Cross-Age Tutoring: Experimental Evidence from Kenya

MAURICIO ROMERO
Instituto Tecnológico Autónomo de México and Abdul Latif Jameel Poverty Action Lab

LISA CHEN and NORIKO MAGARI
Bridge International Academies

I. Introduction

Interventions that 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 Evans and Popova 2016; Glewwe and Muralidharan 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 controlled trial (more than 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

We are grateful to Prashant Bharadwaj, Mitch Downey, Isaac Mbiti, Craig McIntosh, Karthik Muralidharan, Justin Sandefur, and seminar participants at the Pacific Conference for Development Economics and University of California, 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. Mauricio 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. Contact the corresponding author, Mauricio Romero, at mtromero@itam.mx.

Electronically published January 26, 2022

Economic Development and Cultural Change, volume 70, number 3, April 2022.

https://doi.org/10.1086/713940

English or math for 40 minutes. Tutors were five grades above tutees. In some schools, the tutoring focused on math. In others, it focused on English. Whether math or English tutoring took place was randomized across schools. Section II.B 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 of .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 of .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 of .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 .091σ with a confidence of 95% on math test scores. Similarly, we can rule out an effect greater than .087σ with a confidence of 95% on English test scores (for English tutoring relative to math tutoring).

Two central issues to the research design are multitasking 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

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

between knowledge in different subjects. While our research design does not explicitly let us rule out multitasking or spillover effects, we present a series of tests that suggest these are first-order issues in practice. First, tutoring did not take away teaching time from either English or math (or any other subject). Second, had we found effects of English tutoring on English and math tutoring on math, a possible explanation, akin to multitasking, 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 III.A 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 is considerable heterogeneity in learning levels within the same class (Muralidharan, Singh, and Ganimian 2019). Teachers often follow the curriculum, regardless of students' learning levels, making it almost impossible for students lagging behind to catch up (Pritchett and Beatty 2015). Personalized instruction is used to narrow the curriculum gap for students lagging behind. Other interventions aimed at personalized instruction include the use of computer-assisted learning software (e.g., Banerjee et al. 2007; Muralidharan, Singh, and Ganimian 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, Dupas, and Kremer 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

[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.

compared with some of the alternatives to provide personalized instruction. Taking into account that the total cost of the program is around US$3 per student, assuming a linear-dose relationship implies that test scores increase by .02σ per US dollar 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 cross-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, and Kulik 1982). A more recent review looking only at randomized controlled trials comes to the conclusion that cross-age math tutoring has nonsignificant effects on math test scores and that cross-age tutoring in "reading" has a small (statistically significant) positive effect on reading (Shenderovich, Thurston, and Miller 2015). However, only two of the studies reviewed by Shenderovich, Thurston, and Miller (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.

II. Experimental Design

A. 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). 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 et al. 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

[3] For example, Li et al. (2014) found that sitting together high- and low-achieving students in the same class and offering them group incentives for learning improves 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.

quality") of public schools. In response, the demand (and supply) of private primary education increased dramatically (Lucas and 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 that 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, in which schools within its 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 more than 100,000 students.[4]

Bridge tries to take 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]

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 six major exams per academic year. Each academic year has three terms, and each term has a midterm and an end-term exam. Additionally, at the beginning of the academic year, students at the primary level (grades 1–6) take a diagnostic exam. Randomized controlled trials to study the effectiveness of different approaches

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

[5] For example, each Bridge academy has only one employee involved in management. Bridge claims that the vast majority of noninstructional activities the Bridge academy manager would normally have to deal with (billing, payments, expense management, payroll processing, etc.) are automated and centralized. Similarly, Bridge hires experts to develop comprehensive teacher guidelines and training programs, which are then used in all of its schools. Schools charge an average monthly fee of US$6 and cater to families living on US$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 used by teachers across the network. These guides are created by writers in several offices, including Nairobi, Kenya, and Boston. The guides are then streamed to individual teacher tablets. Teachers use tablets to upload students' information (e.g., test scores) to a centralized data warehouse, which can then be accessed by shared services teams.

TABLE 1

TUTORS AND TUTEES

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

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 that in government schools, a school day that is about 2–3 hours longer, and teachers that 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 et al. 2020).

B. *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 p.m.).[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 on 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.

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

[7] At first, the tutoring was designed to be part of the normal school day. However, in 2016, the Kenyan government decreed that class hours end at 3:30 p.m. (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 2
MATH TUTORING INTERVENTION

	Terms 1 and 2	Term 3
Grades 1, 2	Introduction: 3 min Teacher demo: 5 min Tutoring: 30 min Guide with 18 problems One topic	Introduction: 3 min Tutoring 1: 22 min Tutoring 2: 15 min Guide with 60 problems Two topics
Preunit	Introduction: 3 min Warm-up exercise: 10 min Teacher demo: 5 min Tutoring: 15 min Guide with 10 problems One topic	Introduction: 3 min Tutoring 1: 22 min Tutoring 2: 15 min Guide with 56 problems Two topics
Baby class/nursery	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

Note. The math tutoring intervention was scheduled for 3:35–4:15 p.m. for all three terms.

During the first 2 weeks of the 2016 academic year, 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 tutees' 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 2 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 2 weeks, tutors were given guides with problems and activities to do with tutees each day (e.g., addition, counting, and tracing numbers in math and identifying letters, dictation, and reading in English). Roughly,

TABLE 3

ENGLISH TUTORING INTERVENTION

	Term 1	Term 2	Term 3
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
Preunit	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
Baby class, nursery	Introduction and 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 and 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 and 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

Note. The English tutoring intervention was scheduled for 3:35–4:15 p.m. for all three terms.

the tutoring in both English and math had the same structure. First, there is a small introduction (~3 minutes). Afterward, tutors go over the exercises with tutees. Tutors were asked to keep tutees engaged and to 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 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 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

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

the correct solution. Finally, in the first terms, teachers 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 was shifted to reading (writing went from 15 to 10 minutes, while reading went from 9 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 school (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., whether the script translated to classroom practice as envisioned by the master teacher or whether the time allocated for particular tasks was insufficient or too long). 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.

C. *Sampling*

In 2016, Bridge had a network of more than 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 and 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 were schools where other programs were being tested.

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

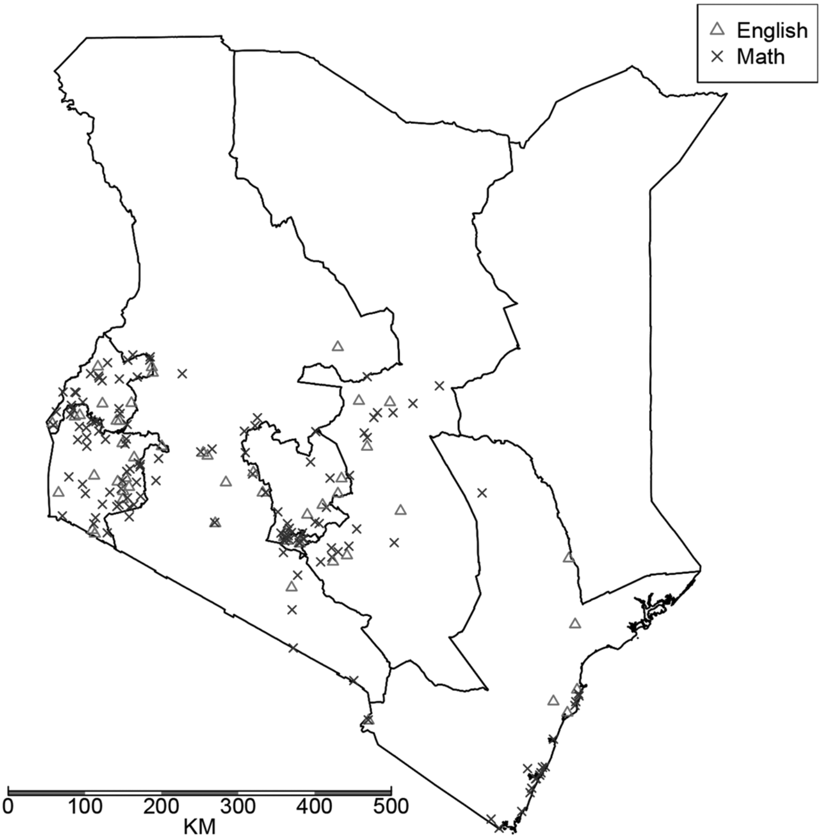

Figure 1. Geographical distribution of schools with math and English tutoring across Kenya. Data on school location were provided by Bridge International Academies. Geographical information from the administrative areas of Kenya comes from DIVA-GIS (2016). A color version of this figure is available online.

D. 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 end-term 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, 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

TABLE 4
LEARNING ASSESSMENTS

Year	Term	Exam	Dates	Code	Grade (2016 Academic Year)
2015	3	End term	November 10–12, 2015	T3ET15	NU, PU, grades 1–6
2016	1	Diagnostic	January 13 and 14, 2016	T1DG16	Grades 1–6
2016	1	Midterm	February 16–18, 2016	T1MT16	BC, NU, PU, grades 1–6
2016	1	End term	April 5–7, 2016	T1ET16	BC, NU, PU, grades 1–6
2016	2	Midterm	June 14–16, 2016	T2MT16	BC, NU, PU, grades 1–6
2016	2	End term	August 9–11, 2016	T2ET16	BC, NU, PU, grades 1–6
2016	3	Midterm	September 26 and 27, 2016	T3MT16	BC, NU, PU, grades 1–6
2016	3	End term	October 25–27, 2016	T3ET16	BC, NU, PU, grades 1–6

Note. NU = nursery; PU = preunit; BC = baby class.

grade the tests and then input the total score into their teacher tablet. The data for students in preunit, nursery, and baby class come 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 (tables A.1–A.6, B.1–B.6 are available in the online appendix) 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 last 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 record only 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 0 and standard deviation of 1 in English tutoring schools) within each grade.

Schools randomly assigned to math tutoring are similar to those assigned to English tutoring. They were inaugurated around the same time (in operation for 2 years by January 1, 2016) and have similar teacher salaries and pupil-teacher ratios of 22 students per teacher (table 5). Tutees (table 6, panels A and B) in English and math tutoring schools are similar across all characteristics. Tutors (table 6, panels C and D) are also similar across English and math

TABLE 5
SCHOOL CHARACTERISTICS IN ENGLISH AND MATH TUTORING SCHOOLS

	English Tutoring (1)	Math Tutoring (2)	Difference (3)	Difference (Fixed Effects) (4)
Days since launch date	672.960	693.310	20.347	−16.257
	(406.417)	(405.017)	(66.887)	(46.838)
Monthly teacher wage of 11,250 KSH	.060	.110	.049	.014
	(.240)	(.313)	(.043)	(.026)
Monthly teacher wage of 10,400 KSH	.180	.100	−.078	−.073
	(.388)	(.304)	(.061)	(.061)
Monthly teacher wage of 7,970 KSH	.760	.790	.028	.058
	(.431)	(.410)	(.070)	(.065)
Teachers	7.440	7.530	.093	.077
	(.541)	(.619)	(.093)	(.092)
Enrollment	167.760	167.180	−.585	−2.554
	(75.793)	(84.627)	(12.894)	(11.451)
Pupil-teacher ratio	22.240	21.980	−.257	−.478
	(9.363)	(10.367)	(1.589)	(1.401)

Note. "Days since launch date" indicates the number of days that have passed since the schools opened on 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" is the number of teachers at the school, and "enrollment" is the enrollment across all grades for the school at the beginning of the school year. Each row presents the mean for schools that receive English tutoring (col. 1), schools that receive math tutoring (col. 2), the difference between the two (col. 3), and the difference taking into account the randomization design (i.e., including strata fixed effects; col. 4). In cols. 1 and 2, the standard deviation is shown in parentheses, while in cols. 3 and 4, the standard error of the difference is in parentheses.

tutoring schools.[11] On average, tutees are 6.5 years old, and tutors are 4.5 years older than their tutees.

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 fails to upload them 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 end-term exam in the second period (T2ET16) is missing more than 60% of test scores for math due to a glitch in the programming update that prevented more than a quarter of the schools from entering test score data. The T2ET16 data-missing rates are different

[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).

[12] In addition, students may be absent from school on the day of the test. However, in most cases, if test score data are 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.

TABLE 6

PUPIL CHARACTERISTICS

	English Tutoring (1)	Math Tutoring (2)	Difference (3)	Difference (Fixed Effects) (4)
	A. Tutees' Time-Invariant Characteristics			
Age	6.600	6.500	−.097*	−.024
	(1.617)	(1.595)	(.054)	(.037)
Male	.520	.520	.002	.000
	(.500)	(.500)	(.011)	(.010)
Age entered Bridge	5.440	5.390	−.057	.013
	(1.669)	(1.643)	(.076)	(.073)
	B. Tutees' Test Scores in T3ET15			
English reading	.000	−.010	−.013	−.058
	(1.000)	(1.021)	(.074)	(.071)
English writing	.000	−.040	−.038	−.064
	(.999)	(1.014)	(.064)	(.058)
Swahili reading	.000	−.020	−.025	−.064
	(1.000)	(1.020)	(.083)	(.082)
Swahili writing	.000	−.070	−.072	−.113
	(1.000)	(1.102)	(.111)	(.090)
Math	.000	.040	.041	.011
	(.999)	(.974)	(.056)	(.052)
	C. Tutors' Time-Invariant Characteristics			
Age	11.040	11.070	.030	.023
	(1.980)	(2.017)	(.097)	(.062)
Male	.500	.520	.020**	.023***
	(.500)	(.500)	(.009)	(.008)
Age entered Bridge	9.660	9.710	.053	.045
	(2.269)	(2.316)	(.140)	(.098)
	E. Tutors' Test Scores in T3ET15			
English reading	.000	.070	.070	.047
	(.999)	(1.038)	(.051)	(.046)
English writing	.000	.070	.069	.034
	(.999)	(.967)	(.054)	(.045)
Swahili reading	.000	.050	.055	.053
	(.999)	(1.042)	(.056)	(.046)
Swahili writing	.000	.140	.138*	.114*
	(.999)	(.941)	(.081)	(.059)
Math	.000	.050	.047	.027
	(.999)	(1.009)	(.063)	(.048)

Note. Math, English, and Kiswahili represent the standardized test scores (mean 0 and standard deviation 1 in English tutoring schools). Each row presents the mean for schools that received English tutoring (col. 1), schools that received math tutoring (col. 2), the difference between the two (col. 3), and the difference taking into account the randomization design (i.e., including strata fixed effects; col. 4). In cols. 1 and 2, the standard deviation is shown in parentheses, while in cols. 3 and 4, 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.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

TABLE 7
NONMISSING DATA

	T1MT16	T1ET16	T2MT16	T2ET16	T3MT16	T3ET16	Total
Math	.751	.591	.711	.399	.570	.532	.590
	(.432)	(.492)	(.453)	(.490)	(.495)	(.499)	(.492)
English writing	.739	.575	.710	.472	.564	.517	.594
	(.439)	(.494)	(.454)	(.499)	(.496)	(.500)	(.491)
English reading	.738	.566	.709	.449	.553	.512	.586
	(.439)	(.496)	(.454)	(.497)	(.497)	(.500)	(.493)
Observations							192,346

Note. Shown are 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.

across math and English tutoring schools (see fig. A.2; figs. A.1–A.6, B.1 are available in the online appendix). However, whether the data are 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 are noisy and have differential attrition across treatments, we remove them from our sample in the main text, but we provide robustness checks that include the data in section B of the online appendix.

Since missing data are prevalent in any given period (more than 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 fig. A.2; tables 8, B.1). Second, there is no evidence of selection bias 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.

TABLE 8
DIFFERENTIAL MISSING DATA RATE BETWEEN TREATMENT AND CONTROL STUDENTS

	Math	English	Swahili
	(1)	(2)	(3)
Math tutoring	−.0027	−.0053	−.0098
	(.022)	(.022)	(.028)
Mean English	.63	.61	.61
Number of observations	81,195	81,209	55,019
Number of schools	187	187	187

Note. Shown is the differential missing data rate between students in math tutoring schools compared with students in English tutoring schools. The estimation data set does not include T2ET16 data. Table B.1 provides estimates that include T2ET16 data. Standard errors, clustered by school, are in parentheses.

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 are missing, we input the value of the average score of T2MT16 and T3MT16 (after standardizing both exams).

III. Results

A. Main Treatment Effects

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

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

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 their 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. Further, X_i is a set of student time-invariant characteristics (month of birth and gender), and X_s indicates school characteristics at baseline (pupil-teacher ratio, monthly school fees, and teachers' wages). We use T_s to indicate whether the student is in a school with a math tutoring program (if not, they are 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 sec. III.B, 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 $\beta_{m,m}$ be the impact of math tutoring on math test scores, $\beta_{e,m}$ the impact of English tutoring on math test scores, $\beta_{m,e}$ the impact of math tutoring on English test scores, and $\beta_{e,e}$ the impact of English tutoring on English test scores. Then $\beta_{\text{math}} = \beta_{m,m} - \beta_{e,m}$ and $\beta_{\text{english}} = \beta_{m,e} - \beta_{e,e}$. 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

TABLE 9
EFFECT ON TEST SCORES

	Tutees			Tutors		
	Math (1)	English (2)	Swahili (3)	Math (4)	English (5)	Swahili (6)
Math tutoring	.063* (.034)	−.0061 (.035)	.035 (.047)	.029 (.031)	−.019 (.035)	−.020 (.036)
Number of observations	50,424	48,204	32,736	48,741	46,938	46,512
Number of schools	187	187	186	187	187	187

Note. The outcome variable is the standardized test score (mean 0 and standard deviation of 1 in English tutoring schools). Student and school controls include students' gender and age, monthly academy fees, dummies for teachers' wage categories, and the pupil-teacher ratio in T1DG16. The number of observations in col. 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.
* $p < .10$.

practicing math in their tutoring sessions. Thus, β_{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, $\beta_{\text{math}} < \beta_{m,m}$. 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 negative of the effect on English scores ($-\beta_{\text{english}}$) is a good estimate for the treatment effect of English tutoring on English ($\beta_{e,e}$).

1. Tutees

Math tutoring relative to English tutoring has a small positive effect on math test scores of .063σ (see col. 1 of 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 col. 2). The difference between the treatment effect of math tutoring on math and English tutoring on English (.069) is statistically significant (p-value of .0024). Math tutoring relative to English tutoring seems to have no effect on Kiswahili (see col. 3). These results

[13] As mentioned in sec. II.B, 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.

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 sec. II.D for details on how the interpolation is done and table B.3 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.

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 .091σ with a confidence of 95% on math test scores. Similarly, we can rule out an effect greater than .087σ with a confidence of 95% on English test scores (for English tutoring relative to math tutoring). See columns 4 and 5 of table 9 for details.

B. 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).

1. Periods

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

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

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.

2. Grade

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

$$\begin{aligned} Y_{isgd,t} = {} & \alpha_0 + \sum_{g=1}^{5} \beta_g T_s \times 1_{\text{grade}=g} + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i \\ & + \alpha_3 X_s + \varepsilon_{isd,t}, \end{aligned} \tag{3}$$

where β_1 measures the treatment effect of math tutoring relative to English tutoring for baby class, β_2 for nursery, β_3 for preunit, 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.

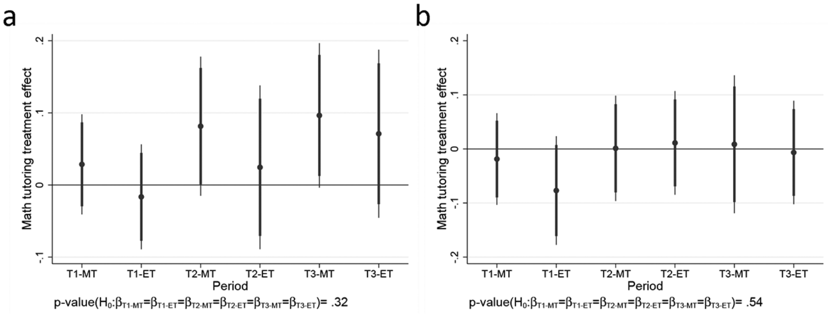

Figure 2. Evolution of the treatment effect of math tutoring relative to English tutoring. Math (*a*) and English (*b*) test scores (*y*-axis) by period (*x*-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). At the bottom of each panel is shown the *p*-value for testing the null hypothesis that the treatment effect is the same in all periods. The raw *p*-value (and the multiple-hypothesis correction-adjusted *p*-value of Romano and Wolf 2005) 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 multiple-hypothesis correction-adjusted *p*-value of Romano and Wolf 2005) 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). A color version of this figure is available online.

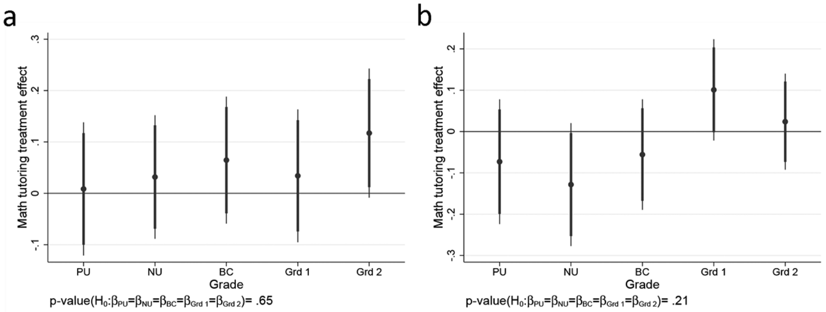

Figure 3. Treatment effect of math tutoring relative to English tutoring by grade. Math (*a*) and English (*b*) test scores (*y*-axis) by grade (*x*-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). At the bottom of each panel is shown the *p*-value for testing the null hypothesis that the treatment effect is the same in all grades. The raw *p*-value (and the multiple-hypothesis correction-adjusted *p*-value of Romano and Wolf 2005) 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 multiple-hypothesis correction-adjusted *p*-value of Romano and Wolf 2005) for English in PU is .34 (.57), in NU is .09 (.37), in 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. A color version of this figure is available online.

3. Baseline Test Scores

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

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

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

Figure A.3 shows the estimates for all the βs that 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 with the average effect of .063σ).[14]

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

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

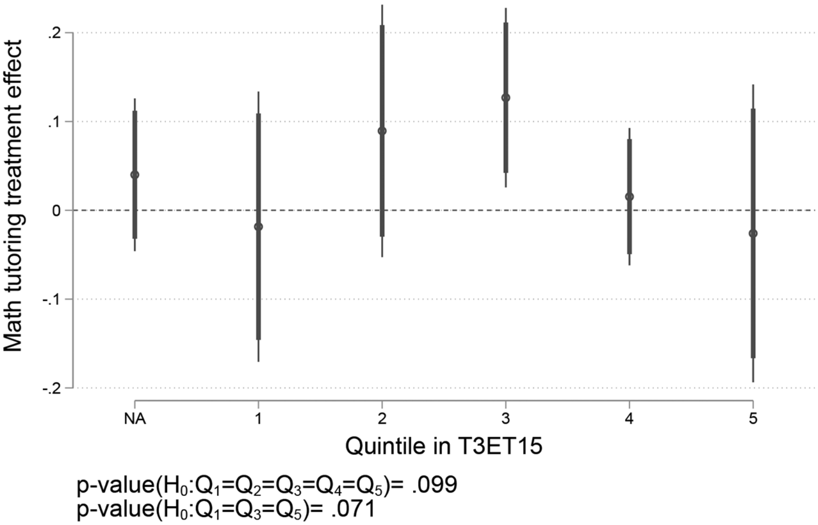

Figure 4. Treatment effect of math tutoring relative to English tutoring by baseline ability quintile. Treatment effect of math tutoring on math test (*y*-axis) scores by ability quintile in T3ET15 (*x*-axis). Vertical bars represent 90% and 95% confidence intervals (thick and thin lines, respectively). At the bottom of the graph is displayed the *p*-value for testing the null hypothesis that the treatment effect is the same across all quintiles, as well as the *p*-value testing whether the treatment effect for the first, third, and fifth quintiles is the same. The raw *p*-value (and the multiple-hypothesis correction-adjusted *p*-value of Romano and Wolf 2005) 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).

That students in the middle of the distribution benefit the most is robust to using deciles (see fig. 4) and terciles (see fig. A.4), as well as to interacting the treatment dummy with a fourth-order polynomial of the baseline test score (see fig. 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 fig. A.6). The effect is indistinguishable from zero for all tutors, regardless of 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.

TABLE 10
HETEROGENEITY: MATH TEST SCORES

	Tutee Characteristics			Tutor Characteristics			School Characteristics		
	Age (1)	Male (2)	Age Joined Bridge (3)	Age (4)	Male (5)	Score in T3ET15 (6)	Pupil-Teacher Ratio (7)	Tutee-Tutor Ratio (8)	Enrollment (9)
Math tutoring × covariate	.023*	−.026	.023*	.018	−.176	−.024	.003	.034	.000
	(.013)	(.029)	(.012)	(.019)	(.218)	(.032)	(.004)	(.029)	(.000)
	[.252]	[.658]	[.252]	[.857]	[.857]	[.857]	[.902]	[.902]	[.902]
Observations	50,820	50,934	50,820	50,538	50,538	40,891	50,934	50,913	50,934
Adjusted R^2	.229	.227	.229	.228	.228	.239	.227	.227	.227

Note. 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 cols. 1–3 are the tutee's age (in 2016), gender, and the age at which they joined Bridge. The covariates used in cols. 4–6 are tutors' average characteristics (age in 2016, gender, and test scores at baseline). Columns 7–9 include school-level characteristics (parent-teacher ratio, tutor-tutee ratio, 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 include 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 brackets. We create three groups of related hypotheses (cols. 1–3, 4–6, and 7–9, respectively) when adjusting for multiple-hypothesis testing.
* $p < .10$.

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:

$$\begin{aligned} Y_{isgd,t} = {} & \alpha_0 + \beta_1 T_s + \beta_2 T_s \times c_i + \alpha_1 Y_{isgd,t=0} + \gamma_g + \gamma_t + \gamma_d + \alpha_2 X_i \\ & + \alpha_3 X_s + \varepsilon_{isd,t}, \end{aligned} \tag{5}$$

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 measure only 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 shows the results from estimating β_2 across different characteristics.[16] Columns 1–3 show heterogeneity by student characteristics, columns 4–6 by the average characteristic of all the possible tutors, and columns 7–9 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 brackets.

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

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

IV. Conclusions

There is an increasing wealth of evidence showing that teaching appropriate to a 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, while 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 controlled trial (more than 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 or 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, and Kremer 2015) and 0.16σ in India (Muralidharan and Sundararaman 2013). Cross-age tutoring is akin to the contract teacher approach (in which nonprofessionally trained teachers are hired), as it delegates older kids to teach. Contract teacher have been found to increase test scores by .0197σ per US dollar invested (Kremer, Brannen, and Glennerster 2013).[18] The total cost of this intervention was US$97,000 for both the math and the English tutoring program.[19] While only

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

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

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

187 schools (more than 15,000 tutees) participated in the field experiment, 405 schools implemented the program (i.e., more than 32,000 students). Thus, the total cost of the program is around US$3 per student, which translates into test score increases of .02σ per US dollar invested. The cost of implementing the program in future years is projected to decrease as the bulk of the cost was a fixed investment: development of lesson guides for tutors. Thus, we expect the program to cost less than US$1 per student in the future, which translates into test score increases of .06σ per US dollar invested. However, computer-assisted learning programs that personalize instruction may be more cost-effective (Muralidharan, Singh, and Ganimian 2019).

Further research could improve upon the limitations of our study. Specifically, further studies could include a pure control group that allows researchers to study the effect of cross-age tutoring compared with a "business-as-usual" counterfactual. In addition, this would allow for directly studying 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., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India." NBER Working Paper no. 22746 (October), National Bureau of Economic Research, Cambridge, MA.

Banerjee, A., S. Cole, E. Duflo, and L. Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122, no. 3:1235–64.

Benjamini, Y., and D. Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics* 29:1165–88.

Bold, T., M. S. Kimenyi, and J. Sandefur. 2013. "Public and Private Provision of Education in Kenya." *Journal of African Economies* 22 (suppl. 2):ii39–ii56.

Bruhn, M., and D. McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1, no. 4:200–232.

Cohen, P. A., J. A. Kulik, and C.-L. C. Kulik. 1982. "Educational Outcomes of Tutoring: A Meta-Analysis of Findings." *American Educational Research Journal* 19, no. 2:237–48.

DIVA-GIS. 2016. Kenya administrative areas. https://biogeo.ucdavis.edu/data/diva/adm/KEN_adm.zip.

Duflo, E., P. Dupas, and M. Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101, no. 5:1739–74.

———. 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., and A. Popova. 2016. “What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews.” *World Bank Research Observer* 31, no. 2:242–70.

Fafchamps, M., and D. Mo. 2018. “Peer Effects in Computer Assisted Learning: Evidence from a Randomized Experiment.” *Experimental Economics* 21, no. 2:355–82.

Figlio, D. N., and M. E. Page. 2002. “School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?” *Journal of Urban Economics* 51, no. 3:497–514.

Glewwe, P., and K. Muralidharan. 2016. “Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications.” In *Handbook of the Economics of Education*, Vol. 5, eds. Eric A. Hanushek, S. Machin, and L. Woessmann, 653–743. Amerstdam: Elsevier.

Gray-Lobe, G., A. Keats, M. Kremer, I. Mbiti, and O. Ozier. 2020. “Evaluation of Bridge International Academies in Kenya.” AEA RCT Registry, May 20. https://doi.org/10.1257/rct.5382-1.7000000000000002.

Jones, S., Y. Schipper, S. Ruto, and R. Rajani. 2014. “Can Your Child Read and Count? Measuring Learning Outcomes in East Africa.” *Journal of African Economies* 23, no. 5:643–72.

Kremer, M., C. Brannen, and R. Glennerster. 2013. “The Challenge of Education and Learning in the Developing World.” *Science* 340, no. 6130:297–300.

Lee, D. S. 2009. “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” *Review of Economic Studies* 76, no. 3:1071–102.

Li, T., L. Han, L. Zhang, and S. Rozelle. 2014. “Encouraging Classroom Peer Interactions: Evidence from Chinese Migrant Schools.” *Journal of Public Economics* 111:29–45.

Lucas, A. M., and I. M. Mbiti. 2012. “Access, Sorting, and Achievement: The Short-Run Effects of Free Primary Education in Kenya.” *American Economic Journal: Applied Economics* 4, no. 4:226–53.

Mbiti, I. M. 2016. “The Need for Accountability in Education in Developing Countries.” *Journal of Economic Perspectives* 30, no. 3:109–32.

Muralidharan, K., A. Singh, and A. J. Ganimian. 2019. “Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India.” *American Economic Review* 109, no. 4:1426–60.

Muralidharan, K., and V. Sundararaman. 2013. “Contract Teachers: Experimental Evidence from India.” NBER Working Paper no. 19440. National Bureau of Economic Research, Cambridge, MA.

Pritchett, L., and A. Beatty. 2015. “Slow Down, You’re Going Too Fast: Matching Curricula to Student Skill Levels.” *International Journal of Educational Development* 40:276–88.

Romano, J. P., and M. Wolf. 2005. “Stepwise Multiple Testing as Formalized Data Snooping.” *Econometrica* 73, no. 4:1237–82.

Secretary for Education, Science and Technology. 2015. “The Basic Education Regulations.” *Kenya Gazette*, suppl. 37. http://kenyalaw.org/kl/fileadmin/pdfdownloads/LegalNotices/39-BasicEducationRegulations_2015.pdf.

Shenderovich, Y., A. Thurston, and S. Miller. 2015. “Cross-Age Tutoring in Kindergarten and Elementary School Settings: A Systematic Review and Meta-Analysis.” *International Journal of Educational Research* 76:190–210.

Snilstveit, B., J. Stevenson, R. Menon, D. Phillips, E. Gallagher, M. Geleen, H. Jobse, T. Schmidt, and E. Jimenez. 2016. *The Impact of Education Programmes on Learning and School Participation in Low- and Middle-Income Countries: Systematic Review Summary 7*. London: 3ie.

Trudell, B. 2016. *The Impact of Language Policy and Practice on Children’s Learning: Evidence from Eastern and Southern Africa*. New York: 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*. Boca Raton, FL: CRC.

World Bank. 2015a. Net primary enrollment—Kenya. World Development Indicators, https://data.worldbank.org/indicator/SE.PRM.NENR?locations=KE.

———. 2015b. School enrollment, primary, private—Kenya. 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, no. 3:307–15.