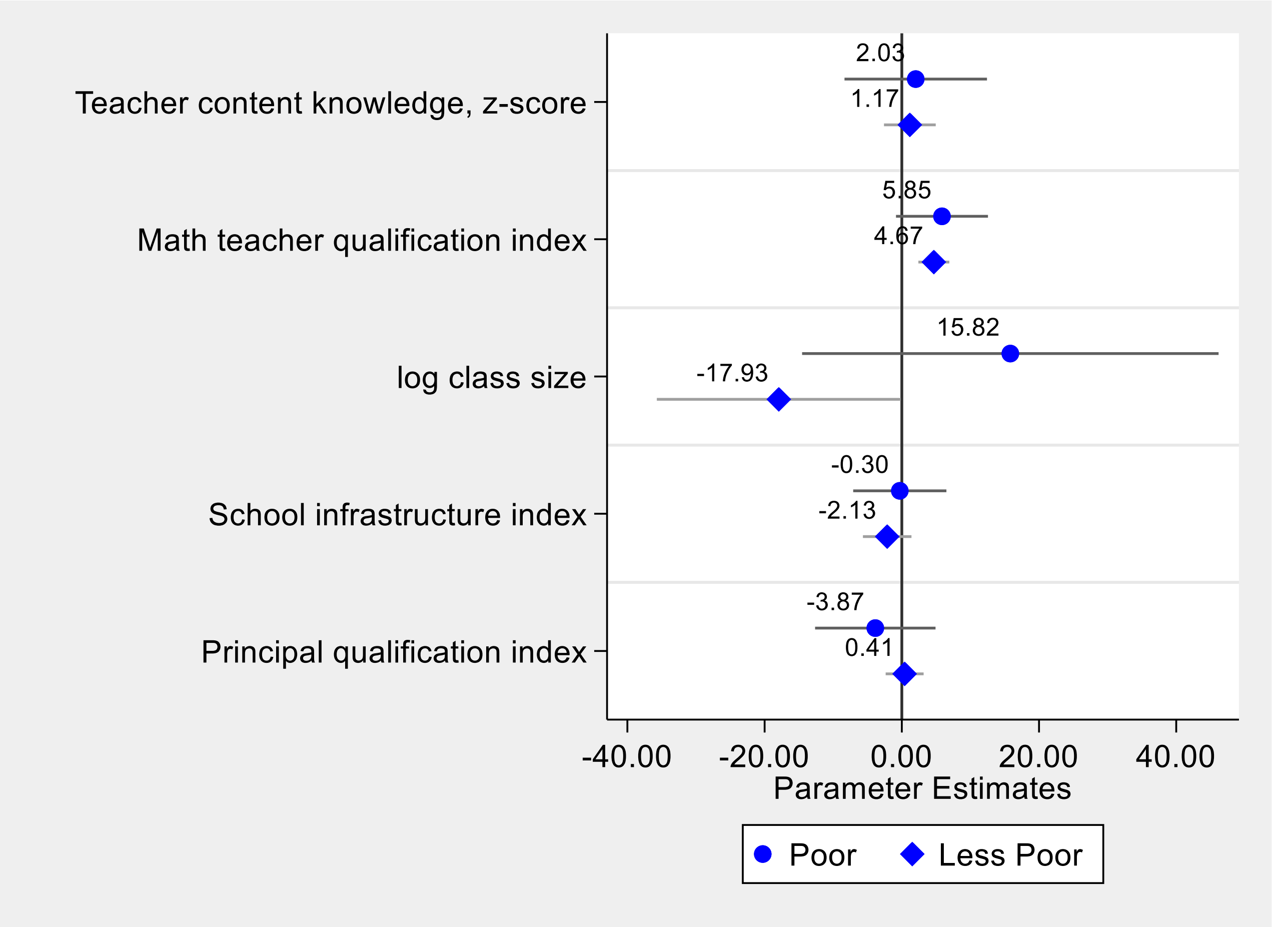


Notes: Sample size is 1457 for girls and 1520 for boys. P values for chi-squared test of null hypotheses that associations between teacher and school characteristics and endline test scores are equal between girls and boys

Characteristic	P value
Teacher content knowledge	0.12
Math teacher qualifications	0.41
Log class size	0.11
School quality index	<0.01
Principal qualifications	0.33

Figure 6: Associations between teacher and school characteristics and endline test scores, by survey and household wealth

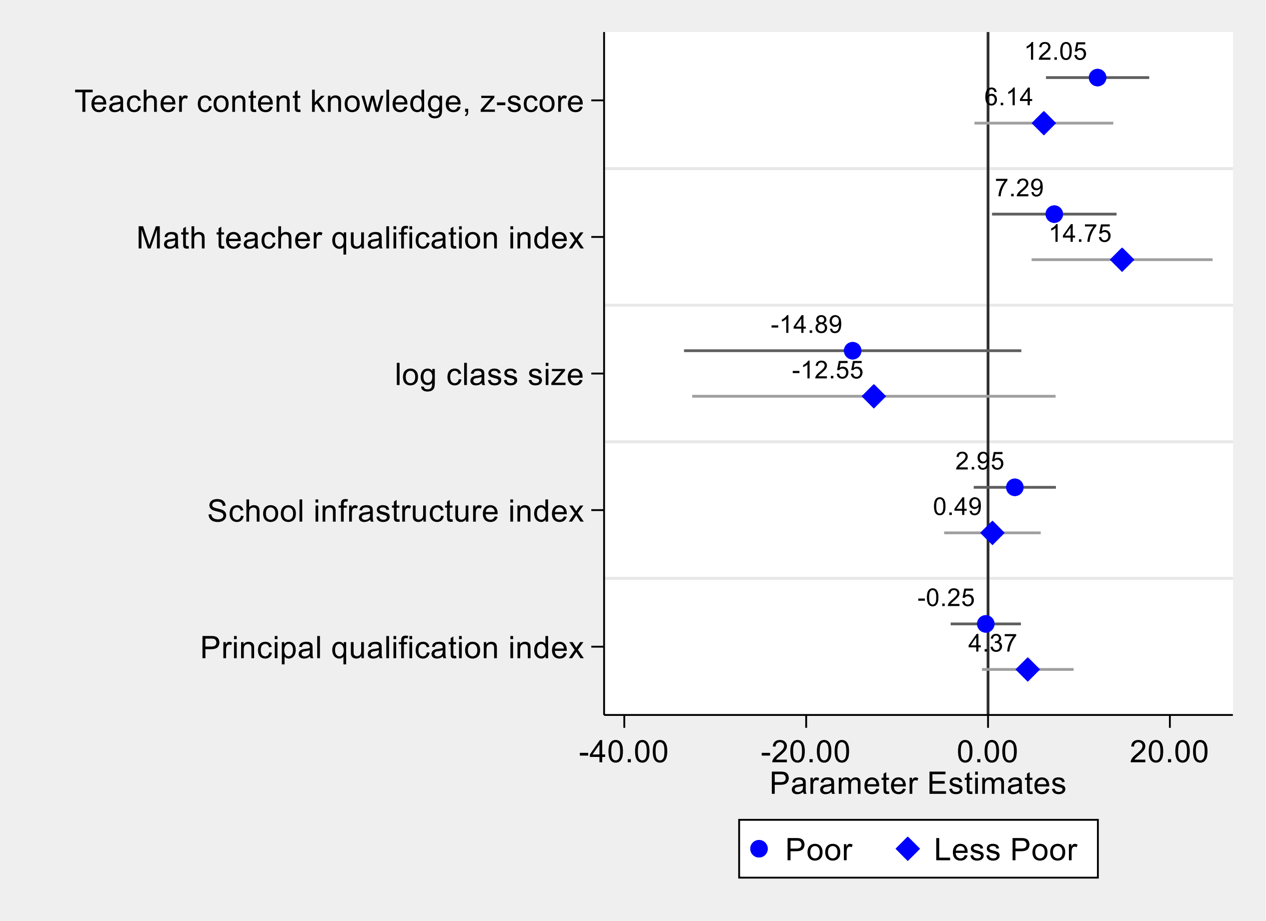
Young Lives



Notes: Poor are the poorest 1/3 of the sample as measured by the wealth index. Sample size is 1084 for the poor sub-sample and 1893 for the less-poor sub-sample. P values for chi-squared test of null hypotheses that associations between teacher and school characteristics and endline test scores are equal between children living in poor and less poor households.

Characteristic	P value
Teacher content knowledge	0.87
Math teacher qualifications	0.71
Log class size	0.03
School quality index	0.60
Principal qualifications	0.33

RISE

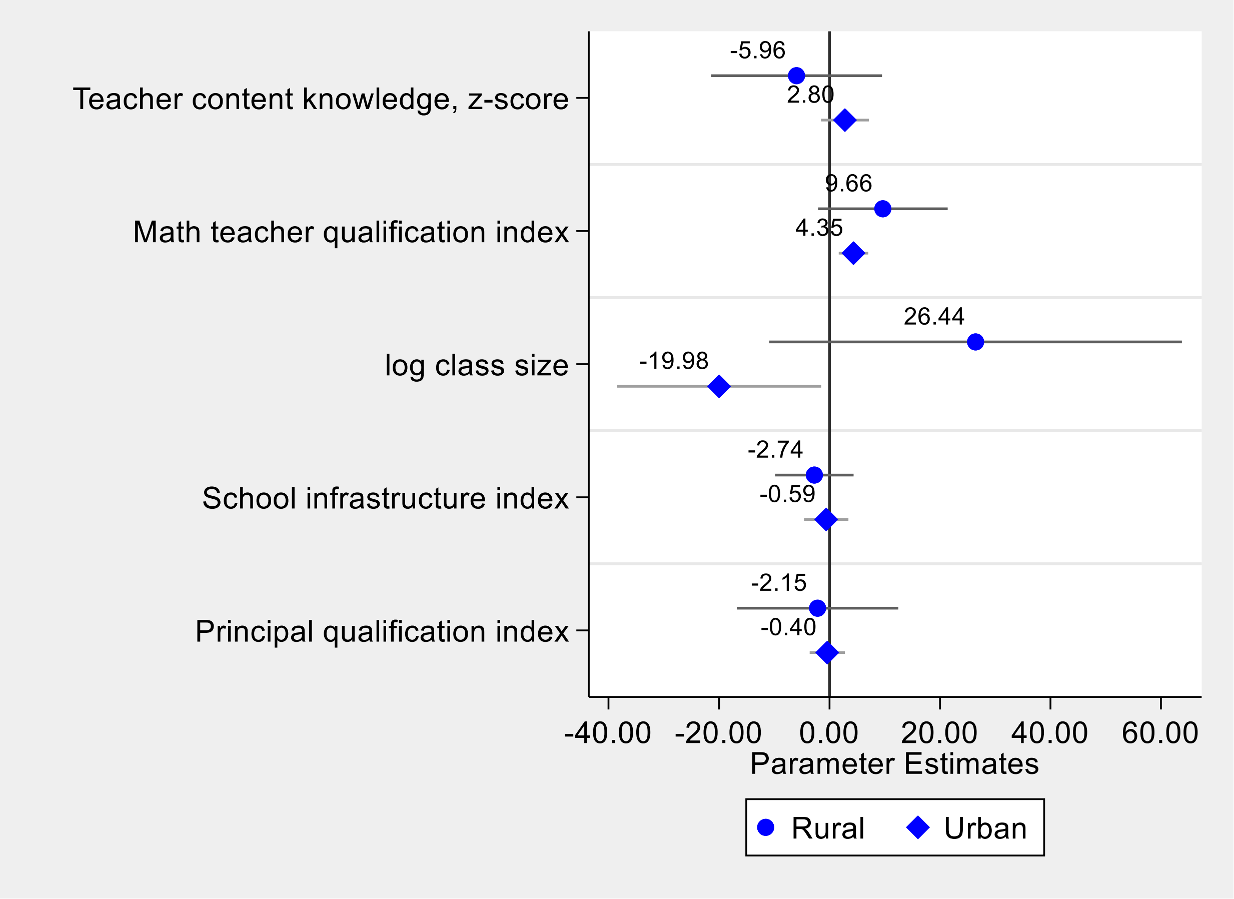


Notes: Poor are the poorest 1/3 of the sample as measured by the wealth index. Sample size is 1084 for the poor sub-sample and 1893 for the less-poor sub-sample. P values for chi-squared test of null hypotheses that associations between teacher and school characteristics and endline test scores are equal between children living in poor and less poor households.

Characteristic	P value
Teacher content knowledge	0.04
Math teacher qualifications	0.29
Log class size	0.99
School quality index	0.46
Principal qualifications	0.34

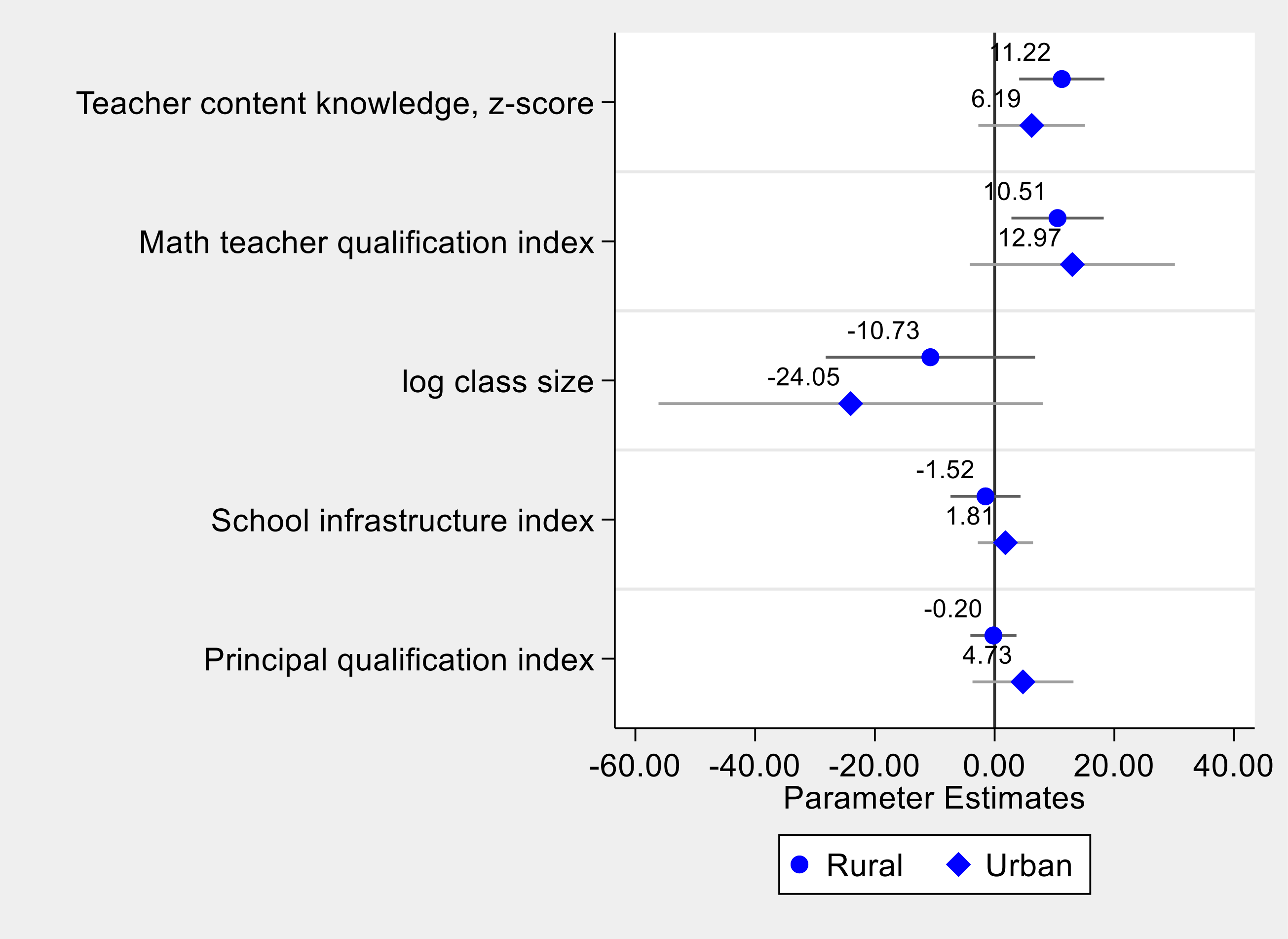
Figure 7: Associations between teacher and school characteristics and endline test scores, by survey and location

Young Lives



Notes: Sample size is 1953 for the rural sub-sample and 1024 for the urban sub-sample. P values for chi-squared test of null hypotheses that associations between teacher and school characteristics and endline test scores are equal between rural and urban areas

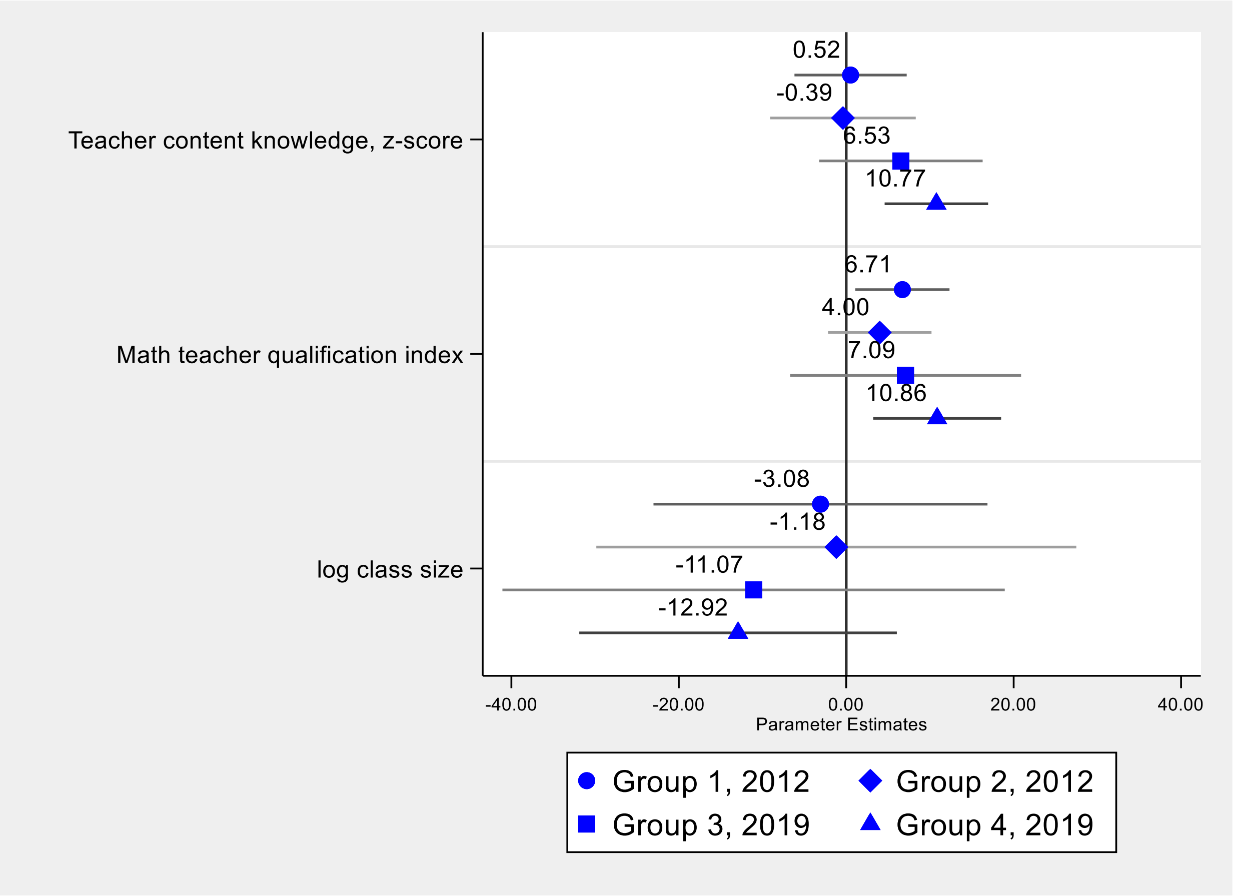
Characteristic	P value
Teacher content knowledge	0.26
Math teacher qualifications	0.36
Log class size	0.02
School quality index	0.59
Principal qualifications	0.81

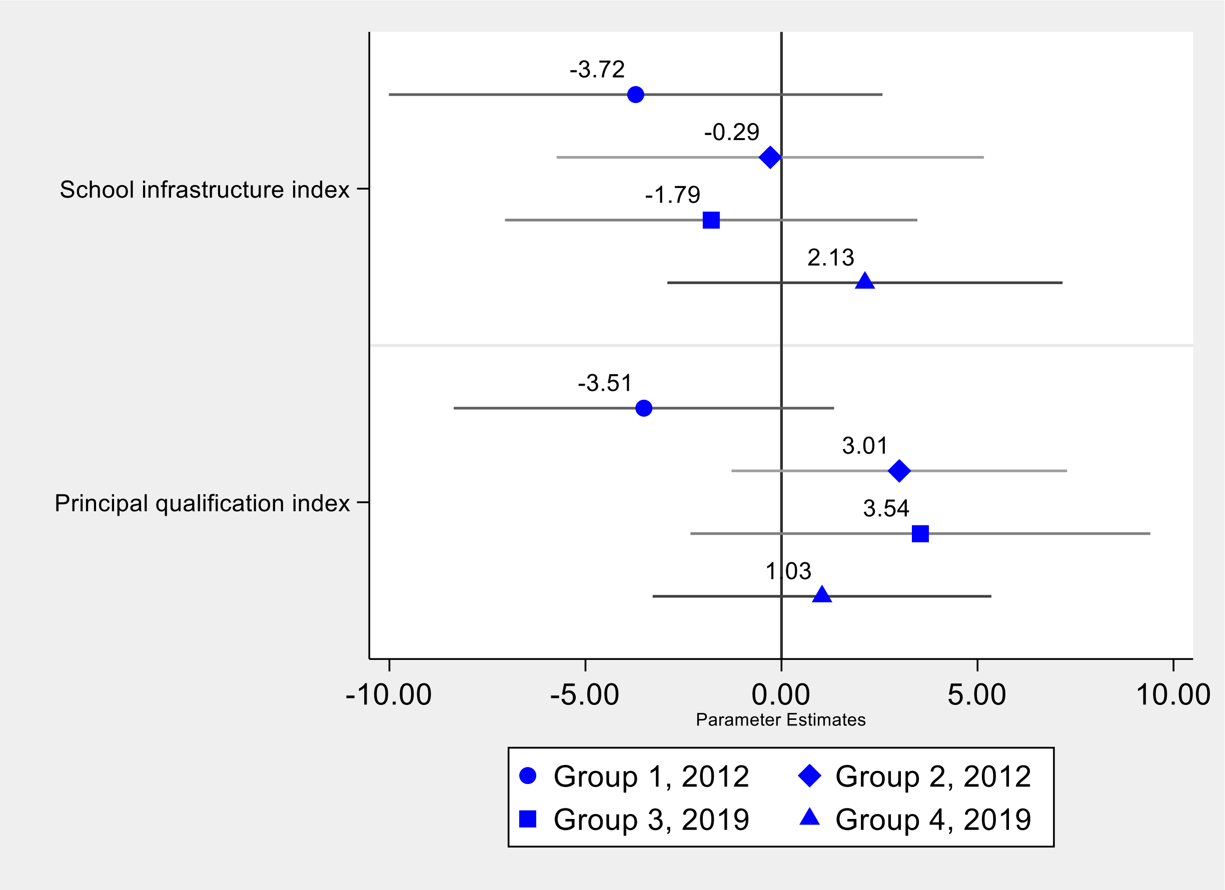


Notes: Sample size is 1953 for the rural sub-sample and 1024 for the urban sub-sample. P values for chi-squared test of null hypotheses that associations between teacher and school characteristics and endline test scores are equal between rural and urban areas

Characteristic	P value
Teacher content knowledge	0.37
Math teacher qualifications	0.79
Log class size	0.46
School quality index	0.37
Principal qualifications	0.28

Figure 8: Associations between teacher and school characteristics and endline test scores, by school type




-3.72
-0.29
-1.79
2.13
-3.51
3.01
3.54
1.03
School infrastructure index
Principal qualification index
-10.00
-5.00
0.00
5.00
10.00
Parameter Estimates
● Group 1, 2012
◆ Group 2, 2012
■ Group 3, 2019
▲ Group 4, 2019

Supplementary Appendix

Supplementary Appendix 1: Sample design, RISE

The design of the RISE sample was based on the need to recognize that Ethiopia's large population lives in diverse agro-climatic conditions including pastoral, highland, and low land areas. Ethiopia is divided into nine regional states and two city administrations. The regional states are defined based on the language people speak. In general, regions can be grouped into four categories:

- Those dominated by or only urban populations: Dire Dawa, Harari, and Addis Ababa.
- 'Emerging' regions largely dominated by pastoral population (Afar and Somali)
- 'Emerging' regions located in the western part of the country (Gambella and Benishangul Gumuz)
- Relative to emerging regions, more developed and rural dominated regions (Tigray, Amhara, Oromia, and SNNP), where more than 90 percent of the Ethiopian population live.

The sample design had to be inclusive conditional on budget constraints. Accordingly, RISE surveys took place in the following regions: Addis Ababa (urban); Somali (pastoral); Benishangul Gumuz (emerging in the west); and Tigray, Amhara, Oromia, and SNNP.

The sampling is school based, specifically state funded public schools. Private schools were purposively excluded. Outside of Addis Ababa, and the cities of Jijiga in Somali and Hawassa in SNNP), there are relatively few private primary schools, and it is unclear to what extent these schools are bound to follow the education reforms that were the focus of the RISE study. Most primary schools offer grades one through eight; since the original design envisaged following students over time, schools that did not offer all eight grades of primary school were excluded.

Power calculations indicated that we needed to survey 168 schools (in practice, we surveyed 166). These were allocated across regions, then selected, in the following way.

1. We sought to include a minimum number of schools for the four least populous regions: Addis Ababa; Benishangul Gumuz; Somali; and Tigray. We set this at 20 for each of these regions.

2. This left 88 schools for the larger regions (168 – (20 x4) = 88). Allocating number of schools in proportion to regional populations, we included 25 schools in Amhara, 41 in Oromia and 22 in SNNP.

3. Given that across the seven regions we worked in, 20% of the population lives in urban areas, we sought to ensure that 20% of our schools were located in urban areas. Thus, the intended d distribution of schools looks like this (Table A2-2):

Table S1.1: Planned distribution of schools across regions

(1)	(2)	(3)	(4)
RISE Regions	Number of schools to be surveyed	Number of rural schools	Number of urban schools
Addis Ababa	20	0	20
Amhara	25	21	4
Benishangul Gumuz	20	16	4

Oromia	41	35	6
SNNP	22	18	4
Somali	20	17	3
Tigray	20	15	5
TOTAL	168	122	46

To select these schools, we needed to account for the phased nature of the intervention (GEQIP E) that RISE sought to study. Phasing occurred in four regions; in these regions, 35% of would receive GEQIP E interventions in the first phase. We also wanted to include schools from the Young Lives sample. We then needed to fill out the remaining schools with a random selection of additional woredas. Doing so was complicated. Woredas are of different size and therefore have different numbers of schools. If we drew a simple random sample of woredas, we would artificially increase the likelihood of selecting schools from less populous woredas and decrease the likelihood of selecting schools from more populous woredas. There were also budget limitations that meant that it would be too costly to do survey work in 168 different woredas. So, we surveyed two schools (and in some cases, 3) per woreda.

Schools were selected in the following manner. We use Tigray as an example.

(1) We want to survey 20 schools in Tigray. Tigray is 27% urban and so we want 15 rural schools and five urban schools.

(2) 35% of our schools in Tigray should be those selected for the first phase of the phased component of GEQIP E. All are rural. Since 35% *x* 20 equals 7, we should include seven phased schools in our Tigrayan sample. To identify these, we used the list of woredas where phase one will take place and using proportionate probability sampling (PPS) based on woreda population, we selected two woredas. Within these woredas, we randomly selected schools.

(3) Next we add in the Young Lives (YL) schools with grades 1-8. There are 7 YL schools in Tigray, five are rural and two are urban.

(4) These leaves us with needing to select 4 additional schools. (Remember we are aiming for 20 schools in Tigray and in this example, we have 7 YL schools and 7 phase one schools.) We took the list of all “non phase one” woredas in Tigray and divide it into rural and urban woredas. We assume that the number of schools across woredas is proportional to woreda size and assume that the number of students per school is roughly similar. The first assumption is likely to be reasonably true (though perhaps less so in more sparsely populated woredas). We have no way of knowing whether the second is true given that there do not seem to be lists of schools which contain the number of students attending each school.

(5) Because Tigray is 27% urban and we are aiming for a sample of 20 schools in Tigray, we need five urban schools. We already have two from YL and so we need another three. We take our list of urban woredas and using PPS, select one urban woreda. This gives us two schools.

(6) We need 15 rural schools. YL provides 5 and phase one provides another 7, leaving us to find 3 additional schools. We take our list of rural woredas and using PPS, select one additional rural woreda from which we will select three schools.

We then repeat this process for the remaining regions.

Lastly, we assume that in each school, there is more than one section(class) per grade. We randomly select two sections(classes) per grade. Tests of learning outcomes are given to all students in each section(class). For the household survey linked to the student, we aimed for 28 students per cohort, 14 students (7 girls, 7 boys) from each section(class).

Supplementary Appendix 2: Accounting for attrition in the RISE sample

We have noted that attrition in RISE is non-random; for example, students with lower baseline scores are less likely to complete the school year and take the endline math test. Consequently, the sample that remains is self-selected in terms of math ability (as measured by the baseline test) and this creates a risk of biasing our parameter estimates. One approach to resolving this is to create sample weights that address this selectivity problem. Fitzgerald, Gottschalk, and Moffitt (1998) show that these can be constructed based on estimating two models of attrition: one where we estimate attrition as a function of the variables that we think affect our outcome of interest (here, endline math scores); and a second where we estimate attrition as a function of the variables that we think affect our outcome of interest ***and*** variables that we think ***only*** affect the likelihood that the child stays in school. We estimate predicted probabilities of remaining in school from both models; Fitzgerald, Gottschalk, and Moffitt show that the weights are the ratio of these predicted probabilities.

There are nine steps to doing this:

Step 1: Identify variables that we think affect the outcome. Call these "**X**"

Step 2: Identify variables that we think influence attrition but do not directly affect the outcome. Call these "**Z**"

Step 3: Run a regression where the dependent variable is the attrition variable. Include both "**X**" and "**Z**"

Step 4: Test whether we can reject the null that "**Z**" are jointly zero

Step 5: Predict likelihood of attrition and store these predicted values

Step 6: Run a regression where the dependent variable is the attrition variable. Include only "**X**"

Step 7: Predict likelihood of attrition and store these predicted values

Step 8: Construct attrition weights where these = predicted values from only "**X**"/ predicted values from "**X**" and "**Z**"

Step 9: Run regressions with and without these weights

Step 1: Our outcome of interest is the endline math score. We assume that this is a function of the same variables that appear in the regressions reported in the main text: baseline math score, the child's age, and sex at baseline, whether the child had attended pre-school, whether the child had previously dropped out for 3 months or more, whether the caregiver was literate; teacher characteristics (math content test score; qualifications); and school (log classroom size, school infrastructure score; principal qualification score) characteristics. Collectively, these are the **X** variables.

Step 2: There are series several that conceivably might only affect the likelihood that the child stays in school (and not the outcome variable itself). These are the **Z** variables:

(a) Variables that capture physical access to schools (or more generally, how physically connected a child and her household is with the community in which she leaves). The assumption here is that physical access affects whether the child can get to school, not what she learns when she gets there. Examples could include:

- Distance from home to school
- A variable that captures how easy it is for a child to get to school (road quality could also work here)
- Does the household have a mobile phone (the logic here is that, conditional on SES, more isolated households would be less likely to own a phone)

(b) Continuing with the theme of connectedness, there are questions about how long the child has lived in the kebele, whether her parents were born in this kebele, how long the household had resided in the kebele etc.

(c) Whether there are other school-aged children attending school. The logic here is that if other children in the household are attending school, that our index child is more likely to be attending school. So this could include:

- Number of children older than the index child who were attending school at baseline
- Number of children younger than the index child who were attending school at baseline

Here, we work with the following variables that we assume affect the likelihood that the child does the endline math test: a dummy variable equaling one if the child has lived in the locality for five years or less (this is an example of the connectedness idea); a dummy variable equaling one if the road to the school is a mud track (another example of connectedness); and the number of older children in the household attending school.

Step 3: Run the attrition regression with both the **X** and **Z** variables as regressors. Here the dependent variable =1 if the child completed the endline test, zero otherwise. So a negative coefficient means that that variable is associated with a lower likelihood that the child completed; a positive coefficient means that that variable is associated with a higher likelihood that the child completed.

Table S2.1: Correlates of attrition, RISE

Characteristics associated with attrition but not with learning	
Child has lived in kebele for ≤ 5 years	-0.054**
	(0.025)
Number of children older than the index child attending school at baseline	0.014**
	(0.006)
Road to the school is a mud track	-0.025
	(0.026)
Child characteristics	
Baseline math score	0.0004***
	(0.00007)
Age	-0.018***
	(0.005)
Sex (=1 if boy)	-0.023*
	(0.012)
Attended pre-school	-0.014
	(0.017)
Stopped schooling for >3 months	0.008
	(0.023)
Household wealth index	0.017**
	(0.007)
Location (=1 if urban)	0.0006
	(0.028)
Caregiver literate	-0.004
	(0.014)
Teacher and school characteristics	
Teacher math content knowledge, z score	0.0005
	(0.011)
Math teacher qualification index	-0.028**
	(0.012)
Log class size	-0.108***
	(0.031)
School infrastructure index	0.004
	(0.007)
Principal qualification index	-0.004
	(0.007)
Intercept	1.325***
	(0.132)
Number of observations	3635

Notes: Cluster robust standard errors in parentheses. *** p<.01, ** p<.05.

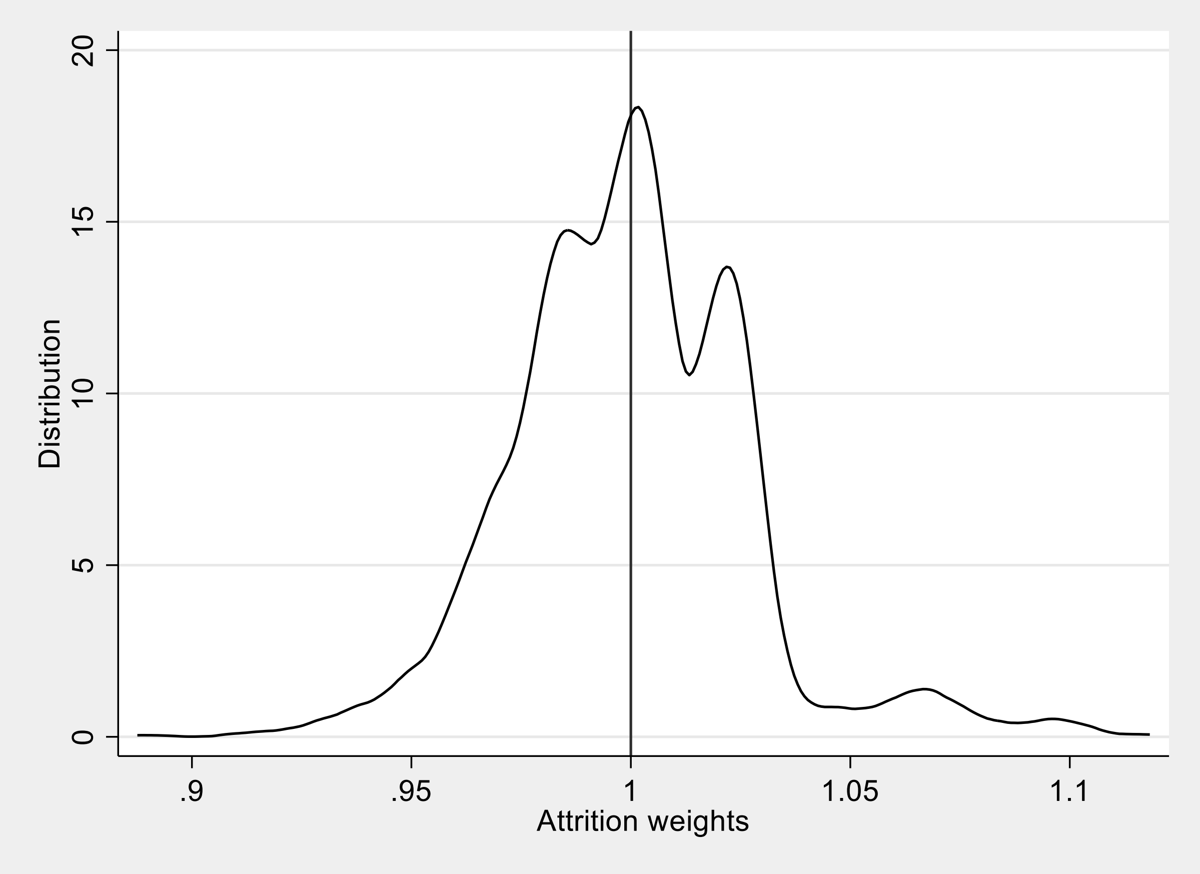
Children who are recent arrivals to the locality are less likely to be tested at endline (put differently, they are more likely to drop out) as are children who need to travel along a mud track in order to get to the school. Children in households with older children also attending school are more likely to complete the endline test (put differently, they are less likely to drop out).

Step 4*:* The F test on the joint significance of the three **Z** variables = 3.86, P ==0.01 and so we are confident that we have variables that affect the likelihood of completion.

Steps 5, 6, and 7. These are available on request.

Step 8. Have calculated the weights, we can graph their distribution (using a kernel density function), shown below. There is some variability in the weights – they range from 0.93 to 1.1 – but the mass of the distribution is around 1. This hints at the possibility that the attrition weighted regression estimates will not differ too much from the unweighted regressions.

Figure S2.1: Kernel density functions for attrition weights, RISE



Step 9. Regression results without and with weights.

Table S2.2: Regression results without and with weights, RISE

Teacher and school characteristics	Unweighted results	Attrition weighted results
Teacher math content knowledge, z score	10.070***	10.120***
	(2.803)	(2.793)
Math teacher qualification index	10.335***	10.315***
	(3.433)	(3.452)
Log class size	-12.670	-12.570
	(8.378)	(8.3746)
School infrastructure index	1.442	1.465
	(2.020)	(2.016)
Principal qualification index	1.415	1.389
	(1.849)	(1.848)
Child characteristics		
Baseline math score	0.731***	0.731***
	(0.025)	(0.025)
Age	4.801***	4.788***
	(0.984)	(0.989)
Sex (=1 if boy)	5.001	4.993
	(2.705)	(2.705)
Ever attended pre-school	-4.760	-4.712
	(3.805)	(3.807)
Stopped schooling for >3 months	-3.041	-3.093
	(4.246)	(4.216)
Household wealth index	-2.830	-2.804
	(1.861)	(1.868)
Location (=1 if urban)	25.765***	25.747***
	(6.383)	(6.372)
Caregiver is literate	1.939	1.942
	(3.311)	(3.302)
Log time (days) elapsed, baseline to endline	-28.930	-28.219
	(38.316)	(38.131)
Intercept	287.920	283.824
	(197.081)	(196.289)

Notes: *** p<.01, ** p<.05. Sample size is 2977

The weighted parameter estimates are the same as the unweighted estimates. So, for the RISE sample, attrition does not appear to be biasing these associations.

Supplementary Appendix 3: Accounting for attrition in the Young Lives sample

Attrition in the Young Lives sample is non-random, as it was in the RISE sample. For example, students with lower baseline scores are less likely to complete the school year and take the endline math test. They were more likely to complete the school year if they lived in an urban area, had attended pre-school, or came from wealthier households. Consequently, the sample that remains is self-selected in terms of math ability (as measured by the baseline test) and this creates a risk of biasing our parameter estimates. As noted above, one approach to resolving this is to create sample weights that address this selectivity problem; these can be constructed based on estimating two models of attrition: one where we estimate attrition as a function of the variables that we think affect our outcome of interest (here, endline math scores); and a second where we estimate attrition as a function of the variables that we think affect our outcome of interest ***and*** variables that we think ***only*** affect the likelihood that the child stays in school. We then estimate predicted probabilities of remaining in school from both models; the attrition adjusted weights are the ratio of these predicted probabilities.

Step 1: Our outcome of interest is the endline math score. We assume that this is a function of the same variables that appear in the regressions reported in the main text: baseline math score, the child's age, and sex at baseline, whether the child had attended pre-school, whether the child had previously dropped out for 3 months or more, whether the caregiver was literate; teacher characteristics (math content test score; qualifications); and school (log classroom size, school infrastructure score; principal qualification score) characteristics. Collectively, these are the **X** variables.

Step 2: There are fewer variables in the Young Lives data set that could conceivably only affect the likelihood that the child stays in school (and not the outcome variable itself). Again using the themes of physical access and connectedness, we identified three (**Z)** variables. These are:

- Does the child sleep in the same house or compound all year round, =1 if yes
- Number of older children in the household (note that unlike RISE, we do not know if these older siblings attend school
- Log travel time, in minutes, from home to the primary school

Step 3: Run the attrition regression with both the **X** and **Z** variables as regressors. Here the dependent variable =1 if the child completed the endline test, zero otherwise. So a negative coefficient means that that variable is associated with a lower likelihood that the child completed; a positive coefficient means that that variable is associated with a higher likelihood that the child completed.

Table S3.1: Correlates of attrition, Young Lives

Characteristics associated with attrition but not with learning	
Child lives in home all year round	-0.001
	(0.018)
Number of children older than the index child in household at baseline	-0.003
	(0.003)
Log travel time to primary school	-0.018**
	(0.007)
Child characteristics	
Baseline math score	0.000**
	(0.000)
Age	-0.033***
	(0.005)
Sex (=1 if boy)	-0.026**
	(0.011)
Attended pre-school	-0.034**
	(0.014)
Stopped schooling for >3 months	-0.021
	(0.015)
Household wealth index	-0.001
	(0.006)
Location (=1 if urban)	0.089**
	(0.038)
Caregiver literate	-0.008
	(0.016)
Teacher and school characteristics	
Teacher math content knowledge, z score	-0.004
	(0.010)
Math teacher qualification index	-0.002
	(0.005)
Log class size	-0.035
	(0.037)
School infrastructure index	-0.005
	(0.009)
Principal qualification index	0.002
	(0.008)
Intercept	1.280***
	(0.159)
Number of observations	4232

Notes: Cluster robust standard errors in parentheses. *** p<.01, ** p<.05.

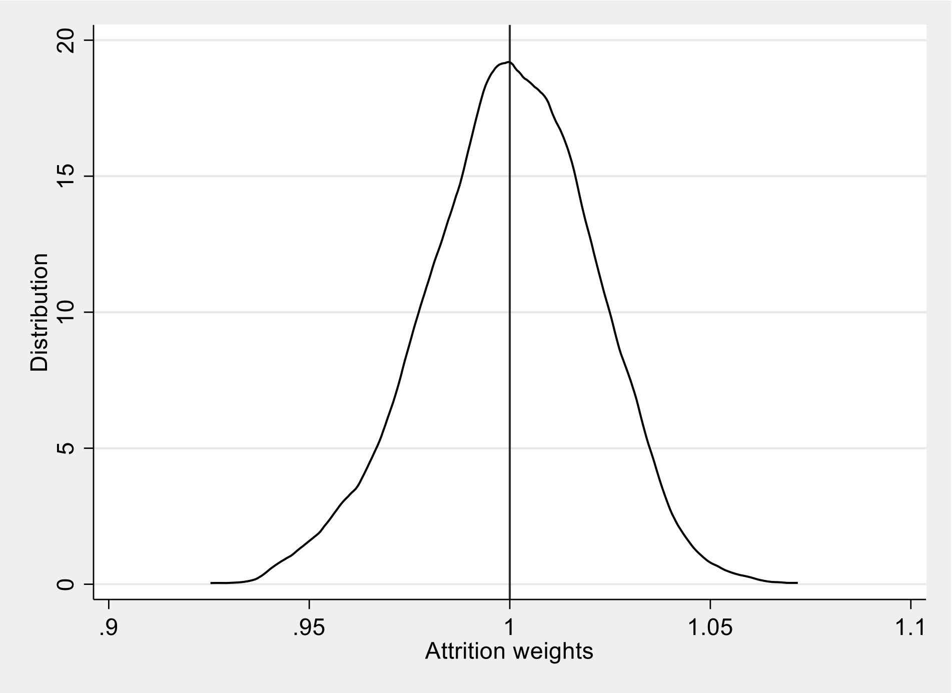
Neither permanence of residence or the number of older siblings in the household affect attrition. However, as travel times to schools increase, children were less likely to continue attending school, and thus less likely to take the endline test (put differently, they were more likely to drop out).

Step 4*:* The travel time variable is statistically significant; however, we cannot reject the null that the coefficients on the other two **Z** variables are equal to zero. The F test on the joint significance of the three **Z** variables = 2.42, P ==0.07. We have some evidence that we have variables that affect the likelihood of completion.

Steps 5, 6, and 7. These are available on request.

Step 8. Have calculated the weights, we can graph their distribution (using a kernel density function), shown below. There is some variability in the weights – they range from 0.90 to 1.07 – but the mass of the distribution is around 1. This hints at the possibility that the attrition weighted regression estimates will not differ too much from the unweighted regressions.

Figure S3.1: Kernel density functions for attrition weights, YL



Step 9. Regression results without and with weights, Young Lives

Table S3.2: Regression results without and with weights, RISE

Teacher and school characteristics	Unweighted results	Attrition weighted results
Teacher math content knowledge, z score	1.598	1.601
	(2.380)	(2.385)
Math teacher qualification index	4.925***	4.907***
	(1.538)	(1.538)
Log class size	-5.788	-5.691
	(8.977)	(8.999)
School infrastructure index	-1.695	-1.702
	(1.879)	(1.883)
Principal qualification index	-0.610	-0.631
	(1.820)	(1.817)
Child characteristics		
Baseline math score	0.773***	0.773***
	(0.022)	(0.022)
Age	0.320	0.324
	(0.896)	(0.891)
Sex (=1 if boy)	0.883	0.903
	(2.294)	(2.309)
Ever attended pre-school	8.618**	8.646**
	(3.449)	(3.456)
Stopped schooling for >3 months	-13.497***	-13.632***
	(3.970)	(4.022)
Household wealth index	0.878	0.921
	(1.041)	(1.038)
Location (=1 if urban)	18.208**	18.224**
	(6.879)	(6.894)
Caregiver is literate	-4.341**	-4.390**
	(2.066)	(2.068)
Log time (days) elapsed, baseline to endline	58.012	56.842
	(52.039)	(51.979)
Intercept	-145.305	-139.595
	(274.892)	(274.560)

Notes: *** p<.01, ** p<.05. Sample size is 3654

The weighted parameter estimates are essentially the same as the unweighted estimates. So, with the caveat that our ability to account for factors that affect attrition (but not outcomes) is more limited in the Young Lives sample, attrition does not appear to be biasing these associations.