Cross-Age Tutoring: Experimental Evidence from Kenya[*]

Mauricio Romero[†] Lisa Chen[‡] Noriko Magari[‡]

September 8, 2020

Abstract

Tailoring teaching to students' learning levels can improve learning outcomes in low-income countries. Cross-age tutoring, where older students tutor younger students, is a potential alternative for providing personalized instruction to younger students, though it comes at the cost of the older students' time. We present results from a large experiment in Kenya, in which schools were randomly assigned to implement either an English or a math tutoring program. Students in grades 3-7 tutored students in grades 1-2 and preschool. Math tutoring, relative to English tutoring, had a small positive effect (.063σ, p-value of .068) on math test scores. These results do not hold for English tutoring: relative to math tutoring, it had no positive effect on English test scores (we can rule out an effect greater than .074σ with 95% confidence). There is heterogeneity by students' baseline learning levels: The effect was largest for students in the middle of the ability distribution (.13σ for students in the third quintile, p-value of .042), while the effect was close to zero for students with either very low or very high baseline learning levels. We do not find any effect (positive or negative) on tutors.

Keywords: Cross-age tutoring
JEL Codes: C93, I20, I25

[*] Corresponding author: Mauricio Romero (mtromero@itam.mx). We are grateful to Prashant Bharadwaj, Mitch Downey, Isaac Mbiti, Craig McIntosh, Karthik Muralidharan, Justin Sandefur, and seminar participants at PacDev and UC San Diego. We are also grateful to Michael Goldstein, Geordie Bracken, and their team at Bridge International Academies for implementing the program. Mark Buttweiler at Bridge International Academies provided excellent data support. Romero acknowledges financial support from the Asociación Mexicana de Cultura. The views expressed here are ours, and not those of Bridge International Academies. Replication code is provided through Dataverse at https://doi.org/10.7910/DVN/IGZJ2J. All errors are our own.

[†] ITAM; J-PAL

[‡] Bridge International Academies

1 Introduction

Interventions which tailor teaching to students' learning levels are consistently signaled by the literature as having the largest effects on learning outcomes across different settings (for three recent reviews of the literature see Glewwe and Muralidharan (2016); Evans and Popova (2016); and Snilstveit et al. (2016)). However, teachers often lack the time (or incentives) to give children personalized instruction tailored to their needs and providing schools with extra teachers to do so is expensive. Cross-age tutoring, where older students tutor younger students, is a potential alternative to providing personalized instruction to younger students. It substitutes a trained instructor (the teacher) with an untrained one (the older student). The cost is the older students' time. However, tutoring can also provide benefits to tutors (e.g., mastering knowledge and increasing social skills). We present results from a large randomized control trial (over 180 schools, 15,000 tutees, and 15,000 tutors) in Kenya, in which schools are randomly selected to implement a cross-age tutoring program in either English or math.

In our setting, tutoring took place each school day of the 2016 academic year. At the end of every day, older students tutored younger students in either English or math for 40 minutes. Tutors were five grades above tutees. In some schools, the tutoring focused on math. In others, it focused in English. Whether math or English tutoring took place was randomized across schools. Section 2.2 provides details on the tutoring interventions. Since all the schools in our sample implement a tutoring program (i.e., there is no "pure" control group that receives no tutoring at all), all of our results should be interpreted as the impact of math tutoring relative to English tutoring (or vice versa).

Cross-age tutoring in math, relative to English tutoring, has a small positive effect (.063σ, p-value .068) on math test scores. These results do not hold true for English tutoring: relative to math tutoring, it has no positive effect on English test scores (we can rule out an effect greater than .074σ with 95% confidence). Moreover, the difference between the treatment effect on math and English (.069) is statistically significant (p-value .0024). There is heterogeneity according to the student's baseline learning level. The effect of math tutoring, relative to English tutoring, on math test scores is largest for students in the middle of the ability distribution (.13σ for students in the third quintile, p-value .042). The point estimate is close to zero for students with either very low or very high baseline learning levels. This suggests tutors are unable to: a) help students who are advanced learners and need an instructor with a high level of expertise to guide them through more advanced material; and b) help tutees lagging behind grade level competencies who may need more specialized instruction to catch up.

In addition, there is no heterogeneity by tutees' gender or age. Similarly, there is no heterogeneity by school characteristics (pupil-teacher ratio, class size, or tutor-tutee ratio). Finally, there is no heterogeneity by tutor's average age, gender or proficiency level (baseline test scores).[1]

There is no evidence that tutoring had an effect (positive or negative) on tutors. We can rule out an effect greater than $.091\sigma$ with a confidence of 95% on math test scores. Similarly, we can rule out an effect greater than $.087\sigma$ with a confidence of 95% on English test scores (for English tutoring, relative to math tutoring).

Two central issues to the research design are multi-tasking and cross-domain spillover effects. For example, treatment could induce tutees to concentrate in the subject they are being tutored on, lowering their performance in other subjects. Tutoring may also increase the performance of students in other subjects by releasing study time from the tutored subject or if there are synergies between knowledge in different subjects. While our research design does not explicitly let us rule out multi-tasking or spillover effects, we present a series of test that suggest these are first order issues in practice. First, tutoring did not take away teaching time from neither English nor math (nor any other subject). Second, had we found effects of English tutoring on English and math tutoring on math, a possible explanation, akin to multi-tasking, would have been that tutoring in one subject erodes performance in the other subject. This was not the case (there is no effect of English tutoring on English). Third, the effect of English tutoring on math test scores is likely either zero or slightly positive. This is because English reading skills may improve performance in math since textbooks are written in English. Since we find positive treatment effects of math tutoring relative to English tutoring on math, the "direct" treatment effect is large enough to compensate for the "indirect language effect". Finally, tutoring has no effect (positive or negative) on Kiswahili. The lack of effect on Kiswahili does not rule out the possibility of cross-domain spillover effects, but the effect on Kiswahili would need to be similar across English and math tutoring to yield no difference when comparing the two. Section 3.1 provides a formal discussion of cross-domain spillovers.[2]

Our results are relevant to two strands in the literature. First, they relate to the literature that studies the effect of personalized instruction on test scores. Across the developing world a large fraction of students are behind their grade-level standard and there

[1] Since we do not have data on tutor/tutee matches and teacher had discretion on how to match tutees to tutors, we show heterogeneity by the average characteristics of possible tutors for a specific tutee.

[2] One limitation of our experimental design is that even if there are benefits on learning outcomes from role model/peer effects, if these are the same across both tutoring programs, they would "cancel out". Similarly, any benefits for tutors coming from confidence or feeling valued may "cancel out" as well.

is considerable heterogeneity in learning levels within the same class (Muralidharan, Singh, & Ganimian, 2019). Teachers often follow the curriculum, regardless of students learning levels, making it almost impossible for students lagging behind to catch up (Pritchett & Beatty, 2015). Personalized instruction is used to narrow the curriculum gap for students lagging behind. Other interventions aimed at personalized instruction include the use computarized-assitance learning software (e.g., Banerjee, Cole, Duflo, and Linden (2007); Muralidharan et al. (2019)), tracking (e.g., Figlio and Page (2002); Zimmer (2003); Duflo, Dupas, and Kremer (2011)), additional contract teachers (e.g., Banerjee et al. (2007); Duflo et al. (2011); Muralidharan and Sundararaman (2013)), and "remedial camps" (Banerjee et al., 2016). We present evidence on a different approach to improve the amount of personalized instruction using resources readily available to schools. Although the program has modest effect sizes, it is relatively low-cost, and therefore may be cost-effective compared to some of the alternatives to provide personalized instruction. Taking into account that the total cost of the program is around 3 USD per student, assuming a linear-dose relationship implies that test score increases by 0.02σ per USD invested, making it a relatively cost-effective intervention.

Finally, we contribute to the literature on peer-learning programs. One variant of this literature focuses on peer-learning programs when students belong to the same grade or age-group.[3] We focus on corss-age tutoring, a subject on which the evidence is mixed and often relies on data from developed countries. An early review of the literature focused on studies based on observational data points to positive effects on student attitudes (Cohen, Kulik, & Kulik, 1982). A more recent review looking only at randomized control trials comes to the conclusion that cross-age math tutoring has non-significant effects on math test scores and cross-age tutoring in "reading" has a small (statistically significant) positive effect on reading (Shenderovich, Thurston, & Miller, 2015). However, only two of the studies reviewed by Shenderovich et al. (2015) had other elementary school students (as opposed to adults, community volunteers, or university students) as tutors and both tutoring programs focus on reading. None of the interventions in those studies were implemented in a low- or middle-income country. To the best of our knowledge, this is the first field experiment implemented on cross-age tutoring in which tutors are students in the same school as tutees. Furthermore, it is the first study of cross-age tutoring from a low-income country.

[3] For example, Li, Han, Zhang, and Rozelle (2014) find that sitting together high- and low-achieving students in the same class, and offering them group incentives for learning, improved test scores. Similarly, Fafchamps and Mo (2018) show, in the context of Chinese students taking a computer remedial course together, that matching children with (past) high and low grades increases the future performance of low achieving students without hurting the performance of the high achieving students.

2 Experimental Design

2.1 Context

Despite high net enrollment rates in primary schools (∼80% in 2012, World Bank (2015a)), the quality of education in Kenya is low: Children often fail to attain proficiency in early grade reading and numeracy (Uwezo, 2015). An annual nationwide learning assessments (the Uwezo test) consistently show that only half of grade 3 students can read a simple story at a grade 2 level in English — one of the national languages and the language of instruction in many schools (Trudell, 2016) — or successfully demonstrate grade 2 numerical skills (Jones, Schipper, Ruto, & Rajani, 2014).

Bold, Kimenyi, and Sandefur (2013) argue that the abolition of fees for primary schools in 2003 led to a decline in the quality ("or at least perceived quality") of public schools. In response, the demand (and supply) of private primary education increased dramatically (Lucas & Mbiti, 2012). According to World Bank statistics, the proportion of students enrolled in private primary schools more than doubled from 4.5% in 2004 to over 16% in 2014 (World Bank, 2015b). Kenya is not the only country where there has seen a surge in private school enrollment. Recently several chains of for-profit, low-cost private schools have emerged around the world. These chains leverage technology to deliver lessons and manage teachers (Mbiti, 2016).

In this study, we work with a large low-cost private school provider, Bridge International Academies (Bridge), in which schools within their network are randomly selected to implement either a math or an English tutoring program. Bridge opened its first school in Nairobi in January 2009. By November 2014, it had opened nearly 400 schools across Kenya and had enrolled over 100,000 students.[4]

Bridge tries to takes advantage of economies of scale in school management, teacher training, and lesson guides to lower the marginal cost of delivering education.[5] English is the language of instruction in all Bridge schools, which are located across East Africa, West Africa, and India, but mainly in Kenya. The company relies heavily on technology-enabled systems and processes and claims to maintain a constant feedback loop.[6]

[4] See http://www.bridgeinternationalacademies.com/company/history/

[5] For example, in each Bridge academy, there is only one employee involved in management. Bridge claims that the vast majority of non-instructional activities that the Bridge "Academy Manager" would normally have to deal with (billing, payments, expense management, payroll processing, and more) are automated and centralized. Similarly, Bridge hires experts to develop comprehensive teacher guidelines and training programs, which are then used in all of their schools. Schools charge on average a monthly fee of USD 6 and cater to families living on USD 2 a day per person or less.

[6] Bridge followed the 8-4-4 curriculum framework mandated by the national government at the time of the study, but provided detailed teacher guides for each lesson which are used by teachers across the

From a research standpoint, an advantage of working with Bridge data is that all students take the same tests across all schools, and Bridge collects data on students' performance to detect levels of content mastery. These data are also used to measure and improve on teacher quality. Students take 6 major exams per academic year. Each academic year has three terms, and each term has a midterm and an endterm exam. Additionally, at the beginning of the academic year, students in primary grades (Grades 1 - 6) take a diagnostic exam. Randomized control trials to study the effectiveness of different approaches to improve learning can be implemented relatively easily with low or no additional cost for data collection (often the most expensive part of a field experiment). This is the first of such trials implemented across schools in the Bridge network in Kenya or any other country in which it works.

A possible concern is that our experiment has limited external validity. Indeed, Bridge schools have a pupil-teacher ratio that is double of that in government schools, a school day that is about 2-3 hours longer, and teachers in Bridge schools are less educated (and paid less) than their counterparts in public schools. However, Bridge schools are similar (in terms of pupil-teacher ratio, school-day length, and teachers' education) to other (low-fee) private schools (Gray-Lobe, Keats, Kremer, Mbiti, & Ozier, 2020).

2.2 Intervention

The intervention took place every school day during the 2016 academic year. At the end of every school day, older students tutored younger students in either English or math for 40 minutes (3:35-4:15 pm).[7] Tutoring replaced an end-of-day independent study period. Tutors were five grades above tutees (Table 1 provides more details). In some schools the tutoring focused on math, while in others it focused in English. Whether math or English tutoring took place was randomized across schools. Therefore, within a school, all grades participated in either math or English tutoring. Table 2 provides details on the math tutoring intervention, while Table 3 provides details on the English tutoring intervention.

network. These guides are created by writers in several offices, including Nairobi, Kenya and Boston, USA. The guides are then streamed to individual teacher tablets. Teachers use tablets to upload students' information (e.g., test scores) to to a centralized data warehouse, which can then be accessed by shared services teams.

[7] At first, the tutoring was designed to be part of the normal school day. However, in 2016 the Kenyan government decreed class hours end at 3:30 pm (Secretary for Education, Science and Technology, 2015). Thus, the tutoring program became an after school program. While it was not mandatory, almost every child attended.

Table 1: Tutors and Tutees

Tutors		Tutees
Grade 3 (N=3,917)	→	Baby Class (BC) (N=2,419)
Grade 4 (N=3,721)	→	Nursery (NU) (N=3,176)
Grade 5 (N=3,341)	→	Preunit (PU) (N=3,534)
Grade 6 (N=2,718)	→	Grade 1 (N=3,906)
Grade 7 (N=2,409)	→	Grade 2 (N=3,919)

The main objective of the math (English) tutoring program was to raise math (English) achievement in tutees (BC-Grade 2 students). A secondary objective was to develop communication and leadership skills in tutors (Grade 3-Grade 7 students) and build a school community through sibling-like relationships between tutees and tutors.

During the first two weeks of the academic year 2016, the tutoring sessions consisted of "tutor training", led by teachers. During this tutor training, teachers instructed tutors to keep tutees focused and use the "ask-tell-show-repeat" procedure to correct the tutee's work. "Ask-tell-show-repeat" is a four-step process following an incorrect answer by the tutee: (1) Tutee is asked to do the problem again; (2) Tutee receives verbal instructions on the correct solution if the mistake is repeated; (3) Tutee is shown the correct solution if they make a mistake again; and (4) Tutee is asked to repeat the problem one last time. The idea was to provide a simple structure for tutor-tutee interaction.

Beyond training tutors during the first two weeks, teachers supervised the tutoring sessions to maintain order and provide assistance. Teachers also chose how to pair tutees with tutors. The matching between tutees and tutors could vary every day. Therefore, any difference in outcomes across treatments could also capture different matching processes across treatments. While we do not have any data on the actual matches, anecdotal evidence from interviews with teachers suggest the matching was more or less random across both treatments.

After the first two weeks, tutors were given guides with problems and activities to do with tutees each day (e.g., addition, counting and tracing numbers in math, identifying letters, dictation, and reading in English). Roughly, the tutoring in both English and math had the same structure. First, there is a small introduction ($\sim$ 3 mins). Afterwards, tutors went over the exercises with tutees. Tutors were asked to keep tutees engaged, and help tutees if they struggled to get the correct answers.

For math, changes were introduced during the last term of the school year.[8] In the first

[8] Academic field officers visit schools on a regular basis to conduct classroom observations to see how lesson guides, tutoring sessions, and other academic can be improved. The changes to math and English tutoring were introduced in response to feedback from these visits.

two terms, teachers gave a demonstration of the topic that was covered that day. In the last term, brief instructions for tutors replaced the teacher demonstration. This was done to shift the focus of the tutoring session from the teacher to the tutoring pairs, giving pupils more time to engage in the productive struggle to learn new skills. In addition, tutors were instructed to shift to a from the “ask-tell-show-repeat” correction method to “ask-show-repeat”. Specifically, instead of first verbally instructing the tutee in how to obtain a correct solution in case of a repeat mistake, the tutor went straight to showing them the correct solution. Finally, in the first terms were asked to circulate and make sure both tutees and tutors were behaving, and tutees were understanding the material covered. In the last term, teachers were instructed to “check-respond-leave” with tutors exclusively, thus empowering tutors to take responsibility for their tutees’ performance (Table 2 provides more details).

For English, changes were introduced during the last two terms of the school year. Most of the changes varied how much time was allocated to different activities. Some of the time allocated to writing in Grades 1 and 2 (it went from 15 minutes to 10 minutes), was shifted to reading (which went from 9 minutes to 15 minutes). Dialogue practice, which took place in Preunit, Grade 1 and Grade 2 was also removed after the second term to allocate more time to other activities. Finally, more time was allocated to finding rhyme words in the last term for baby class and nursery (Table 3 provides more details).

Regarding compliance, we have data on subject matter, time, and tutor-tutee ratio from school visits. Bridge academic field officers visited Bridge schools on a regular basis. Their job was to observe classrooms and see how the scripted lessons were taught by the teachers(e.g., did the script translate to classroom practice as envisioned by the master teacher, is the time allocated for particular tasks insufficient or too long, etc.). During their visits, they also collected data from the cross-age tutoring scheme. Overall, every school complied with their treatment status: tutoring took place in the subject they were assigned to. On average, tutoring took place for 28 minutes (as opposed to the 40 minutes scheduled). In less than 20% of schools, there was more than one tutee per tutor.

Table 2: Math tutoring intervention

	Term 1 and Term 2	Term 3
Timing	3:35 - 4:15pm.	3:35 - 4:15pm.
Grades 1/2	Introduction: 3 min Teacher demo: 5 min Tutoring: 30 min Guide with 18 problems 1 topic	Introduction: 3 min Tutoring 1: 22 min Tutoring 2: 15 min Guide with 60 problems 2 topics
PU	Introduction: 3 min Warm-up Exercise: 10 min Teacher demo: 5 min Tutoring: 15 min Guide wtih 10 problems 1 topic	Introduction: 3 min Tutoring 1: 22 min Tutoring 2: 15 min Guide with 56 problems 2 topics
BC/NU	Introduction: 3 min Counting with tutors: 7 min Rhyme: 3 min ID numbers with tutors: 7 min ID frames with tutors: 7 min Rhyme: 3 min ID shapes with tutors: 8 min Closing: 2 min	Introduction: 3 min Counting with tutors: 7 min Rhyme: 3 min Writing numbers with tutors: 7 min Drawing frames with tutors: 7 min Rhyme: 3 min Drawing shapes with tutors: 8 min Closing: 2 min
Tutor duties	Keep tutees focused Use ask-tell-show-repeat	Correct tutee after every two problems Use ask-show-repeat
Teacher duties	Do teacher demo Circulate	Check-respond-leave with tutors only

Table 3: English tutoring intervention

	Term 1	Term 2	Term 3
Timing	3:35 - 4:15pm.	3:35 - 4:15pm.	3:35 - 4:15pm.
Grades 1/2	Introduction: 2 min Dialogue practice: 5 min Tutoring instructions: 3 min Words: 5 min Reading: 15 min Writing: 9 min	Introduction: 2 min Dialogue practice: 5 min Words: 8 min Writing: 15 min Reading: 9 min	Introduction: 3 min Words: 10 min Writing: 10 min Reading: 15 min Closing: 2 min
PU	Introduction: 3 min Dialogue practice: 5 min Practice book: 7 min Sight words: 5 min Reading: 15 min	Introduction: 3 min Dialogue practice: 5 min Sight words: 12 min Reading: 15 min	Introduction: 2 min Words: 8 min Reading: 15 min Sight words: 15 min
BC/NU	Introduction & song: 5 min Practice set: 7 min Finding words: 4 min Rhyme: 3 min Finding letters: 5 min Letter sound chant: 2 min Dialogue practice: 5 min Closing: 2 min	Introduction & song: 5 min Words: 11 min Rhyme: 3 min Finding letters: 5 min Letter sound chant: 2 min Dialogue practice: 5 min Closing: 2 min	Introduction & song: 5 min Words: 11 min Rhyme: 3 min Finding rhyme words: 7 min Letter sound chant: 2 min Finding letters: 5 min Dialogue practice: 5 min Closing: 2 min
Tutor duties	Keep tutees focused Use ask-tell-show-repeat Correction method	Keep tutees focused Use ask-tell-show-repeat Correction method	Keep tutees focused Use ask-tell-show-repeat Correction method
Teacher duties	Circulate	Circulate	Circulate

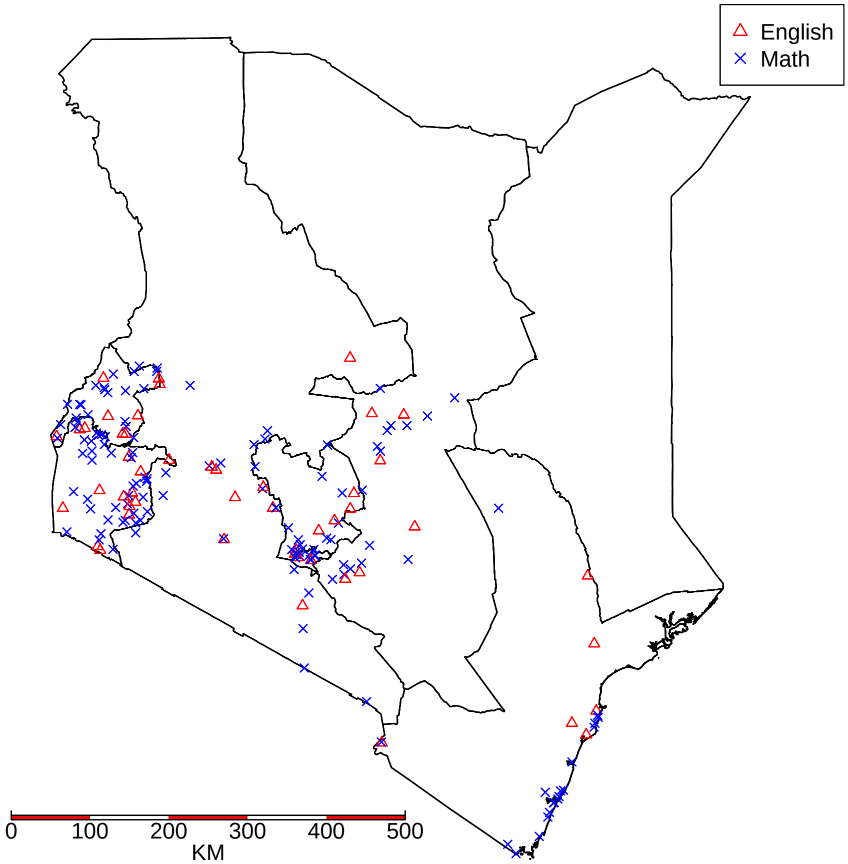
2.3 Sampling

In 2016, Bridge had a network of over 400 schools across Kenya. However, only 187 schools were eligible to participate in the trial.[9] Randomization was stratified at the "former province" level (Kenya's provinces were replaced by a system of counties in 2013) and by average baseline test scores at each academy. Estimations take into account the randomization design by including the appropriate fixed effects (Bruhn & McKenzie, 2009). Figure 1 shows the distribution of schools across the country. Math tutoring took place in 137 academies, while English tutoring took place in 50 academies.[10]

[9] Schools in which a pilot of the program was tested during the 2015 academic year were excluded, as well as schools where other programs were being tested.

[10] Math tutoring took place in more schools as Bridge expected this intervention to be more effective.

Figure 1: Geographical distribution of schools with math and English tutoring across Kenya



Note: Data on school location was provided by Bridge International Academies. Geographical information from the administrative areas of Kenya comes from DIVA-GIS (2016).

2.4 Data and summary statistics

As mentioned above, students have six major exams per academic year. Each academic year has three terms, and each term has a midterm and an endterm exam. Additionally, at the beginning of the academic year students in primary grades (Grades 1 - 6) take a diagnostic exam. Table 4 shows the dates of each exam. Two exams (T3ET15 and T1DG16) were taken by students before tutoring began, and six exams were taken after. Since students in Preunit, Nursery and Baby-Class are not tested at the beginning of 2016 (T1DG16), we use both T1DG16 and T3ET15 as our baseline test scores. For students in Baby Class (BC) we have no baseline test scores.

The exams for all grades are designed by education professionals working at Bridge. Teachers are given answer-keys to minimize grading errors. Teachers grade the tests and then input the total score into their teacher tablet. The data for students in Preunit, Nursery and Baby-Class comes from one-on-one tests in which a teacher sits with the student, asks questions, and records the answers. These exams test emerging numeracy and literacy skills (e.g., a picture vocabulary test for literacy and counting for numeracy, see Table A.1 for details on the skills tests). For Grades 1-7 students are given a more standard written exam. Exams are predominantly multiple choice for primary school kids (averaging 45 questions per exam, depending on subject and grade level) and generally lasts 30-40 minutes. These exams cover grade-appropriate content (e.g., reading comprehension of a grade-appropriate story, or single-digit addition for grade 1 and two-digit addition for grade 3). We provide specific details of what skills are tested in each grade in Table A.1.

All students at each grade level across schools in Bridge's network take the same exam, making test scores for students in different schools comparable. However, the exams are not vertically linked (i.e., there are no overlapping questions across exams in different grades or across time). As mentioned above, teachers only record the total score for the students, and not the answer to individual questions. Thus, we are unable to use Item Response Theory to estimate students' abilities (van der Linden, 2017). Therefore we standardized test scores in each term (to obtain mean zero and standard deviation of 1 in English tutoring schools) within each grade.

Table 4: Learning assessments

Year	**Term**	**Exam**	**Dates**	**Code**	Grade (2016 academic year)
2015	3	Endterm	Nov 10-12, 2015	T3ET15	NU, PU, Grades 1-6
2016	1	Diagnostic	Jan 13-14, 2016	T1DG16	Grades 1-6
2016	1	Midterm	Feb 16-18, 2016	T1MT16	BC, NU, PU, Grades 1-6
2016	1	Endterm	Apr 5-7, 2016	T1ET16	BC, NU, PU, Grades 1-6
2016	2	Midterm	Jun 14-16, 2016	T2MT16	BC, NU, PU, Grades 1-6
2016	2	Endterm	Aug 9-11, 2016	T2ET16	BC, NU, PU, Grades 1-6
2016	3	Midterm	Sept 26-27, 2016	T3MT16	BC, NU, PU, Grades 1-6
2016	3	Endterm	Oct 25-27, 2016	T3ET16	BC, NU, PU, Grades 1-6

Schools randomly assigned to math tutoring are similar to those assigned to English tutoring: They were inaugurated around the same time (in operation for two years by January 1, 2016), and have similar teacher salaries and pupil-teacher ratios (PTR) of 22 students per teacher (Table 5). Tutees (Panel A and B, Table 6) in English and math tutoring schools are similar across all characteristics. Tutors (Panel C and D, Table 6) are also similar across English and math tutoring schools.[11] On average tutees are 6.5 years old and tutors are 4.5 years older than their tutees.

Table 5: School characteristics in English and math tutoring schools

	(1)	(2)	(3)	(4)
	English	Math	Difference	Difference
	Tutoring	Tutoring		(F.E)
Days since launch date (since 01/01/2016)	672.960	693.310	20.347	-16.257
	(406.417)	(405.017)	(66.887)	(46.838)
Monthly teacher wage of 11250 KSH	0.060	0.110	0.049	0.014
	(0.240)	(0.313)	(0.043)	(0.026)
Monthly teacher wage of 10400 KSH	0.180	0.100	-0.078	-0.073
	(0.388)	(0.304)	(0.061)	(0.061)
Monthly teacher wage of 7970 KSH	0.760	0.790	0.028	0.058
	(0.431)	(0.410)	(0.070)	(0.065)
Teachers	7.440	7.530	0.093	0.077
	(0.541)	(0.619)	(0.093)	(0.092)
Enrollment	167.760	167.180	-0.585	-2.554
	(75.793)	(84.627)	(12.894)	(11.451)
PTR	22.240	21.980	-0.257	-0.478
	(9.363)	(10.367)	(1.589)	(1.401)

Days since launch date indicates the number of days that have passed since the schools opened, as of January 1, 2016. Bridge had three teacher wage categories at the time. “Monthly teacher wage” shows the proportion of schools within each wage schedule. Teachers was the number of teachers at the school, and Enrollment was the enrollment across all grades for the school at the beginning of the school year. PTR is the pupil-teacher ratio. Each row presents the mean for schools which receive English tutoring (Column 1), schools which receive math tutoring (Column 2), the difference between the two (Column 3), and the difference taking into account the randomization design (i.e., including strata fixed effects) in Column 4. In the first two columns the standard deviation is shown in parentheses, while in the third and fourth columns the standard error of the difference is in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[11] Tutors are more likely to be male in math tutoring schools. However, given that we are testing for differences across 23 school, tutee, and tutor characteristics it is unsurprising that the difference across English and math tutoring schools in one characteristic is statistically significant. Indeed, this difference is not statistically significant after adjusting for multiple hypothesis testing following Romano and Wolf (2005).

Table 6: Pupil's characteristics

	(1)	(2)	(3)	(4)
	English	Math	Difference	Difference
	Tutoring	Tutoring		(F.E)
Panel A: Tutees' time invariant characteristics				
Age	6.600	6.500	-0.097*	-0.024
	(1.617)	(1.595)	(0.054)	(0.037)
Male	0.520	0.520	0.002	0.000
	(0.500)	(0.500)	(0.011)	(0.010)
Age entered Bridge	5.440	5.390	-0.057	0.013
	(1.669)	(1.643)	(0.076)	(0.073)
Panel B: Tutees' test-scores in T3ET15				
English (reading)	0.000	-0.010	-0.013	-0.058
	(1.000)	(1.021)	(0.074)	(0.071)
English (writing)	0.000	-0.040	-0.038	-0.064
	(0.999)	(1.014)	(0.064)	(0.058)
Swahili (reading)	0.000	-0.020	-0.025	-0.064
	(1.000)	(1.020)	(0.083)	(0.082)
Swahili (writing)	0.000	-0.070	-0.072	-0.113
	(1.000)	(1.102)	(0.111)	(0.090)
Math	0.000	0.040	0.041	0.011
	(0.999)	(0.974)	(0.056)	(0.052)
Panel C: Tutors' time invariant characteristics				
Age	11.040	11.070	0.030	0.023
	(1.980)	(2.017)	(0.097)	(0.062)
Male	0.500	0.520	0.020**	0.023***
	(0.500)	(0.500)	(0.009)	(0.008)
Age entered Bridge	9.660	9.710	0.053	0.045
	(2.269)	(2.316)	(0.140)	(0.098)
Panel E: Tutors' test scores in T3ET15				
English (reading)	0.000	0.070	0.070	0.047
	(0.999)	(1.038)	(0.051)	(0.046)
English (writing)	0.000	0.070	0.069	0.034
	(0.999)	(0.967)	(0.054)	(0.045)
Swahili (reading)	0.000	0.050	0.055	0.053
	(0.999)	(1.042)	(0.056)	(0.046)
Swahili (writing)	0.000	0.140	0.138*	0.114*
	(0.999)	(0.941)	(0.081)	(0.059)
Math	0.000	0.050	0.047	0.027
	(0.999)	(1.009)	(0.063)	(0.048)

Math, English, and Kiswahili represent the standardized test scores (mean zero and standard deviation 1 in English tutoring schools). Each row presents the mean for schools which received English tutoring (Column 1), schools which received math tutoring (Column 2), the difference between the two (Column 3), and the difference taking into account the randomization design (i.e., including strata fixed effects) in Column 4. In the first two columns the standard deviation is shown in parentheses, while in the third and fourth columns the standard error, clustered at the school level, of the difference is in parentheses. Table A.2 shows tutees and tutors' test scores are also balanced in T1DG16. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We have an unbalanced panel, where few students have test score data for all periods. This is due to a combination of compliance (i.e., teachers not entering the data), software updates, and internet failures in which the teacher enters the data but it fails to upload to Bridge's servers.[12] Table 7 shows the fraction of students tested each time. More than 25% of the data are missing (and often more than 30%). In particular, the endterm exam in the second period (T2ET16) is missing over 60% of test scores for math due to a glitch in the programming update which prevented over a quarter of the schools from entering test score data. The T2ET16 data-missing rates are different across math and English tutoring schools (see Figure A.2). However, whether the data is missing is uncorrelated to whether the student is receiving math or English tutoring in other periods (see Table 8). Given that the data from T2ET16 is noisy and has differential attrition across treatments we remove it from our sample in the main text, but we provide robustness checks that include the data in Appendix B.

Table 7: Non-missing data

	T1MT16	T1ET16	T2MT16	T2ET16	T3MT16	T3ET16	Total
Math	0.751	0.591	0.711	0.399	0.570	0.532	0.590
	(0.432)	(0.492)	(0.453)	(0.490)	(0.495)	(0.499)	(0.492)
English (Writing)	0.739	0.575	0.710	0.472	0.564	0.517	0.594
	(0.439)	(0.494)	(0.454)	(0.499)	(0.496)	(0.500)	(0.491)
English (Reading)	0.738	0.566	0.709	0.449	0.553	0.512	0.586
	(0.439)	(0.496)	(0.454)	(0.497)	(0.497)	(0.500)	(0.493)
Observations	192346						

This table shows the fraction of students in the data set (i.e., those tested at some point in the 2016 academic year) with scores for math, English (reading), and English (writing) in each test. A glitch in the software prevented more than 25% of the schools from entering test-score data for T2ET16.

Since missing data is prevalent in any given period (over 30%) we do not perform Lee (2009) bounds as these are too wide to be informative. However, we do not believe differential attrition is a first order concern when interpreting our results. First, as mentioned above, the rate of missing data is the same across treatments (see Figure A.2, as

[12] In addition, students may be absent from school on the day of the test. However, in most cases if test score data is missing for a student, it is also missing for their entire grade. For the purposes of this paper, the missing data numbers include tutees who are not currently active (i.e., have not paid fees) in a given period.

well as Tables 8 and B.1). Second, there is no evidence of selection bias among as student characteristics (age and gender) are not correlated with attrition (see Table A.3).

Since a large number of students do not have baseline test scores we avoid dropping these observations by adding a dummy variable to all our regressions for whether the baseline test score was missing, replacing the missing test score with zero (but the replacement value does not affect the estimates), and interacting the dummy with the modified test score.

Our results are also robust to using interpolation to reduce sample attrition due to missing outcome data. If the outcome data for a student in a given term is missing, but we have outcome data for the terms before and after, we input the average score for the missing term using a simple linear interpolation. For example, if data for T2ET16 is missing, we input the value of the average score of T2MT16 and T3MT16 (after standardizing both exams).

Table 8: Differential missing data rate between treatment and control students

	(1)	(2)	(3)
	Math	English	Swahili
Math tutoring	-0.0027	-0.0053	-0.0098
	(0.022)	(0.022)	(0.028)
Mean English	0.63	0.61	0.61
N. of obs.	81195	81209	55019
Number of schools	187	187	187

This table shows the differential missing data rate between students in math tutoring schools compared to students in English tutoring schools. The estimation data set does not include T2ET16 data. Table B.1 provides estimates that includes T2ET16 data. Clustered standard errors, by school, in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3 Results

3.1 Main treatment effects

In order to estimate the effect of tutoring on test scores we use the following specification:

$$Y_{ijsgd,t} = \alpha_0 + \beta_j T_s + \alpha_1 Y_{ijsgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t} \tag{1}$$

where $Y_{ijsgd,t}$ is the test score of student i in subject j in grade g at school s located in province d at time t (and $Y_{isgd,t=0}$ is his test score before treatment), γ_d is a set of province and strata fixed effects, γ_t are time fixed effects, and γ_g are grade fixed effects. We include time fixed effects as the test scores are not comparable across time. Likewise, we include grade fixed effects as the test scores are not comparable across grades. However, as the test scores are standardized within each term for each grade, these fixed effects have almost no effect on the estimated treatment effects. X_i is a set of student time-invariant characteristics (month of birth and gender), and X_s are school characteristics at baseline (pupil-teacher ratio, monthly school fees and teachers' wages). T_s indicates whether the student is in a school with a math tutoring program (if not, he is in a school with English tutoring). Standard errors are clustered at the school level. The coefficient of interest is β_j, which estimates the effect of math tutoring relative to English tutoring on test scores in subject j. This specification assumes that the treatment effect (β_j) is time invariant and grade invariant (in Section 3.2 we relax these assumptions).

As mentioned above, β_j estimates the effect of math tutoring relative to English tutoring on test scores in subject j. Formally, let $\beta_{m,m}$ is the impact of math tutoring on math test scores, $\beta_{e,m}$ is the impact of English tutoring on math test scores, $\beta_{m,e}$ is the impact of math tutoring on English test scores, and $\beta_{e,e}$ is the impact of English tutoring on English test scores. Then $\beta_{math} = \beta_{m,m} - \beta_{e,m}$ and $\beta_{english} = \beta_{m,e} - \beta_{e,e}$. This is what the experimental design allows us to estimate. However, the effect of math tutoring on English test scores is likely zero (i.e., $\beta_{m,e} = 0$). We do not expect students to improve their English skills, while practicing math in their tutoring sessions. Thus, $\beta_{english}$ is likely a good proxy for $-\beta_{e,e}$.

In addition, the effect of English tutoring on math test scores ($\beta_{e,m}$) is likely zero or slightly positive (English reading skills could help students on math tests since the tests and textbooks are written in English). Therefore, $\beta_{math} < \beta_{m,m}$. Thus, if we find any positive effect of math tutoring, relative to English tutoring, on math test scores, this will be a lower bound of $\beta_{m,m}$.[13]

In sum, formally we can only estimate the effect of math tutoring, relative to English tutoring. However, under some reasonable assumptions, the effect on math test scores (β_{math}) is as a lower bound of the effect of math tutoring on math ($\beta_{m,m}$). Likewise, the

[13] As mentioned in Section 2.2, tutoring took place at the end of the school day and did not take away teaching time from either English or math (or any other subject in particular). If this was not the case, $\beta_{e,m}$ and $\beta_{m,e}$ could be negative.

negative of the effect on English scores ($-\beta_{English}$) is a good estimate for the treatment effect of English tutoring on English ($\beta_{e,e}$).

3.1.1 Tutees

Math tutoring, relative to English tutoring, has a small positive effect on math test scores of .063σ (see Column 1 in Table 9). English tutoring, relative to math tutoring, has no effect on English test scores — we can rule out an effect greater than .074σ with a confidence of 95% (see Column 2 in Table 9). The difference between the treatment effect of math tutoring on math and English tutoring on English (.069) is statistically significant (p-value .0024). Math tutoring, relative to English tutoring, seem to have no effect on Kiswahili (see Column 3 in Table 9). These results are robust (effect sizes and p-values are similar) to including data from all terms, including T2ET16 (see Table B.2), using interpolation to reduce sample attrition due to missing outcome data (see Section 2.4 for details on how the interpolation is done, and Table B.3 in the Appendix for the results), and to different controls (see Table A.4).

To summarize, our findings suggest math tutoring is more effective than English tutoring in raising test scores (in the subject of tutoring) in this setting.

Table 9: Effect on test scores

	Tutees			Tutors		
	(1) Math	(2) English	(3) Swahili	(4) Math	(5) English	(6) Swahili
Math tutoring	0.063*	-0.0061	0.035	0.029	-0.019	-0.020
	(0.034)	(0.035)	(0.047)	(0.031)	(0.035)	(0.036)
N. of obs.	50424	48204	32736	48741	46938	46512
Number of schools	187	187	186	187	187	187

The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. The number of observations in Column 3 is smaller as students in Baby Class are not tested in Kiswahili. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does not include T2ET16 data. Tables B.2 and B.4 provide versions of these estimates that include T2ET16 data. Tables B.3 and B.5 provide versions of these estimates that include T2ET16 data and use interpolation to reduce sample attrition due to missing outcome data. Table A.4 provides treatment estimates varying the controls used in the regression. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.1.2 Tutors

We do not find an impact of math mentoring, relative to English tutoring, on tutor test scores. We can rule out an effect greater than $.091\sigma$ with a confidence of 95% on math test scores. Similarly, we can rule out an effect greater than $.087\sigma$ with a confidence of 95% on English test scores (for English tutoring, relative to math tutoring). See Columns 4 and 5 in Table 9 for details.

3.2 Heterogeneity

In this section we test for heterogeneous treatment effects in tutees. Overall, there is some evidence that the math tutoring program, relative to the English tutoring program, is most effective after the first term (except for T2ET16, the exam with a high missing data rate and therefore unreliable results). However, the difference in the treatment effect across periods is not statistically significant. In addition the evidence suggests that math tutoring, relative to English tutoring, is most effective for students in the middle of the ability distribution at baseline. We do not find any heterogeneity by grade, age, gender, average tutor characteristics (age, gender, baseline test scores), or average school characteristics (pupil-teacher ratio, school size, or tutor-tutee ratio).

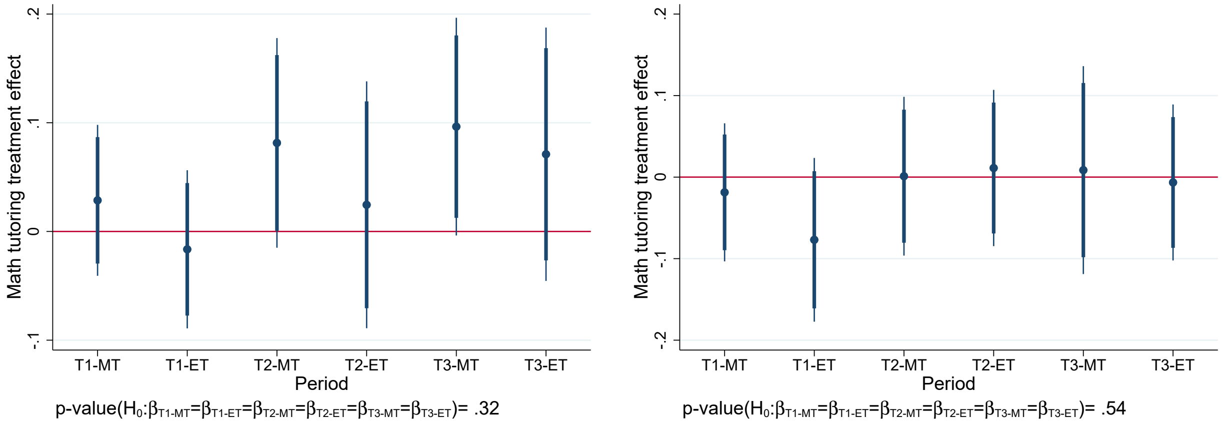
3.2.1 Periods

In order to estimate the effect of tutoring on test scores across time we use the following specification

$$Y_{isgd,t} = \alpha_0 + \sum_{\tau=1}^{6} \beta_\tau T_s \times \mathbb{1}_{t=\tau} + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \qquad (2)$$

where $\mathbb{1}_{t=\tau}$ is equal to one when the time period is equal to τ and zero otherwise. Thus, β_1 measures the treatment effect of math tutoring, relative to English tutoring, in period T1MT15, β_2 measures the effect in T1ET15, and so on, until β_6 which measures the effect in period T3ET15. The treatment effect on math test scores of math tutoring relative to English tutoring increases after the first marking period (except for T2ET16, the period with a high missing data rate). However, we cannot reject the null that the treatment effect is the same across all periods and after adjusting for multiple hypothesis testing the treatment effect is not significant in any period. On the other hand, math tutoring, relative to English tutoring, does not seem to have a negative effect on English test scores, with point estimates close to zero after the first marking period. See Figure 2 for more details.

Figure 2: Evolution of the treatment effect of math tutoring, relative to English tutoring



(a) Math

(b) English

Note: Math (left panel) and English (right panel) test scores (y-axis) by period (x-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). Each figure displays at the bottom the p-value for testing the null hypothesis that the treatment effect is the same in all periods. The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for math in T1-MT is .42 (.67), in T1-ET is .66 (.83), in T2-MT is .097 (.41), in T2-ET is .67 (.83), in T3-MT is .059 (.36), and in T3-ET is .23 (.63). The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for English in T1-MT is .66 (1), in T1-ET is .13 (.66), in T2-MT is .98 (1), in T2-ET is .82 (1), in T3-MT is .89 (1), and in T3-ET is .89 (1).

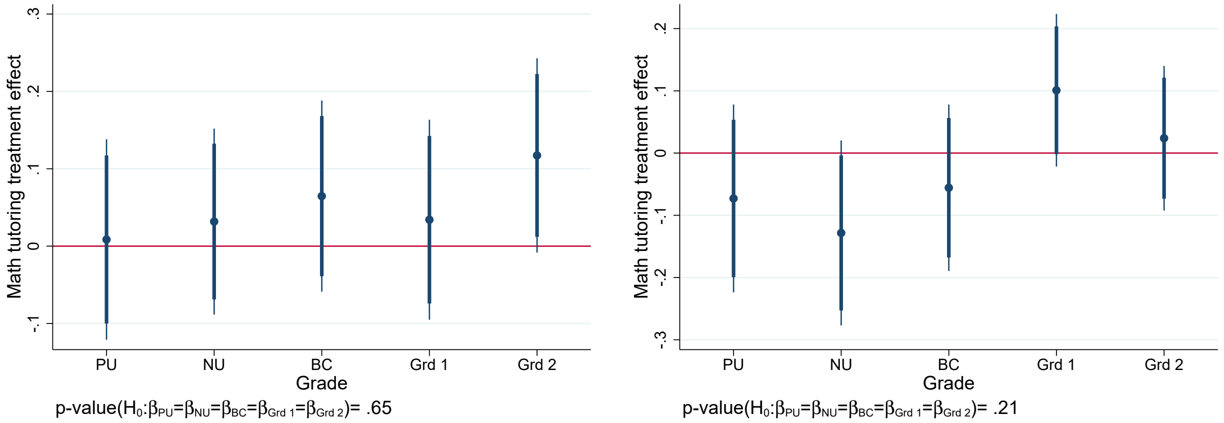
3.2.2 Grade

In order to estimate the effect of tutoring on test scores across grades we use the following specification

$$Y_{isgd,t} = \alpha_0 + \sum_{g=1}^{5} \beta_g T_s \times 1_{grade=g} + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \quad (3)$$

where β_1 measures the treatment effect of math tutoring, relative to English tutoring, for BC, β_2 for NU, β_3 for PU, β_4 for Grade 1 and β_5 for Grade 2. Although the point estimate of the treatment effect on math test scores is the largest for Grade 2, there does not seem to be a systematic pattern in which oldest students benefit more than younger ones from math tutoring, and we cannot reject the hypothesis that the effect is the same across grades. Similarly, there seems to be no systematic pattern in the effect on English test scores. See Figure 3 for more details.

Figure 3: Treatment effect of math tutoring, relative to English tutoring, by grade



(a) Math

(b) English

Note: Math (left panel) and English (right panel) test scores (y-axis) by grade (x-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). Each figure displays at the bottom the p-value for testing the null hypothesis that the treatment effect is the same in all grades. The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for math in preunit (PU) is .9 (1), in nursery (NU) is .6 (.85), in baby class (BC) is .3 (.75), in Grade 1 is .6 (.9), and in Grade 2 is .067 (.29). The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for English in preunit (PU) is .34 (.57), in nursery (NU) is .09 (.37), in baby class (BC) is .41 (.62), in Grade 1 is .11 (.45), and in Grade 2 is .69 (.63). Figure B.1 provides a version of these estimates that includes T2ET16 data.

3.2.3 Baseline test scores

In order to estimate the effect of tutoring on test scores across baseline test scores we use the following specification:

$$Y_{isgd,t} = \alpha_0 + \sum_{i=0}^{5} \beta_i T_s \times c_i + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \tag{4}$$

where c_i is the decile of the student's test score in math in T3ET15. We have 6 categories for c_i: 5 quintiles and a category for those students with missing test scores.

Figure A.3 shows the estimates for all the βs which correspond to the treatment effect of math tutoring, relative to English tutoring, for students in a given category. Students in the middle of the distribution benefit more from math tutoring (.13σ for students in the third quintile compared to the average effect of .063σ).[14]

We can reject the null that the treatment effect is the same for all quintiles (p-value .099) and the null that the treatment effect for students in the first, the third, and the fifth

[14] For students in the bottom 25% and the top 25% at baseline there is a small, insignificant, negative effect.

quintiles is the same (p-value .071). The treatment effect for students in the middle of the distribution is statistically significant after adjusting for multiple hypothesis testing with an adjusted p-value of .042 (the raw p-value is .014).

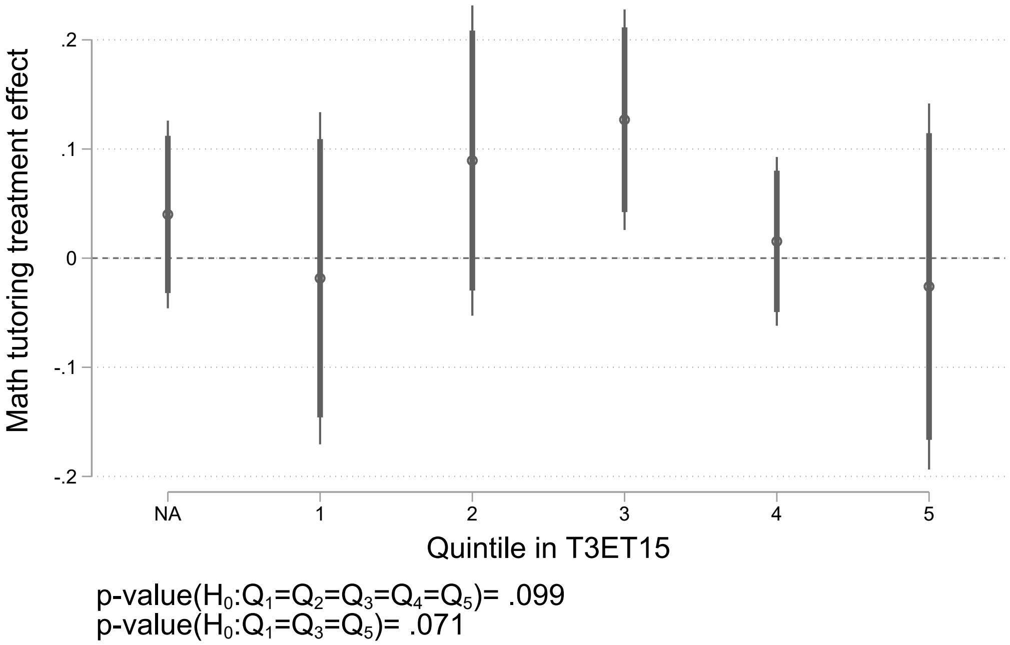
That students in the middle of the distribution benefit the most is robust to using deciles (see Figure 4) and terciles (see Figure A.4), as well as to interacting the treatment dummy with a fourth-order polynomial of the baseline test score (see Figure A.5).

Students in the middle benefiting the most is consistent with tutors unable to: a) help students who are advanced learners and need an instructor with a high level of expertise to guide them through more advanced concepts; and b) help tutees lagging behind grade level competencies who may need more specialized instruction to catch up.[15]

While low achieving tutors may benefit from reviewing material they do not master completely, we do not find evidence of this (see Figure A.6 in Appendix A). The effect is indistinguishable from zero for all tutors, regardless from baseline test scores, without any discernible pattern.

[15] In addition, there is some evidence that more advanced tutees, when matched with more advanced tutors, benefit more from math tutoring (Table A.6). This aligns with the intuition above. That is, more advanced tutors are able to help students who are advanced learners and need an instructor with a high level of expertise to guide them through more advanced concepts.

Figure 4: Treatment effect of math tutoring, relative to English tutoring by baseline ability quintile



Note: Treatment effect of math tutoring on math test (y-axis) scores by ability quintile in T3ET15 (x-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). The figure displays at the bottom the p-value for testing the null hypothesis that the treatment effect is the same in across all quintiles, as well as the p-value testing whether the treatment effect for the first, the third and the fifth quntile is the same. The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for the first quintile is .81 (.57), for the second quintile is .22 (.36), for the third quintile is .014 (.042), for the fourth quintile is .7 (.83), and for the fifth quintile is .76 (.56).

3.2.4 Tutee, tutor and school characteristics

In order to estimate the effect of tutoring on test scores across tutee, tutor and school characteristics we use the following specification:

$$Y_{isgd,t} = \alpha_0 + \beta_1 T_s + \beta_2 T_s \times c_i + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \qquad (5)$$

where c_i denotes the characteristics along which we wish to measure heterogeneity and β_2 allows us to test whether there is any differential treatment effect. Since we do not know how teachers matched students we can only measure heterogeneity across the average characteristic of all the possible tutors a tutee might have (e.g., all the Grade 5 students for Preunit tutees). Table 10 show the results from estimating β_2 across different characteristics.[16] The first three columns show heterogeneity by student characteristics, the middle three columns by the average characteristic of all the possible tutors, and the last three columns by school characteristics. Given the large number of hypothesis tested, the table presents adjusted q-values that account for multiple-hypothesis testing following Benjamini and Yekutieli (2001) in square brackets.

There is no evidence of heterogeneity by tutee's age (see Column 1), gender (see Column 2), or how long tutees have been attending Bridge schools (see Column 3).[17] Column 4-6 show that there is no differential effect by tutors' average age, gender, or baseline test score (a PCA index across all subjects), while Column 7-9 show that there is no differential effect by the tutors' pupil-teacher ratio (PTR), tutee-tutor ratio (TTR) or school size (number of enrolled students).

[16] Table 10 provides results for math test scores. Table A.5 provides the results for English test scores.

[17] In this context the age distribution in each grade has wide tails and they often overlap (see Figure A.1 in Appendix A).

Table 10: Heterogeneity: Math test scores

	Tutee characteristics			Tutor characteristics			School characteristics		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Age	Male	Age joined	Age	Male	Score in	PTR	TTR	Enrollment
			Bridge			T3ET15			
Math tutoring × Covariate	0.023*	-0.026	0.023*	0.018	-0.176	-0.024	0.003	0.034	0.000
	(0.013)	(0.029)	(0.012)	(0.019)	(0.218)	(0.032)	(0.004)	(0.029)	(0.000)
	[0.252]	[0.658]	[0.252]	[0.857]	[0.857]	[0.857]	[0.902]	[0.902]	[0.902]
Observations	50820	50934	50820	50538	50538	40891	50934	50913	50934
Adjusted R^2	0.229	0.227	0.229	0.228	0.228	0.239	0.227	0.227	0.227

The outcome variable is the standardized math test score (mean 0 and standard deviation of 1 in English tutoring schools). Each column shows heterogeneity by a different covariate. The covariates in Columns 1-3 are the tutee's age (in 2016), gender, and the age at which they joined Bridge. The covariates used in Columns 4-6 are tutors' average characteristics (age in 2016, gender and test scores at baseline). Columns 7-9 include school level characteristics (pupil teacher ratio (PTR), tutee-tutor ratio (TTR), and number of enrolled students. Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does not include T2ET16 data. Table B.6 provides estimates that includes T2ET16 data. Standard errors, clustered at the school level, are in parentheses. The adjusted q-value taking into account multiple hypothesis testing following Benjamini and Yekutieli (2001) is in square brackets. We create three groups of related hypotheses (Columns 1-3, 4-6, and 7-9) when adjusting for multiple hypothesis testing. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Conclusions

There is an increasing wealth of evidence showing that teaching appropriate to the student's learning level can improve learning outcomes in low-income countries. However, teachers often lack the time (or incentives) to give each child personalized instruction tailored to their needs and providing schools with extra teachers to do so is expensive. Cross-age tutoring, where older students tutor younger students, is a potential alternative to providing personalized instruction to younger students in that it substitutes a trained instructor (the teacher) with an untrained one (the older student). However, it comes at the cost of the older students' time.

We present results from a large randomized control trial (over 180 schools, 15,000 tutees, and 15,000 tutors) in Kenya, in which schools are randomly selected to implement a cross-age tutoring program in either English or math. Our results suggest cross-age tutoring is not a very effective personalized instructional intervention. While tutoring seems to be more effective for math than languages, even for math the treatment effect is modest. However, our results also suggest cross-age tutoring in math helps students in the middle of the ability distribution (but not top-performing students nor those who are far behind). Finally, although the program has modest effect sizes, it is relatively low-cost. As a comparison, contract teachers have been shown to increase student learning by 0.26σ in Kenya (Duflo, Dupas, & Kremer, 2015) and 0.16σ in India (Muralidharan & Sundararaman, 2013). Cross-age tutoring is akin to the contract teacher approach (in which non-professionally trained teachers are hired), as it delegates older kids to teach. Contract teacher have been found to increase test scores by 0.0197σ per USD invested (Kremer, Brannen, & Glennerster, 2013).[18] The total cost of this intervention was 97,000 USD for both the math and the English tutoring program.[19]. While only 187 schools (over 15,000 tutees) participated in the field experiment, 405 schools implemented the program (i.e., over 32,000 students). Thus, the total cost of the program is around 3 USD per student, which translates into test score increases of 0.02σ per USD invested. The cost of implementing the program in future years is projected to decrease as the bulk of the cost was a fix investment: development of lesson guides for tutors. Thus, we expect the program to cost less than 1 USD per student in the future, which translates into test score increases of 0.06σ per USD invested. However, computer-assisted learning programs that personalize instruction may be more cost-effective (Muralidharan et al.,

[18] See https://www.povertyactionlab.org/policy-lessons/education/increasing-test-score-performance for cost-effectiveness comparisons across interventions for details on these calculations.

[19] This includes the cost of the original pilot, the development and testing of lesson guides for tutors, and the monitoring of the program.

2019).

Further research could improve upon the limitation of our study. Specifically, further studies could include a pure control group that allows researchers to study the effect of cross-age tutoring compared to a "business-as-usual" counterfactual. In addition, this would allow them to study directly the possibility that tutoring in one subject has spillovers on other subjects. Finally, studying different "matching" algorithms between tutors and tutees would allow researchers to understand how to optimize these matches.

References

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., ... Walton, M. (2016, October). *Mainstreaming an effective intervention: Evidence from randomized evaluations of "teaching at the right level" in India* (Working Paper No. 22746). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w22746 doi: 10.3386/w22746

Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*(3), 1235-1264.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Bold, T., Kimenyi, M. S., & Sandefur, J. (2013). Public and private provision of education in Kenya. *Journal of African Economies*, *22*(suppl 2), ii39-ii56.

Bruhn, M., & McKenzie, D. (2009, October). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, *1*(4), 200-232.

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*(2), 237-248.

DIVA-GIS. (2016). *Kenya administrative areas.* Retrieved 06/01/2016, from http://biogeo.ucdavis.edu/data/diva/adm/KEN_adm.zip

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739-74.

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

Evans, D. K., & Popova, A. (2016). What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews. *The World Bank Research Observer*, *31*(2), 242-270.

Fafchamps, M., & Mo, D. (2018). Peer effects in computer assisted learning: evidence from a randomized experiment. *Experimental Economics*, *21*(2), 355–382.

Figlio, D. N., & Page, M. E. (2002). School choice and the distributional effects of ability tracking: does separation increase inequality? *Journal of Urban Economics*, *51*(3), 497–514.

Glewwe, P., & Muralidharan, K. (2016). Chapter 10 - improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In S. M. Eric A. Hanushek & L. Woessmann (Eds.), (Vol. 5, p. 653 - 743). Elsevier.

Gray-Lobe, G., Keats, A., Kremer, M., Mbiti, I., & Ozier, O. (2020). *Evaluation of Bridge International Academies in Kenya: Preliminary analysis and plan.* Retrieved from https://doi.org/10.1257/rct.5382-1.0

Jones, S., Schipper, Y., Ruto, S., & Rajani, R. (2014). Can your child read and count? measuring learning outcomes in East Africa. *Journal of African Economies*.

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

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*(3), 1071-1102.

Li, T., Han, L., Zhang, L., & Rozelle, S. (2014). Encouraging classroom peer interactions: Evidence from Chinese migrant schools. *Journal of Public Economics*, *111*, 29-45. Retrieved from http://www.sciencedirect.com/science/article/pii/S0047272713002569 doi: https://doi.org/10.1016/j.jpubeco.2013.12.014

Lucas, A. M., & Mbiti, I. M. (2012, July). Access, sorting, and achievement: The short-run effects of free primary education in Kenya. *American Economic Journal: Applied Economics*, *4*(4), 226-53. Retrieved from https://www.aeaweb.org/articles?id=10.1257/app.4.4.226 doi: 10.1257/app.4.4.226

Mbiti, I. M. (2016, August). The need for accountability in education in developing countries. *Journal of Economic Perspectives*, *30*(3), 109-32.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019, April). Disrupting education? experimental evidence on technology-aided instruction in India. *American Economic Review*, *109*(4), 1426-60. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20171112 doi: 10.1257/aer.20171112

Muralidharan, K., & Sundararaman, V. (2013, September). *Contract teachers: Experimental evidence from India* (Working Paper No. 19440). National Bureau of

Economic Research. Retrieved from http://www.nber.org/papers/w19440 doi: 10.3386/w19440

Pritchett, L., & Beatty, A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. *International Journal of Educational Development*, *40*, 276–288.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.

Secretary for Education, Science and Technology. (2015). *The basic education regulations.*

Shenderovich, Y., Thurston, A., & Miller, S. (2015). Cross-age tutoring in kindergarten and elementary school settings: A systematic review and meta-analysis. *International Journal of Educational Research*.

Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., ... Jimenez, E. (2016). The impact of education programmes on learning and school participation in low-and middle-income countries.

Trudell, B. (2016). *The impact of language policy and practice on children's learning: Evidence from Eastern and Southern Africa*. UNICEF.

Uwezo. (2015). Are our children learning? *Uwezo Kenya Sixth Learning Assessment Report. Nairobi: Twaweza East Africa*.

van der Linden, W. J. (2017). *Handbook of item response theory*. CRC Press.

World Bank. (2015a). *Net primary enrollment - Kenya.* (data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SE.PRM.NENR?locations=KE)

World Bank. (2015b). *School enrollment, primary, private (Kenya.* (data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SE.PRM.PRIV.ZS?locations=KE)

Zimmer, R. (2003). A new twist in the educational tracking debate. *Economics of Education Review*, *22*(3), 307–315.

Online Appendix for “Cross-Age Tutoring: Experimental Evidence from Kenya” by Romero, Chen and Magari

A Additional tables and figures

Table A.1: Test content

	Baby class	Nursery	Preunit	Grd 1	Grd 2	Grd 3	Grd 4	Grd 5	Grd 6	Grd 7
Panel A: Math										
Rote counting	X	X	X	X						
Number identification	X	X	X	X						
Counting	X	X	X	X						
Identify Shapes	X									
Tracing shapes	X									
Tracing numbers	X	X								
Writing numbers		X								
Inequalities		X	X	X	X					
Addition			X	X	X	X	X	X	X	X
Subtraction					X	X	X	X	X	X
Ranks of digits						X	X	X	X	X
Word problems						X	X	X	X	X
Multiplication								X	X	X
Fractions								X	X	X
Division									X	X
Panel B: English										
Picture vocabulary test	X	X								
Tracing letters	X									
Identify words		X	X							
Read letters			X							
Read words			X							
Read sentences			X							
Letter sounds				X	X					
Word sounds				X	X					
Prepositions				X	X					
Reading comprehension				X	X	X	X	X	X	X
Dictation						X	X	X	X	X
Nouns/Verbs						X	X	X	X	X
Adjectives/Adverbs							X	X	X	X
Composition							X	X	X	X

This table displays the skills tested for different grades. While several skills overlap across grades, the test items are grade-appropriate. For example, the complexity of the stories students are asked to read increases across grades. Similarly, students in Grade 1 are asked single-digit addition questions without carrying over, while students in Grade 2 are asked questions with carrying over, and students in Grade 3 are asked two-digit addition questions.

Table A.2: Pupil and tutor test scores during T1DG16

	(1) English Tutoring	(2) Math Tutoring	(3) Difference	(4) Difference (F.E)
Panel A: Tutees				
English	0.000	-0.050	-0.047	-0.077
	(1.000)	(1.061)	(0.082)	(0.078)
Math	0.000	-0.060	-0.056	-0.086
	(1.000)	(1.060)	(0.085)	(0.087)
Swahili	0.000	0.030	0.026	-0.001
	(1.000)	(1.053)	(0.078)	(0.076)
Panel C: Tutors				
English	0.000	0.040	0.041	0.027
	(0.999)	(1.030)	(0.050)	(0.048)
Math	0.000	0.050	0.046	0.030
	(0.999)	(0.999)	(0.050)	(0.044)
Swahili	0.000	-0.010	-0.010	-0.010
	(0.999)	(1.017)	(0.065)	(0.041)

Math, English, and Kiswahili represent standardized test scores (mean zero and standard deviation 1 in English tutoring schools). Each row presents the mean for schools that receive English tutoring (Column 1), schools that receive math tutoring (Column 2), the difference between the two (Column 3), and the difference taking into account the randomization design (i.e., including strata fixed effects) in Column 4. In the first two columns the standard deviation is shown in parentheses, while in the third and fourth columns the standard error of the difference is in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.3: Differential attrition by student characteristics

	Tested in math		Tested in English		Tested in Kiswahili	
	(1) Male	(2) Age	(3) Male	(4) Age	(5) Male	(6) Age
Tested	0.013	-0.016	0.012	-0.018	-0.00033	-0.032
	(0.011)	(0.027)	(0.010)	(0.026)	(0.012)	(0.038)
Math tutoring	0.011	-0.0040	0.011	-0.0064	-0.00089	0.0028
	(0.013)	(0.038)	(0.013)	(0.037)	(0.012)	(0.035)
Tested × Math tutoring	-0.018	-0.024	-0.020	-0.019	0.0034	-0.053
	(0.012)	(0.030)	(0.012)	(0.030)	(0.015)	(0.041)
Mean not tested	0.52	6.39	0.52	6.38	0.51	5.98
N. of obs.	101880	101880	101880	101880	101880	101880
Number of schools	187	187	187	187	187	187

The outcome variable in the odd columns is the students gender (=1 if male, 0 if female) and in the even columns its the age. The first two columns interact an indicator for whether a student is missing data for the math exam with the treatment to look for differential attrition. Column 3-4 interact whether the student is missing data in an English exam, and the last two columns interact whether the student is missing data in the Kiswahili exam. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: Effect on tutees' test math scores: Robustness to different controls

	(1)	(2)	(3)	(4)	(5)
Math tutoring	0.052	0.052	0.055	0.063*	0.064*
	(0.033)	(0.034)	(0.034)	(0.034)	(0.035)
N. of obs.	50934	50820	50820	50424	50424
Number of schools	187	187	187	187	187
Strata fixed effects	Yes	Yes	Yes	Yes	Yes
Baseline scores	Yes	Yes	Yes	Yes	Yes
Student controls	No	Yes	Yes	Yes	Yes
Time and grade fixed effects controls	No	No	Yes	Yes	No
School controls	No	No	No	Yes	Yes

The outcome variable is the math standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation includes T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: Heterogeneity: English test scores

	Tutee characteristics			Tutor characteristics			School characteristics		
	(1) Age	(2) Male	(3) Age joined Bridge	(4) Age	(5) Male	(6) Score in T3ET15	(7) PTR	(8) TTR	(9) Enrollment
Math tutoring × Covariate	0.030**	-0.006	0.022	0.035*	-0.208	0.013	-0.000	0.121***	0.000
	(0.015)	(0.029)	(0.014)	(0.019)	(0.215)	(0.034)	(0.004)	(0.037)	(0.000)
	[0.260]	[1.000]	[0.304]	[0.348]	[0.918]	[1.000]	[1.000]	[0.006]	[1.000]
Observations	48597	48704	48597	48311	48311	39029	48704	48683	48704
Adjusted R^2	0.300	0.300	0.301	0.299	0.299	0.311	0.299	0.302	0.299

The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation data set does not include T2ET16 data. Table B.6 provides estimates that includes T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure A.1: Age distribution across grades

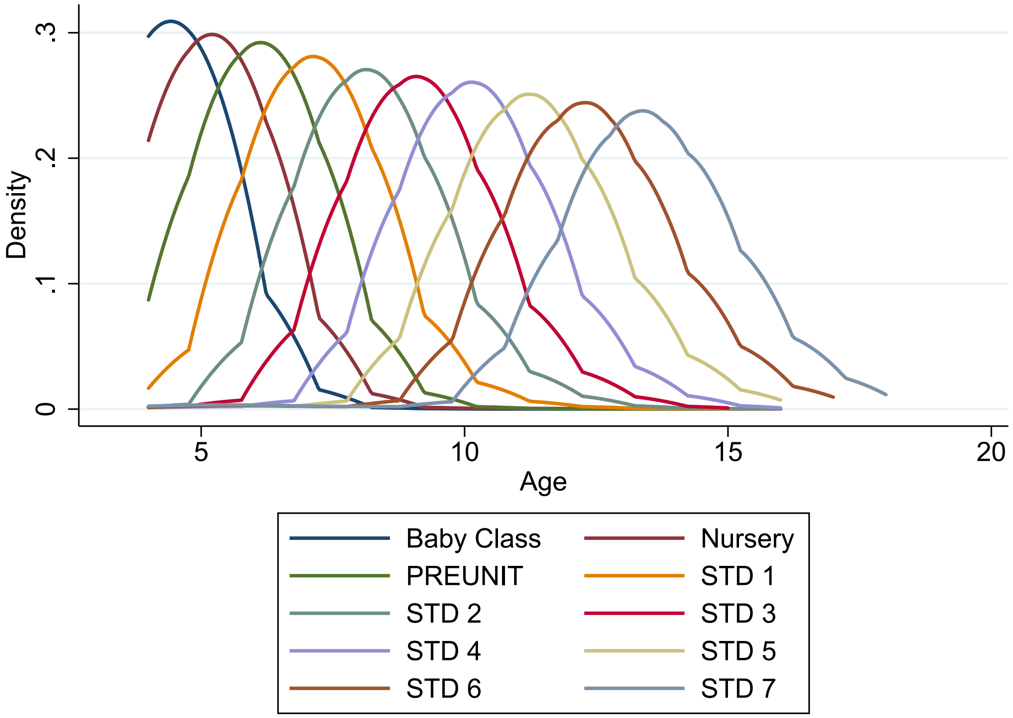
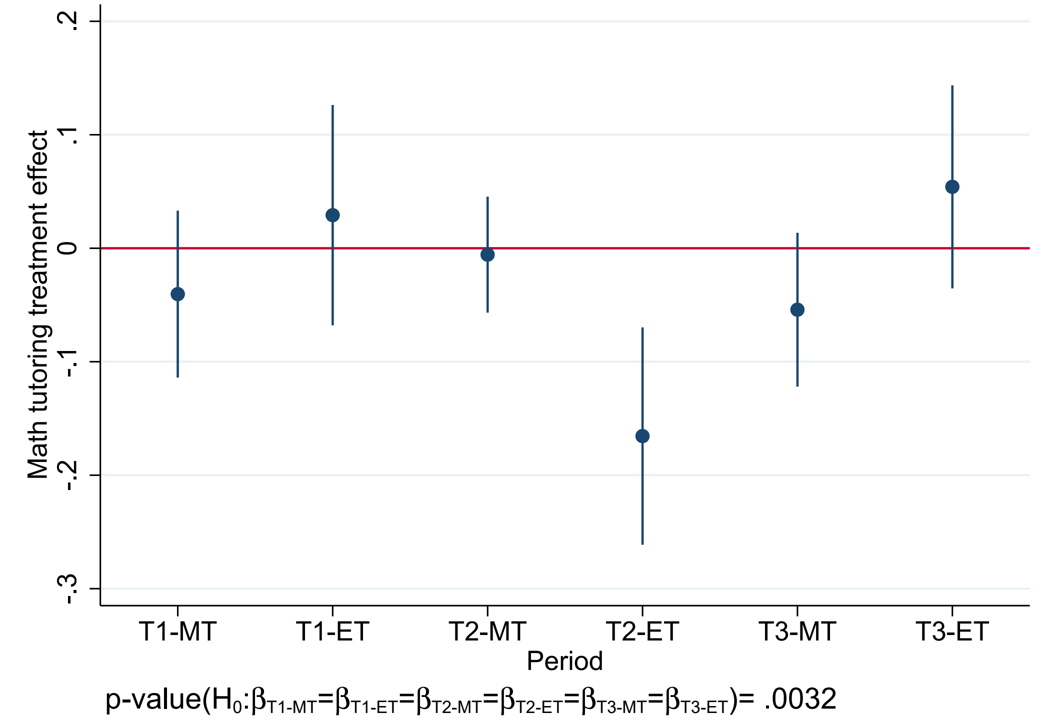
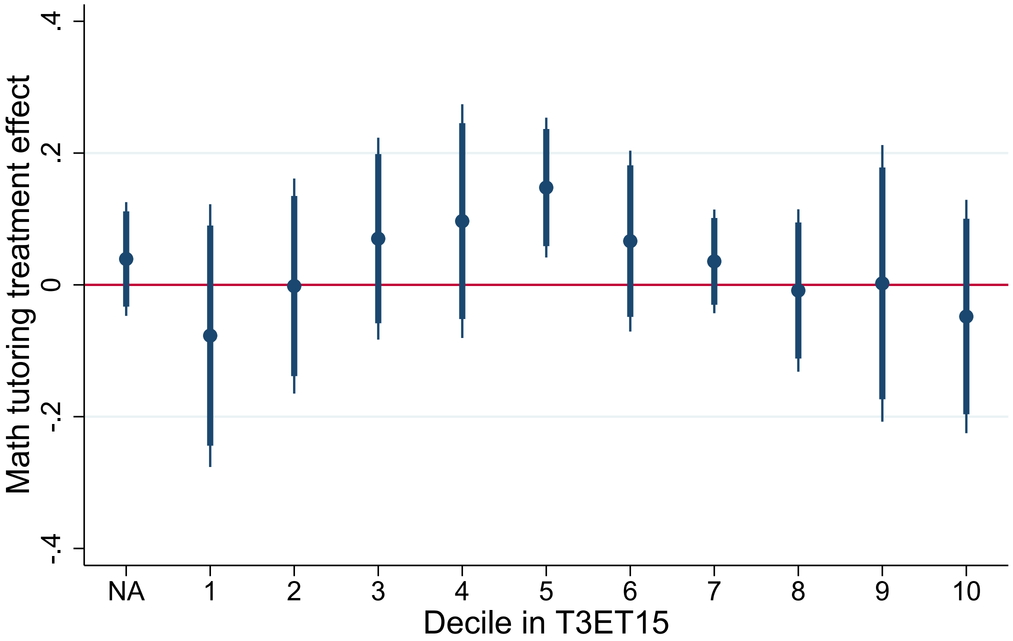


Figure A.2: Difference in testing rates across English and math tutoring in each period



Vertical bars show 95% confidence intervals.

Figure A.3: Treatment effect of math tutoring, relative to English tutoring



p-value(H_0:$Q_1=Q_2=Q_3=Q_4=Q_5=Q_6=Q_7=Q_8=Q_9=Q_{10}$)= .2
p-value(H_0:$Q_1=Q_5=Q_{10}$)= .026

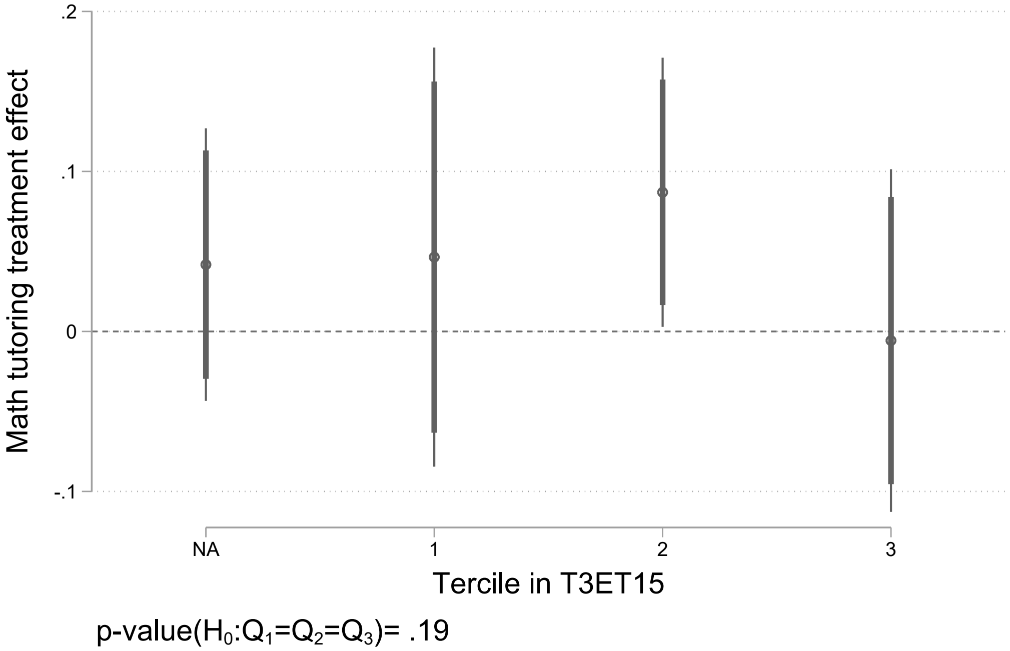
Note: Treatment effect of math tutoring on math test (y-axis) scores by ability decile in T3ET15 (x-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). The figure displays at the bottom the p-value for testing the null hypothesis that the treatment effect is the same in across all deciles, as well as the p-value for testing the null that the treatment effect is the same for the first, the fifth, and the tenth decile. The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for the first decile is .45 (.99), for the second decile is .98 (.99), for the third decile is .37 (.99), for the fourth decile is .28 (.87), for the fifth decile is .0066 (.057), for the sixth decile is .34 (.99), for the seventh decile is .37 (.99), for the eight decile is .89 (.99), for the ninth decile is .98 (.99), and for the tenth decile is .59 (.96).

Table A.6: Heterogeneity by baseline ability of tutors and tutees

	Math test scores		
	(1)	(2)	(3)
Math tutoring	0.038	0.060	0.059
	(0.043)	(0.037)	(0.038)
Tutors' score in T3ET15	0.045	0.034	0.024
	(0.034)	(0.032)	(0.031)
Math tutoring × Tutors' score in T3ET15	-0.056	-0.043	-0.041
	(0.040)	(0.038)	(0.037)
Tutee score in T3ET15	0.505***	0.309***	0.308***
	(0.029)	(0.027)	(0.027)
Math tutoring × Tutee score in T3ET15	-0.025	-0.017	-0.017
	(0.034)	(0.030)	(0.030)
Tutors' score in T3ET15 × Tutee score in T3ET15	-0.054*	-0.051*	-0.053*
	(0.029)	(0.028)	(0.029)
Math tutoring × Tutors' score in T3ET15 × Tutee score in T3ET15	0.063*	0.066*	0.070**
	(0.036)	(0.034)	(0.035)
Observations	22918	22918	22893
Adjusted R^2	0.306	0.366	0.371
Controls	No	Student	Student+ School

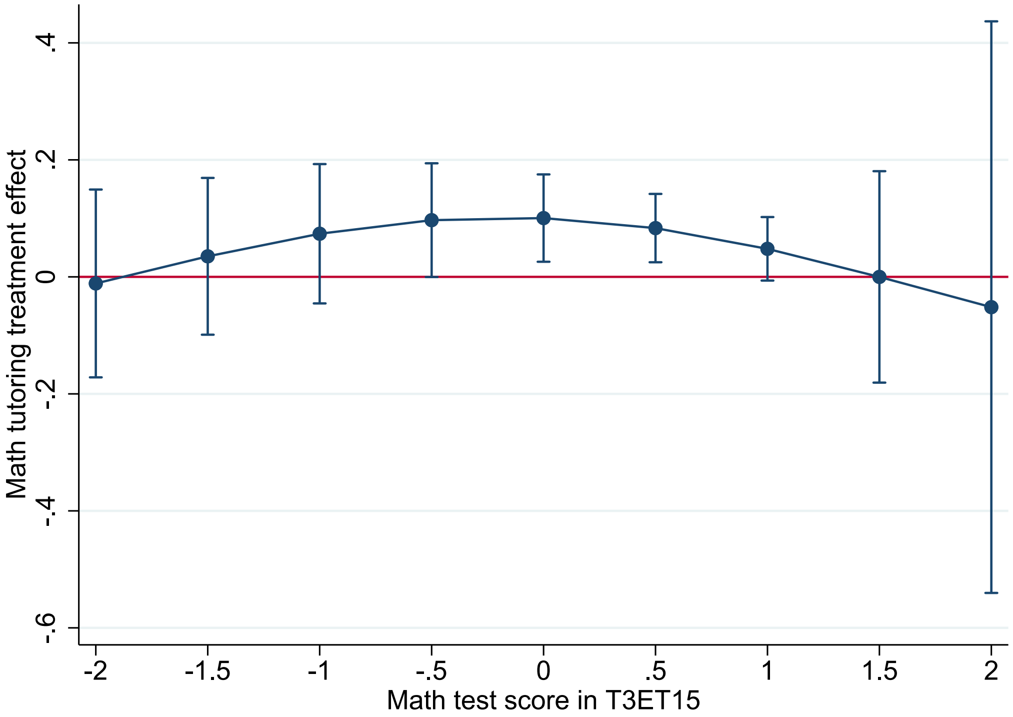
The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). "Tutors' score in T3ET15" is the median score of tutors at baseline. "Tutee score in T3ET15" is the tutee baseline test score in math. Student controls include student's gender and age, and a flexible third-order polynomial controlling for lagged test scores in other subjects. School controls include monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. The estimation data set does not include T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure A.4: Treatment effect of math tutoring, relative to English tutoringby baseline ability tercile



Note: Treatment effect of math tutoring on math test (y-axis) scores by ability terciles in T3ET15 (x-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). The figure displays at the bottom the p-value for testing the null hypothesis that the treatment effect is the same in across all terciles. The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for the first tercile is .49 (.67), for the second tercile is .043 (.02), and for the third tercile is .92 (.94).

Figure A.5: Treatment effect of math tutoring, relative to English tutoring by baseline ability

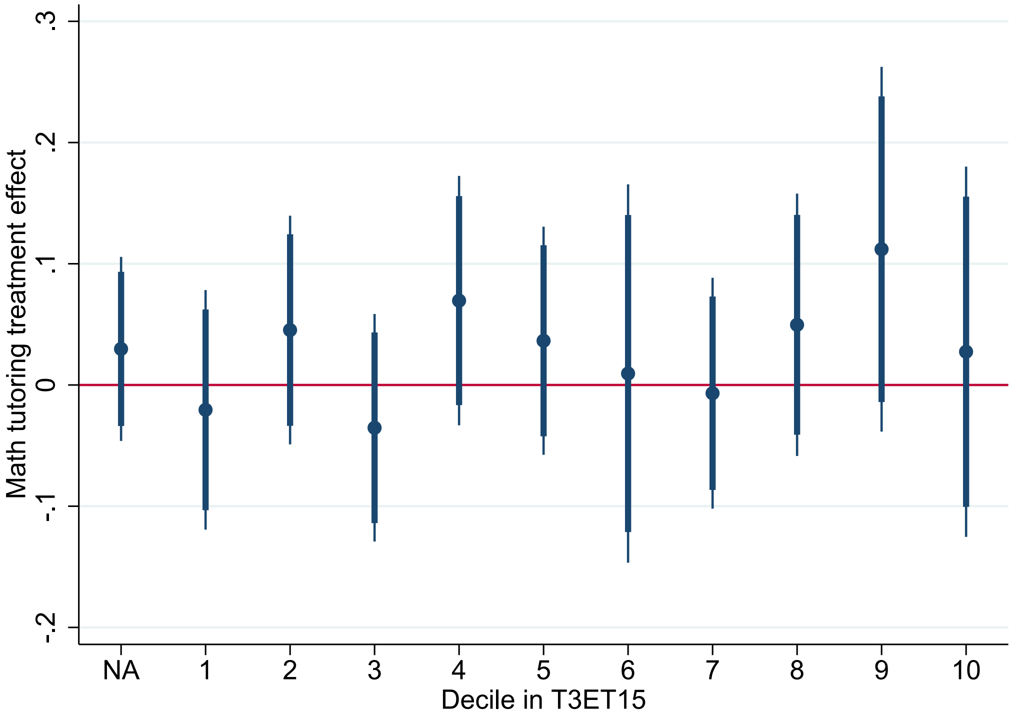


Note: Treatment effect of math tutoring on math test (y-axis) scores by ability in T3ET15 (x-axis). Vertical bars represent 90% confidence interval. The estimating equation is

$$\begin{aligned} Y_{isgd,t} \quad = \quad & \alpha_0 + \beta_0 T_s + \beta_1 T_s \times Y_{isgd,t=0} + \beta_2 T_s \times Y^2_{isgd,t=0} + \beta_3 T_s \times Y^3_{isgd,t=0} + \beta_4 T_s \times Y^4_{isgd,t=0} + \\ & \beta_5 Y_{isgd,t=0} + \beta_6 Y^2_{isgd,t=0} + \beta_7 Y^3_{isgd,t=0} + \beta_8 Y^4_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i + \alpha_3 X_s + \varepsilon_{isd,t}, \end{aligned}$$

The figure shows the difference in outcomes for the average student in math tutoring compared to the average student in English tutoring, conditional on the baseline test score in math.

Figure A.6: Treatment effect of math tutoring on tutors, relative to English tutoring, on math test scores by ability decile in T3ET15



Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively).

B Estimates including T2ET16

A glitch in the software prevented more than 25% of the academies from entering test-score data for T2ET16. Since this is noisy data, we remove it from our sample in the main text, but we provide robustness checks that include the T2ET16 data in this section.

Table B.1: Differential missing data rate between treatment and control students

	Math	English	Swahili
Math tutoring	-0.031	-0.0026	-0.0067
	(0.023)	(0.024)	(0.028)
Mean English	0.61	0.58	0.59
N. of obs.	97742	97756	66149
Number of schools	187	187	187

This table shows the differential missing data rate between students in math tutoring schools compared to students in English tutoring schools. The estimation includes T2ET16 data. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.2: Effect on tutees' test scores

	Math	English	Swahili
Math tutoring	0.057	-0.0038	0.017
	(0.034)	(0.034)	(0.048)
N. of obs.	56834	55937	37835
Number of schools	187	187	186

The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation includes T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.3: Effect on tutees' test scores: Interpolating to input missing outcome data

	Math	English	Swahili
Math tutoring	0.063*	-0.0076	0.021
	(0.035)	(0.033)	(0.047)
N. of obs.	67700	65575	44446
Number of schools	187	187	186

The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation includes T2ET16 data and uses interpolation to input missing outcome data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.4: Effect on tutors' test scores

	Math	English	Swahili
Math tutoring	0.038	-0.0097	-0.014
	(0.031)	(0.035)	(0.036)
N. of obs.	55066	53222	52560
Number of schools	187	187	187

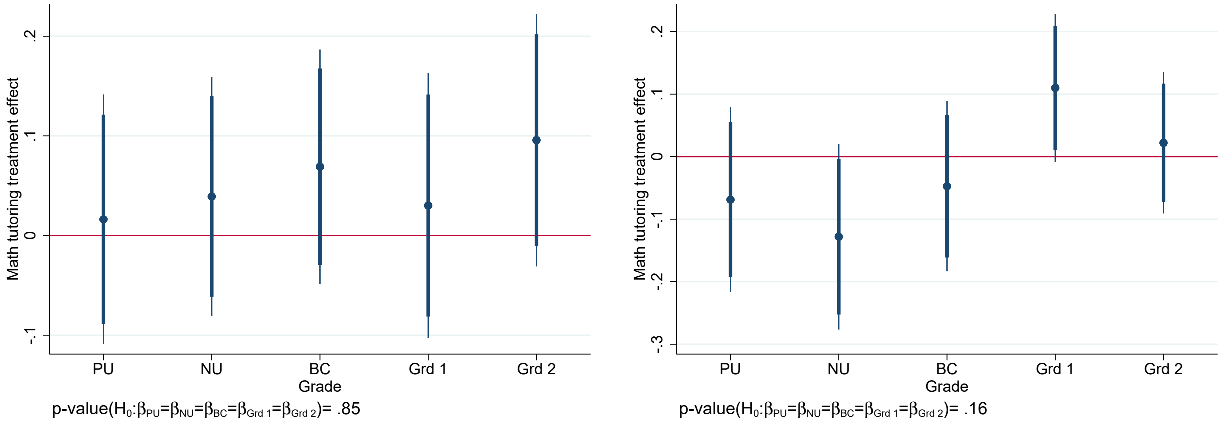
The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation includes T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.5: Effect on tutors' test scores: Interpolating to input missing outcome data

	Math	English	Swahili
Math tutoring	0.042	-0.0072	-0.017
	(0.031)	(0.034)	(0.036)
N. of obs.	65483	63790	63318
Number of schools	187	187	187

The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation includes T2ET16 data and uses interpolation to input missing outcome data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure B.1: Treatment effect of math tutoring, relative to English tutoring, by grade



(a) Math

(b) English

Note: Math (left panel) and English (right panel) test scores (y-axis) by grade (x-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). Each figure displays at the bottom the p-value for testing the null hypothesis that the treatment effect is the same in all grades. The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for math in preunit (PU) is .8 (.91), in nursery (NU) is .52 (.79), in baby class (BC) is .25 (.66), in Grade 1 is .66 (.91), and in Grade 2 is .14 (.56). The raw p-value (and the Romano and Wolf (2005) multiple hypothesis correction adjusted p-value) for English in preunit (PU) is .36 (.56), in nursery (NU) is .091 (.35), in baby class (BC) is .49 (.72), in Grade 1 is .068 (.35), and in Grade 2 is .7 (.72).

Table B.6: Heterogeneity

	Tutee characteristics			Tutor characteristics			School characteristics		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Age	Male	Age joined Bridge	Age	Male	Score in T3ET15	PTR	TTR	Enrollment
Panel A: Math									
Math tutoring × Covariate	0.018	-0.030	0.020	0.013	-0.111	-0.016	0.003	0.040	0.000
	(0.014)	(0.027)	(0.012)	(0.019)	(0.212)	(0.032)	(0.004)	(0.028)	(0.000)
	[0.503]	[0.503]	[0.503]	[1.000]	[1.000]	[1.000]	[0.900]	[0.836]	[0.900]
Observations	57258	57390	57258	56966	56966	46346	57390	57363	57390
Adjusted R^2	0.229	0.228	0.229	0.228	0.228	0.242	0.228	0.228	0.228
Panel B: English									
Math tutoring × Covariate	0.030**	-0.006	0.022	0.034*	-0.199	0.015	0.001	0.127***	0.000
	(0.015)	(0.029)	(0.013)	(0.018)	(0.209)	(0.033)	(0.004)	(0.034)	(0.000)
	[0.234]	[1.000]	[0.294]	[0.310]	[0.939]	[1.000]	[1.000]	[0.001]	[1.000]
Observations	56363	56490	56363	56064	56064	45513	56490	56463	56490
Adjusted R^2	0.300	0.300	0.300	0.299	0.299	0.311	0.299	0.302	0.299

The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include student's gender and age, monthly academy fees, dummies for teachers' wage categories and the pupil-teacher ratio in T1DG16. A flexible third-order polynomial is used to control for lagged test scores. The estimation includes T2ET16 data. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$