PRELIMINARY DRAFT

Improving Primary School Quality Across Countries: Experimental Evidence from Kenya and Uganda[1]

Adrienne M. Lucas
University of Delaware

Patrick J. McEwan
Wellesley College

Moses Ngware
African Population and Health Research Center

Moses Oketch
Institute of Education, University of London

October 2012

Abstract: Primary school enrollments have increased rapidly in sub-Saharan Africa, spurring new concerns about low levels of quality. This study uses a field experiment to assess whether the Reading to Learn (RtL) program improved student achievement in Kenya and Uganda. In each country, RtL provided teacher training and school materials to randomly selected primary schools. Using common tests and surveys across countries, we find that the program successfully increased the teaching and learning materials in treatment schools. In Uganda, achievement in written and oral literacy increased by nearly 20 percent of a standard deviation. In Kenya, we find a smaller effect (8 percent), just for oral literacy. Across both countries, schools whose implementation of the program was closer to the RtL ideal had better achievement effects. Among those treated, the adherence to the ideal RtL model was endogenous to baseline characteristics of student ability and teacher absenteeism.

Keywords: education, impact evaluation, teacher training, student resources, Kenya, Uganda, randomized controlled trial
JEL Codes: I2, O15, H52

[1] The authors gratefully acknowledge financial support from The William and Flora Hewlett Foundation's Quality Education in Developing Countries (QEDC) program. We thank Maurice Mutisya for his exceptional research assistance. For useful comments, we thank Kristin Butcher, Ellen Green, Saul Hoffman, Paul Larson, Isaac Mbiti, Lynn Murphy, Chloe O'Gara, Maria Perez, Dana Schmidt, and seminar participants at IFPRI, Middlebury College, Swarthmore College, University of Delaware, and the Improving Learning Outcomes in Developing Countries workshop at Wellesley College.

1. Introduction

In the past twenty years, sub-Saharan African governments and other international organizations have invested heavily in increasing the number of students enrolled in primary school. With increased enrollment rates and years of completed schooling, the focus has now shifted to ensuring that students are learning while they are enrolled. Such efforts are severely hampered by a lack of knowledge about the investments that might increase learning among primary school children. This paper uses a randomized controlled trial to evaluate the impact of the popular Reading to Learn program, which provided staff training and resource provision, on student absenteeism and achievement.

Both Kenya and Uganda, the two study countries, have successfully eliminated primary school fees in government schools, but primary school exit examinations and independently collected data show that some students are learning very little while in school.[2] This study randomly assigned schools in two low-achieving districts in each country to participate in a lower primary literacy intervention. The Reading to Learn (RtL) treatment was designed to improve literacy levels in lower primary grades through the provision of teacher training and materials. Lower primary teachers in treatment schools were provided with centralized teacher training on the RtL scaffolding model of instruction and lower primary classrooms were provided with instructional materials, primarily books and stationery supplies to create visual aids, to support this model. Exams were administered to lower primary students in both the treatment and control group prior to the intervention and then approximately 1.5 years after the start of the intervention. Based on a value-added model, we find that the intervention increased oral and written literacy in Uganda by 0.18-0.2 standard deviations. In contrast, there are smaller

[2] In Kenya in 2011, only 30 percent of grade 3 students could proficiently read a grade 2 story in Swahili (Uwezo 2011). In 2010 in Uganda, 3 percent of grade 3 students could read a grade 2 story in English (Uwezo 2010).

effects on oral literacy in Kenya (0.08 standard deviations), and none on other assessments. Additionally, we find higher achievement gains, in both countries, for students in schools that more closely adhered to the "ideal" RtL model. This might simply reflect that schools with better implementation also had higher baseline scores and fewer problems with teacher absenteeism. However, the implementation results are robust to controls for these and a wide range of other baseline school and student covariates.

2. The Reading to Learn Treatment in Kenya and Uganda

Primary school in Kenya has 8 grades, beginning at age 6 or later. In Uganda, students may begin a 7 grade primary education at age 5. Grades 1-3—the focus of this study—are considered "lower primary" in both countries. Though most schools offer all primary grades, some offer only the lower primary grades. In both countries the school year starts in January. Both countries removed school fees in all government primary schools, Uganda in 1997 and Kenya in 2003. These policies have led to remarkable growth in primary school enrolment and graduation (Oketch and Rollestone 2007; Lucas and Mbiti 2012). Now that access has been increased, concerns are emerging about the quality of education being provided.

The Reading to Learn (RtL) program was first implemented in Australia approximately ten years ago, with the goal of increasing the reading level of those well behind grade level, especially the Aboriginal population. Since its start, it has been expanded to Afghanistan, Kenya, Uganda, and Sweden. The model was designed to support students whose mother tongues have a strong oral tradition but who have very limited prior exposure to written language, aptly describing the children in the study districts in Kenya and Uganda. The RtL design in Kenya and Uganda includes three components: teacher preparedness and practice, school leadership, and

classroom learning environments. Early-grade teachers were centrally trained within each country on a child-centered, systematic, and social-interaction focused instructional approach that emphasized the use of local materials. The intervention applies a 5 step "scaffolding" approach: (1) preparation before reading when the teacher tells the pupils the story, (2) reading the story, (3) sentence making, (4) spelling and phonics, and (5) writing. The approach is a balanced approach to the teaching of reading, with elements that include look-say, whole language, and phonics methodology. Prior education literature has emphasized the inclusion of such elements as important for literacy attainment (Dubeck et al. 2012). In addition to training, schools were provided with teaching and learning materials. Finally, project technical staff worked with head teachers, key classroom teachers, and district education staff on methods to support the program.[3] The Aga Khan Foundation (AKF) implemented the program.

The instructional model and resources were based on the same RtL model in both countries, although each country's implementation was unique in some ways. In Kenya the RtL materials were in Swahili and English, and they supplemented other Ministry of Education materials in the same languages.[4] According to the official curriculum for primary school students, the language of instruction during the first three years of primary school should be the mother tongue of the learners. In the study schools in Kenya, multiple mother tongues were often in the same school and none of them have an established lexicography. Therefore, Swahili was the official language of the first three grades. In practice, most students, parents, and teachers preferred English as the language of instruction and in almost all schools Swahili was taught as a foreign language. As

[3] The model described is the "core" model. All treatment schools received at least this intervention. In half of the treatment schools in Kinango in Kenya and all of the schools in Amolatar in Uganda treatment consistent of a "core plus" model that included all of the elements of the core model plus a parental involvement component that established mini-libraries in each community and encouraged parents to borrow books and read and tell stories to their children. Because of the randomization scheme we are not able to separately identify the effects of the two treatments.

[4] In the baseline survey, 73% of grade 2 classrooms had visible textbooks.

implemented, only thirty minutes each day were set aside each day for RtL in Swahili and this was often final subject of the day. This allocation was inadequate to cover the entire 5 step RtL methodology in a single setting. Therefore, a single cohesive lesson was often spread across multiple days. Teachers were also encouraged and supported in implementing RtL in their English lessons. Unfortunately, due to the official ministry of education policy emphasizing Swahili as the language of instruction, students were not tested in English.

In Uganda the literacy materials were in Lango, the primary local language spoken in the study districts and the language of instruction in the study schools. For many of the study schools in Uganda, these were the only local language texts. Existing books in the schools were limited in number and often in Swahili, a second or third language for many students. Additionally, existing materials were often not tailored for the thematic curriculum that Uganda adopted in 2006. The curriculum contains central themes around which daily lessons are to be based, regardless of the subject. Therefore, instruction in English, Lango, or even mathematics might involve the use of similar words. Lango lessons occurred during a 60 minute block of time, usually as one of the first two lessons of the day after announcements and current events, allowing all 5 steps of RtL to occur in a single lesson block.

Those schools not selected for treatment continued to follow the government-prescribed curriculum that mandated what was to be taught, but not the methods of instruction. Based on visits to lower primary classrooms in the one of the study districts in Kenya, Dubeck et al. (2012) found in practice an emphasis on word recognition, oral language, and choral repetition where students would repeat sentences but not read them. Teachers were not comfortable teaching or using phonics-based instruction.

3. Empirical Strategy

3.1. Random Assignment

Two districts in each country were chosen because of their poor historical performance on the national primary school exit examinations and high poverty rates. Kwale and Kinango were the two districts selected for the intervention in Kenya. The districts are adjacent to each other in the Coast Province of Kenya. In 2007 the districts had the lowest average scores in the country on the primary school exit examinations. Kinango has among the highest levels of poverty in the country with half of the population living below the poverty line and two thirds considered food poor (KNBS 2009). One of the education challenges in the districts is a lack of parental involvement in their children's education due to a low valuation of education and the misconception that parents have no responsibilities for their children's education.

As with the districts in Kenya, Amolatar and Dokolo districts were selected in Uganda because of their historically poor educational performance. These two adjacent districts are in the northern region of Uganda that suffered displacement, murder, abductions, and torture during the twenty year Lord's Resistance Army (LRA) insurgency. With the elimination of the LRA in the region, the government and a number of international partners have sought to provide effective educational opportunities. Despite these efforts, literacy rates remain low in northern Uganda, and literacy rates in these districts were well below the national average.

Within each district in both Kenya and Uganda, schools are divided into geographically-proximate clusters and monitoring and support are provided by common education officials to all schools in a given cluster. In order to encourage official support and minimize contamination, randomization occurred at the cluster level. In Kenya, the 28 clusters contained 1 to 8 schools

each, for a total of 112 schools (see Figure 1). The Kenyan clusters belonged to one of three strata: (1) clusters located in Kwale; (2) clusters located in Kinango, where schools participated in the Kenya School Improvement Project (KENSIP) intervention; and (3) non-KENSIP clusters in Kinango.[5] KENSIP is a separate school-improvement program implemented by AKF and funded by the United States Agency for International Development (USAID). That program started in Kinango in 2000 and its effects have not been empirically tested. In Uganda, the 10 clusters were administrative sub-counties that contained 2 to 16 schools each, for a total of 109 schools. Uganda's clusters were sub-regions of two strata, the districts of Amolatar and Dokolo. The randomized assignment to treatment or control groups occurred at the cluster level, within each of the five strata. In Kenya, 12 and 16 clusters were randomly assigned, respectively, to treatment and controls groups. In Uganda, 4 and 6 clusters were randomly assigned to treatment and control groups (see Figure 1A and 1B).

All schools located within treatment clusters should have received the treatment, while control schools should have been studied but not received treatment. However, one school in Amolatar and one in Dokolo were randomly assigned to a control cluster, but were later re-assigned to the treatment by the implementing agency. We assign these two schools to their initial condition—the control group—in subsequent analyses. Our treatment effects can therefore be interpreted as intention-to-treat effects. Overall, the small amount of crossover suggests that this is a minor issue in the interpretation of effects.[6]

[5] In Kenya not all clusters in the districts of Kwale and Kinango were in the experimental sample. In Kwale, three clusters were omitted because at the time of randomization they were scheduled to be included in another study (conducted by the HALI project) and an entire administrative zone was additionally omitted. In Kinango, a number of clusters were excluded ex ante from randomization.

[6] In the following regression: $treated = \beta_0 + \beta_1 itt + \varepsilon$ where *treated* is whether the school was treated and *itt* is intention to treat (i.e. original randomization) the estimate of the coefficient on *itt* is 0.982 with a standard error of 0.013 and a regression R-squared of 0.96.

In practice, the implementation of the program began in October of 2009 with teachers receiving three days of in-service training led by separate teams in Kenya and Uganda. After this training teachers received basic stationery (e.g. manila paper, newsprint, markers, glue, etc.) to be used to make their own instructional materials. At about this time school management committees and head teachers were trained in the RtL approach, how to development a School Development Plan, and how to support teachers through leadership, mentoring, and supervision. Due to unforeseen delays, most schools did not receive the classroom mini-libraries with books in English and the local language (Swahili in Kenya and Lango in Uganda) and the lockable storage units in which to store these books until April 2010.

As the intervention continued, AKF-trained teams visited each treatment school monthly to provide in-class mentoring support. Teachers were invited to quarterly review trainings to meet with peers and members of the implementation team to share ideas, see model lessons, and receive refresher trainings. At the start of the new school year in 2010 and 2011, teachers assigned to the lower primary classrooms who had not previously received training were trained locally. Figure 2 provides a schematic of the program timeline and the treated cohorts.

3.2. Estimation

The usual difficulty in estimating the effect of a school treatment on student outcomes is the likely correlation between unobservable attributes of students that affect achievement—such as motivation or ability—and the treatment status of the school. In contrast, the random assignment of the school treatment ensures that students are, on average, similar across treated and untreated schools. Given random assignment, estimation is straightforward. Our main regression specification in each country is:

$$posttest_{isj} = \alpha + \beta itt_{sj} + \sum_{e=1}^{3} \gamma_e pretest_{e,isj} + \gamma X_{isj} + \delta_j + \varepsilon_{isj} \quad (1)$$

where the dependent variable is a posttest—either written literacy, oral literacy, or numeracy—administered to student *i* enrolled in school *s* in experimental strata *j*. The variable *iit* indicates initial assignment to the treatment group; the variables $pretest_e$ are pretests administered at the baseline in each subject *e*; the X_{isj} are a vector of controls including dummy variables indicating students' gender and cohort (see Figure 2); the δ_j are dummy variables indicating experimental strata; and ε is the idiosyncratic error assumed to be independent between school clusters but allowed to be correlated within them, following Bertrand et al. (2004). β is the nominally unbiased average effect of the program on student test scores. It is not, strictly speaking, necessary to control for baseline scores, though we always do so to improve the precision of estimated treatment effects, and to adjust for imbalance—albeit statistically insignificant—in the baseline scores across treatment and control groups (described in the next section).

4. Data Collection and Baseline Balance

4.1. APRHC data on schools and teachers

The African Population and Health Research Center (APHRC)—independently of the implementing organization—applied a variety of exams and surveys in treatment and control schools. The baseline data collection occurred in two phases. The first phase was conducted in July and August 2009 and included questionnaires addressed to the head teachers (principals) and classroom teachers, visual classroom observations, and achievement exams for students in grades 1 and 2. The second baseline occurred in February and March 2010 for newly-enrolled

grade 1 students.[7] For all cohorts, follow-up data collection was conducted in June and July 2011, 1.5 to 2 years after the baseline. During the follow-up, visual classroom observations occurred and subject examinations were given to students from the original three cohorts. Figure 2 provides an overview of data collection activities.

During the baseline, head teachers were asked a number of questions about their school and themselves.[8] In Table 1, panel A reports summary statistics for treatment and control schools. Because the randomization occurred at the cluster level, the standard errors of the differences in Table 1—and subsequent tables—are adjusted to allow for correlation at that level. Across treated and untreated schools, about one-third of the head teachers think teacher absenteeism is a problem; schools charge about US$1-$2 for miscellaneous schooling costs; about 4 hours each day are devoted to instruction; and the majority of the schools offer a primary school exit exam (the KCPE in Kenya or the PLE in Uganda). The means of these and other variables are not statistically different across treatment and control groups.

During the baseline conducted in July-August 2009, surveyors performed visual inspections of mathematics and language classrooms that served grades 1 and 2, noting the visibility of various classroom elements. In Table 1, panel B compares the visibility of these various elements across the treatment and the control group. Treatment schools were 11 percentage points more likely to have exercise books visible in the classroom, the only marginally statistically significant difference between the two groups. Most classrooms in both types of schools had chalkboards, exercise books, lesson plans, text books, visual teaching aids, and wall charts. The same survey

[7] Technically these students were subject to up to 2 months of the RtL trained teachers prior to the baseline survey. According to AKF, such exposure at the start of a child's schooling career should not have conveyed substantial benefit over the non-RtL early grade 1 instructional model. Students were not subject to the treatment of the mini-libraries because they were not delivered until April 2010. We provide pooled estimates and separate estimates by cohort.

[8] Of the 221 schools in the experiment, 96 percent (213) of head teachers completed the survey. Head teachers of control schools were 3.7 percentage points more likely to complete the survey than head teachers at treatment schools.

was repeated at the follow-up for the classrooms in which study cohorts should have been students (i.e., grades 2-4).[9] In section 5, we will use grade 2 data from the baseline and follow-up—the only common grade across the two surveys—to estimate the effect of the treatment on the various classroom attributes.

Additionally, teachers were surveyed at the baseline. They were asked questions about themselves, their teaching, and their classes. We use their answers to check for baseline balance in panel C. They are generally similar in their levels of experience and pre-service training, although teachers in control schools are 3 percentage points less likely to have at least a high school diploma.

Finally, at the baseline surveyors used the classroom roster to take attendance. For pupils who were not present, surveyors tried to ascertain from those pupils present whether those not in attendance were truly absent or no longer attending school. Based on this measure, absenteeism was approximately 10 percent in Kenya and 25 percent in Uganda at the baseline with no statistically significant difference between the treatment and the control groups. A similar exercise was conducted at the follow-up. Unfortunately, the status of individual students was not recorded.

4.3 APHRC data on students

APHRC administered student tests at the baseline and follow-up. To reduce data collection costs, a subsample of students were tested within each grade. If a school had multiple sections (or streams) in a given grade, a single stream was randomly selected at the baseline. During the July-August 2009 baseline of grades 1 and 2, a total of 20 students per grade per section were randomly sampled (blocking by gender, and sampling the number of boys and girls so as to

[9] Some students repeated one or more grades. Therefore, study pupils were in grades 1-4 at the follow-up.

preserve the gender proportions in the entire section). If fewer than 20 students were in a section, all were selected. The same procedure was followed during the 2010 baseline among newly-enrolled grade 1 students, although 25 students were randomly selected from each section.

Each student was administered exams in numeracy, written literacy, and oral literacy. The exams were prepared in consultation with the implementing organization (AKF), the independent evaluators (APHRC), national assessment experts, national curriculum experts, academics, and practitioners. Several stages were involved in developing the assessment tools. First, a pool of questions was developed in English, drawing on the primary school curriculum from both countries. For instance, in numeracy, the team came up with a pool of 50 test items in each grade. Second, the pool was pre-tested and refined to be sure it correctly assessed the competency levels of pupils in grades 1-3. Third, the final test items were selected from the pools and compiled into a single exam for each subject. The test was designed so that grade 1 students took only the grade 1 portion of the exam, grade 2 students took the grade 1 and grade 2 portions of the exam, and grade 3 pupils took the entire exam. Fourth, the literacy test items were translated into Kiswahili and Lango, the languages of instruction in the Kenya and Uganda study sites respectively. The numeracy exams in Kenya remained in English, the language of mathematics instruction in Kenya. AKF was aware of the types of questions that would be asked, given their input, but they did not know the exact questions. At each study school, the numeracy exams were administered first, followed by the written literacy exam. The final exam was the oral literacy exam that involved interaction between an enumerator and a student.

At the follow-up the same students from the baseline were tested if they were present on the day of the survey, regardless of their grade level at the time of the follow-up. Students were given the same test items taken at baseline, as well as any new items specific to their current

grade. Thus, a student from the January 2010 baseline who was in grade 2 at the follow-up would respond to the same grade 1 questions as in 2010 and also answer the grade 2 questions. Any baseline students that were absent were replaced by another randomly selected student of the same gender in their expected grade (e.g. absent students from the 2009 grade 2 baseline were replaced with students in grade 4 in 2011). Section 5 will describe baseline attrition in greater detail. Our preferred estimates use only students who were present at both baseline and follow-up, although a robustness check in section 6 will also include the replacement students.

The baseline and follow-up test forms varied depending on the grades in which they were applied. However, the use of repeated "anchor" items facilitates the linking of all tests to a common scale. Within each subject, we estimated a separate one-parameter (i.e., Rasch) item response theory model. For example, the numeracy model is estimated concurrently in a pooled sample of test item data across the baseline and follow-up numeracy tests applied to all grades in Kenya and Uganda. We then standardized the resulting logit scale in each exam by the mean and the standard deviation of the baseline. Hence, all subsequent effects—in Kenya and Uganda—can be interpreted as a proportion of the pooled standard deviation on the baseline test.

Table 2 presents the baseline characteristics of students across the treatment and control groups. Students are equally likely to be male, although students in control schools tended to have higher scores on all three tests than the treatment schools, though none of these differences is statistically significant. When disaggregated by country, we do find a statistically significant difference between the treatment and control on the numeracy exam in the Kenyan sample, favoring control schools. The immediate implication is that it is always preferable to control for baseline test scores in a regression that estimates the effect of the intervention on test scores. Given the common scale, we can further conclude that students in Kenya have markedly higher

average scores on all three baseline exams than the students in Uganda. The largest differences occur in the written literacy scores where Kenyan students score on average 1.3 standard deviations higher.

4.3. AKF data on implementation

Before the follow-up, enumerators from AKF rated the adherence of each treatment school to the preferred RtL model. A single enumerator rated all of the treated schools in each country. Each school was given a score from 0-11 based on the number of statements about implementation that were answered in the affirmative (see Table 3). The statements concerned teacher behavior, classroom learning environments, and school leadership. The raw scores were then converted by AKF into an implementation score of "high" for those with scores of 7-11, "medium" for schools with scores of 5-6, or "low" for schools with scores 0-4. Approximately 50% of the sample in each country was designated medium implementers, with about 25% in each of the other two designations. Section 5 assesses whether average treatment effects vary by the category of implementation.

5. Results

5.1. Attrition

Overall, 24 and 47 percent of the Kenyan and Ugandan baseline students, respectively, were not present at the follow-up testing (see Table 4). We do not have student-level data on the reason for their absence on the test day: it may be due to student drop-outs or simply due to absence on a single day. Based on classroom rosters of students examined at the follow-up, the estimated absenteeism, instead of dropping out, among enrolled students was 13 percent in

Kenya and 26 percent in Uganda. These figures imply that roughly half of the attriting students were absent on the test day, while the others had dropped out. Whatever the case, the natural concern is that non-random follow-up attrition introduces imbalance in observed or unobserved student attributes that are correlated with treatment status, perhaps biasing estimates of treatment effects. To partly assess the likelihood of bias, Table 4 reports on whether attrition differed across treatment and control groups, and whether it was correlated with baseline scores.

In columns 1 and 4, a dummy variable indicating attrition in each country is regressed on a treatment indicator and dummy variables indicating experimental strata. The probability of attriting is similar, in practical and statistical terms, for Kenyan treatment and control groups. However, students in the Ugandan treatment group were 5 percentage points less likely to attrit, with the difference statistically significant at 10%. Columns 2 and 5 further control for baseline test scores, although the evidence of differential test scores across attritors and non-attritors is mixed. Even if attritors are, on average, low-achieving, we are mainly concerned about potential imbalance in baseline attributes across attritors in treatment and control groups. Thus, columns 3 and 6 include interactions between the treatment group indicator and the baseline test scores. None of these coefficients are statistically significant. In conclusion, while levels of attrition are high—and not surprisingly so, given the context—differential attrition does not seem to be an important threat to the internal validity of the estimated treatment effects. Even so, our subsequent estimates always control for baseline test scores, and we provide two falsification tests in Section 6 in which we apply the higher attrition rates in the control group to the treatment group by first eliminating the highest and then the lowest baseline ability students in the treatment group.

5.2. Achievement

In Table 5, columns 1 to 3 present the coefficients from the estimation of equation 1. The sample in each column reflects students who took the follow-up test and at least one baseline test.[10] We provide separate estimates for Kenya (panel A) and Uganda (panel B). In Kenya (panel A) we find no statistically significant effect of the program on numeracy or written literacy scores.[11] We find that the program increased oral literacy score by 0.077, or 8 percent of the baseline standard deviation (overall, the control group scores increased by 1.23 standard deviations). In contrast, for Uganda (panel B) we find that the treatment increased written literacy scores by 19.9 percent of a standard deviation, and oral literacy by 17.9 percent of a standard deviation (column 3). We also find that the control group, overall, increased by much less than in Kenya: 78 percent of a standard deviation. Since RtL is mainly a literacy intervention, the lack of an effect in both countries for numeracy is not unexpected. Further, the lack of a numeracy effect—despite large effects on literacy in Uganda—provides indirect evidence that the results are not being driven by selective attrition at the follow-up.

Given the small number of clusters in Uganda, the standard formula for cluster-robust standard errors could result in inappropriately small standard errors. Therefore, we calculate the p-values associated with the wild cluster bootstrap-T method developed by Cameron, Gelbach, and Miller (2008), and report these in brackets underneath the standard cluster-robust standard errors. The p-values confirm that the literacy results are statistically significant at better than a 5% level in both cases, and that the numeracy results remain statistically insignificant. Given the

[10] At least 96 percent of each sample completed all three baseline exams. When students did not take one of the baseline exams, we recode the missing value to zero, and include a dummy variable indicating observations with missing values.

[11] The degrees of freedom for the critical values in all tables have been adjusted following Cameron, Gelbach, and Miller (2008).

consistency of the findings with those using typical cluster-robust standard errors formula and the adjusted critical value, subsequent tables omit these p-values.

Table 6 reports treatment effects by the three cohorts in each country, since each cohort was treated in different grades and for a different duration. As Figure 2 illustrated, cohort 1 (July 2009 baseline) started RtL late in the 2nd grade, their classroom RtL exposure ended at the conclusion of 3rd grade in 2010, and they were tested for the second time during 4th grade. Cohort 2 (July 2009 baseline) started RtL late in their 1st grade year with the follow-up exams occurring during 3rd grade. Finally, cohort 3 (February 2010 baseline) was treated for a month before the baseline survey in 1st grade and were tested in the follow-up during 2nd grade. We create three new treatment variables that are the interaction of the treatment indicator and each cohort dummy variable. In Kenya (panel A), for both measures of literacy, the effect on the oldest cohort (cohort 1) is statistically significant. But, in all cases we fail to reject the hypothesis that the point estimates of the three presented coefficients are the same. For Uganda (panel B), we find evidence of a positive treatment effect for both written and oral literacy across all cohorts, with the largest point estimates for the youngest cohort 3. But, as with Kenya, we fail to reject the equality of coefficients. For the sake of parsimony and statistical power, we pool all cohorts together for the remainder of the estimates.

Table 7 tests for heterogeneity in treatment effects by interacting the treatment indicator with students' baseline test scores and their gender. The results for baseline scores appear in columns 1 to 3. For numeracy in Kenya (panel A, column 1) we find a differential effect of treatment by baseline test score with higher baseline students accruing more benefit. We do not find a differential effect for any of the other examinations in Kenya. On the literacy exams in Uganda (panel B, columns 2 and 3) all treated students benefitted equally from the intervention,

regardless of baseline score. Additionally in Table 6 we test for differential effects by gender. Lucas and Mbiti (forthcoming) found that Kenyan girls scored on average 25 percent of a standard deviation lower than boys on the primary school exit exam. In a setting with such achievement disparities, heterogeneity by sex could have important implications. We find in Kenya that girls performed about 0.05 to 0.10 standard deviations better than boys on the follow-up exam, conditional on their baseline scores (panel A, columns 4-6), although the treatment effect is not different by sex. In Uganda the program effect is similarly homogeneous by sex, but the gender gap is reversed. In both treatment and control groups, boys score 0.07 to 0.09 standard deviations higher than girls on the follow-up exam, conditional on their baseline test scores (panel B, columns 4-6).

5.3. Classroom Instructional Materials

One hypothesis for the differential treatment effects across Kenya and Uganda is that classrooms received different quantities of prescribed instructional inputs, such as instructional materials and training. Table 8 partially assesses this by re-estimating equation 1 in the sample of 2^{nd} grade classrooms, the only grade that was visually assessed by APHRC surveyors at both baseline and follow-up. Each regression is a linear probability model with the dependent variable equal to 1 if the particular characteristic was visible in the classroom. The control variables include dummy variables indicating the presence of attributes in the baseline, including the lagged value of the dependent variable. Panel A contains the estimates for Kenya and Panel B for Uganda.

Consistent with the stated goals of the program we find that the likelihood of observing visual teaching aids (column 3), picture books (column 6), story books (column 7), other reading

materials (column 9), and wall charts (column 10) increased more in the treated schools across both countries. In Kenya, the probability of visible student made materials increased (column 8). Additionally, in Uganda, the likelihood of lesson notes (column 2), textbooks (column 5), and a chalkboard being visible (column 11) increased. Chalkboards were not part of the intervention, but could have been provided by the treatment schools themselves because of decreased budget expenditures on other items or due to teacher requests as a complement to the RtL approach. Table 8 also contains the average prevalence of each attribute in the baseline. On all measures except lesson plans and notes, Kenyan classrooms had more materials at the baseline.

The provision of classroom instructional materials was only one component of the intervention. To partially assess whether differences in materials can "explain" RtL treatment effects, Table 9 controls for changes in classroom instructional materials (see the table footnote for details). As with the original specifications for Kenya, we still find that the program had a small and statistically insignificant effect on numeracy and written literacy achievement. The point estimate on oral literacy has a similar magnitude but is less precisely estimated. For Uganda we still find no effect of treatment on numeracy scores and a positive effect on written and oral literacy scores. The point estimates for both literacy exams are still at least 85 percent the size of the original coefficients. We conclude that the average treatment effects within and across countries are likely explained by other features of the students, context, or treatments.

5.4. Implementation

AKF rated the implementation of RtL in each treated school, using criteria described in Table 3. The criteria are based on the perceived attitudes and classroom behaviors of head teachers, teachers, and students, rather than the physical presence of instructional materials. In columns 1

to 3 of Table 9, we test whether schools judged to be better implementers had superior achievement outcomes by replacing the single *itt* treatment variable in equation 1 with three separate dummy variables that interact the treatment indicator with indicators of high, medium, or low degrees of implementation fidelity. We cannot discard the possibility that implementation fidelity is endogenous to observed or unobserved variables that would lead to higher test scores even in the absence of treatment. We discuss and empirically test some of these characteristics below.

Panel A contains the results for Kenya. Across all test scores the effect of the treatment is monotonically decreasing in the quality of implementation with the point estimates for high implementing schools the largest. For literacy in Kenya, the high-implementing schools had positive and statistically significant effects on achievement beyond the improvements observed in control schools (columns 2 and 3). We reject that the point estimates are the same across the three levels of implementation. Panel B repeats the same analysis for Uganda. We find no evidence of any effect on numeracy (column 1). For the two literacy tests the point estimates monotonically decrease with lower-implementation, and both high and medium implementers had statistically significantly improved student achievement scores. Additionally, even the lowest implementing schools improved oral literacy. Even though the point estimates are different for the three levels of implementation we fail to statistically reject that they are equal.

Despite the appeal of Table 10's results, they do not conclusively show that implementation mediates the magnitude of treatment effects, since implementation is endogenous to attributes of students and schools. To assess this, Table 11 reports school-level means—including baseline test scores and head teacher data—across the sample of treatment schools that are high-, medium-, and low-implementers. There are large and statistically significant differences in

baseline test scores, as well as head-teacher-reported data on teacher absenteeism, the likelihood of completing the national curriculum, and the existence of excess demand for enrollment. In all cases, high-implementers are the "better" schools, suggesting that the greater effectiveness of high-implementers is an artifact of their heterogeneity in other regards. To partly assess this, columns 4 to 6 in Table 10 include the additional set of baseline controls described in Table 11. Interestingly enough, the pattern of effects is substantively similar, which may simply indicate that baseline student test scores already controlled for much of the heterogeneity across high- and low-implementers.[12] We cautiously interpret the results as evidence that implementation quality may mediate the size of treatment effects. The results also highlight the importance of understanding the determinants of implementation. Implementation quality covaries with the baseline attributes of school and students, in ways that are broadly suggestive of barriers to implementation in poorer and lower-quality schools across both countries.

6. Robustness

Table 12 provides evidence that the main experimental findings are robust to alternative specifications. Results are presented separately for Kenya (panel A) and Uganda (panel B), with the exception of column 7. Each column contains the coefficient of interest from 6 separate regressions. Column 1 repeats the prior findings from Table 5, and serves as a baseline against which to compare the following results.

Column 2 employs a revised specification that includes student fixed effects, as in:

$$test_{ist} = \alpha + \beta itt_s * followup_t + \gamma followup_t + \delta_i + \varepsilon_{ist}$$

[12] In other regressions, not reported here, we replicated the specifications in columns 1 to 3, excluding baseline test scores. In this case, the statistically insignificant coefficients on low-implementing Kenyan schools became strongly negative and significant across all tests (at least 20 percent of a standard deviation). The robustly positive coefficients for low-implementing Ugandan schools became statistically insignificant.

where $test_{ist}$ is the posttest or pretest in a particular subject of student i in school s at time t (baseline or follow-up). The dummy variable $followup_t$ indicates the follow-up period, and the δ_i are student fixed effects. The variable itt_s indicates whether the student is in a school that was ever assigned to the treatment. Hence, β identifies the average treatment effect. The regression is estimated in a stacked sample of student observations, restricted to students who took both the pretest and posttest of the specified subject.

The next estimates, in column 3, leverage all available students in a repeated cross-section specification. They include students even when they do not appear in the posttest (because of attrition) or the pretest (because they were "replacement" students randomly sampled from the same cohort, as described in section 4). The regression, estimated in a stacked sample of student-level data, is:

$$test_{ist} = \alpha + \beta itt_s * followup_t + \gamma followup_t + \lambda X_{isj} + \delta_s + \varepsilon_{ist}$$

where the δ_s are school fixed effects, and the other variables are as previously described.

Columns 4 to 6 return to the specification in column 1, but with alternative sample designations. Column 4 limits the sample to students who were present at the follow-up and took all three baseline exams. To assess whether differential attrition influences the pattern of results, columns 5 and 6 impose the (higher) proportion of attrition from the control schools on the treatment schools. The sample in column 5 "attrits" students from the treatment group with the highest baseline scores, while the sample in column 6 removes the lowest-scoring students from the treatment group.

In general, the robustness checks do not overturn the main pattern of results. The Kenyan treatment effects continue to be small and statistically insignificant in numeracy and written literacy. The effect of 0.077 in oral literacy, statistically significant at 10%, varies from 0.044 to

0.08 in other columns, though it is often estimated with less precision, as in columns 2 and 3. The Ugandan results are also consistent with the preferred estimate. There are small effects on numeracy (0.117-0.157) that are nonetheless not statistically distinguishable from zero. Across all columns, the large and statistically significant effects on written literacy are 0.20-0.232. For oral literacy, the range is 0.157-0.24, although coefficients for the student fixed-effects and repeated cross-section specifications have larger standard errors.

Lastly, column 7 uses a single sample that pools data from both countries, made possible by the use of commonly-scaled test instruments. Across both countries, the RtL treatment produced modest effect sizes on oral and written literacy of 0.10 and 0.12 standard deviations. Even after pooling the samples, the numeracy coefficient is not distinguishable from zero.

8. Conclusions

We used a randomized controlled trial to evaluate the Reading to Learn strategy for increasing early primary school literacy across multiple districts in Kenya and Uganda. Overall, we find that the approach, as implemented by AKF, resulted in higher student achievement in written and oral literacy in Uganda, by at least 18 percent of a standard deviation, but not numeracy. In Kenya, it had, at best, a small effect of 8 percent of a standard deviation on oral literacy. Additionally we found that those schools that adhered most closely to the ideal RtL model had the largest achievement gains, although we confirmed that implementation quality is highly correlated with observed student and school attributes, such as baseline test scores, and potentially with unobserved attributes as well.

The experiment provides a rare opportunity to gain insight into the external validity of impact evaluation results. The Kenyan and Ugandan experiments both assessed a common model for

improving learning. In each setting, a single organization, AKF, conducted and supervised the model's implementation, using common criteria to judge its success across schools. While implementation varied in each settings—and plausibly mediated program effects—it was not markedly better or worse in either setting. An independent organization, APRHC, managed the evaluation design and data collection. Each experiment employed a common set of instruments to measure student learning in numeracy and literacy, facilitating the linking of tests to a common scale using item response theory models.

Despite these similarities, the program's effects were larger in the Ugandan context. There are several remaining hypotheses as to why program effects varied by country. The first is that Kenyan and Ugandan students differed dramatically in their baseline levels of achievement (see Table 2), suggesting that the substantially lower-scoring Ugandan students were more able to benefit from the RtL instructional methods. But, in this case, one would expect to observe larger effects among relatively lower-scoring Kenyan children, and Table 10 provided no evidence of this type of treatment effect heterogeneity in either Kenya or Uganda.

Second, the Kenyan and Ugandan classrooms differed along several key dimensions at baseline. Most obviously, the Ugandan classrooms had lower levels of key instructional resources such as textbooks (see Table 8). There is less data to discern baseline levels of teacher content knowledge or instructional ability, but suppose that Ugandan schools had lower levels of instructional materials and teacher capacity, and that there are diminishing returns to these investments. In this case, the relatively larger effects in Uganda are an artifact of that country's more deprived instructional settings.

Lastly, the different results may partly stem from the language of the literacy assessments. In Uganda, the language was students' native language, and also one of primary languages of

instruction. In Kenya, the literacy assessments were in Swahili which, as it turned out, was not the mother tongue of most students. Indeed, Swahili was only taught as a foreign language, for just 30 minutes a day, and English was the primary language of instruction. It cannot be ascertained whether RtL effects might have been observed if English literacy assessments had been applied, but it is a plausible hypothesis.

References

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden (2007) "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, 122(3): 1235-1264.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. (2008) "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics*, 90(3): 414–427.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola (2005). "The Central Role of Noise in Evaluating Interventions that use Test Scores to Rank Schools." *American Economic Review*, 95(4): 1237-1258.

Dubeck, Margaret M., Matthew C. H. Jukes, and George Okello (2012). "Early Primary Literacy Instruction in Kenya." *Comparative Education Review*, 56(1): 48-68.

Glewwe, Paul, Eric A. Hanushek, Sarah D. Humpage, and Renato Ravina (2011). "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010." NBER Working paper 17554.

Glewwe, Paul, and Michael Kremer (2006). "Schools, Teachers, and Educational Outcomes in Developing Countries." In *Handbook of the Economics of Education*, edited by Eric A. Hanushek and Finis Welch. Amsterdam: North Holland: 943-1017.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin (2009). "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1(1): 112-35.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz (2004). "Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya." *Journal of Development Economics* 74(1): 251-268.

Lucas, Adrienne M. and Isaac Mbiti (2012). "Access, Sorting, and Achievement: the Short-Run Effects of Free Primary Education in Kenya." *American Economic Journal: Applied Economics*, October.

Lucas, Adrienne M. and Isaac Mbiti (Forthcoming). "Does Free Primary Education Narrow Gender Differences in Schooling Outcomes? Evidence from Kenya." *Journal of African Economies*.

Lucas, Adrienne M. and Isaac Mbiti (2012). "Effects of Attending Selective Secondary Schools on Student Achievement: Discontinuity Evidence from Kenya." Unpublished.

Tan, Jee-Peng, Julia Lane, and Gerard Lassibille (1999). "Student Outcomes in Philippine Elementary Schools: An Evaluation of Four Experiments." *The World Bank Economic Review* 13(3): 493-508.

Figure 1A: Randomization Framework, Kenya

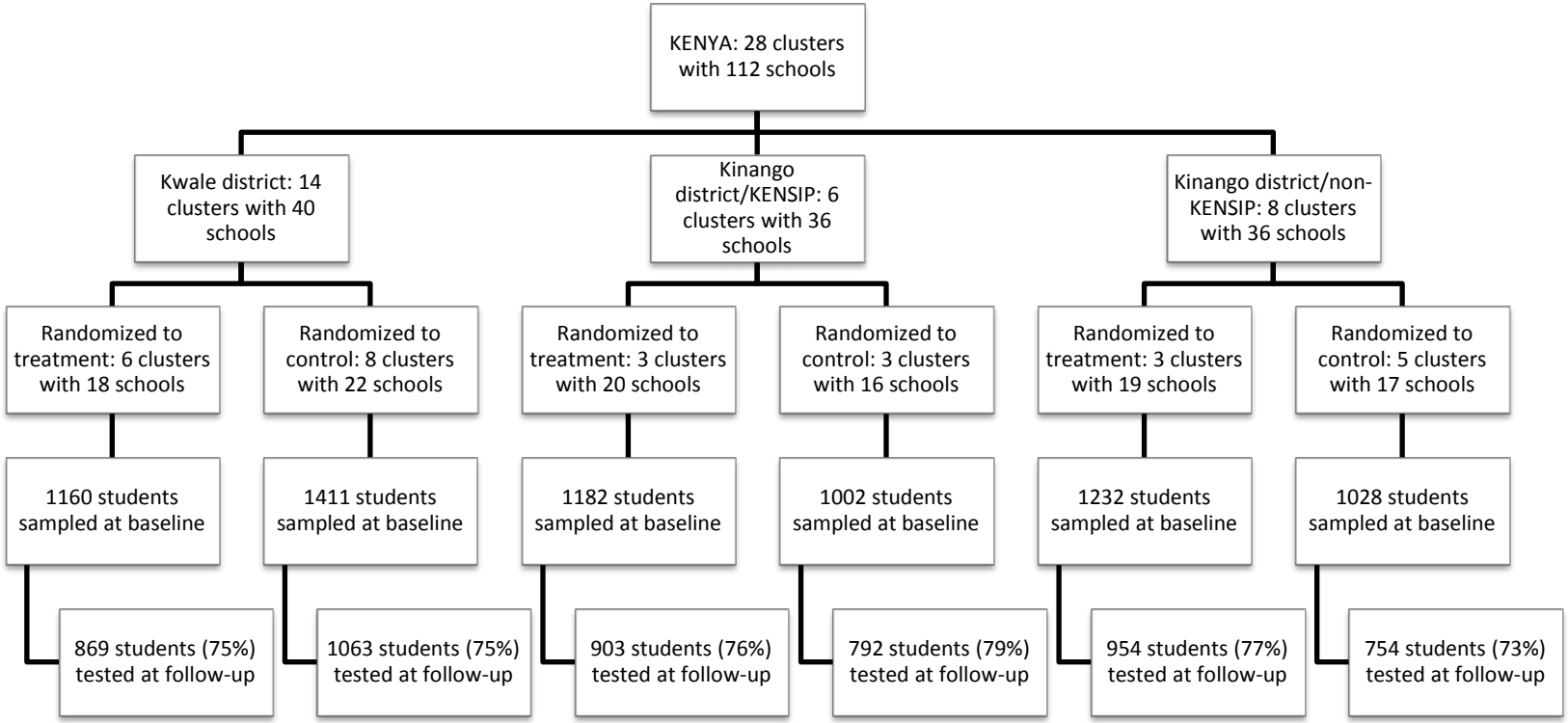

Figure 1B: Randomization Framework, Uganda

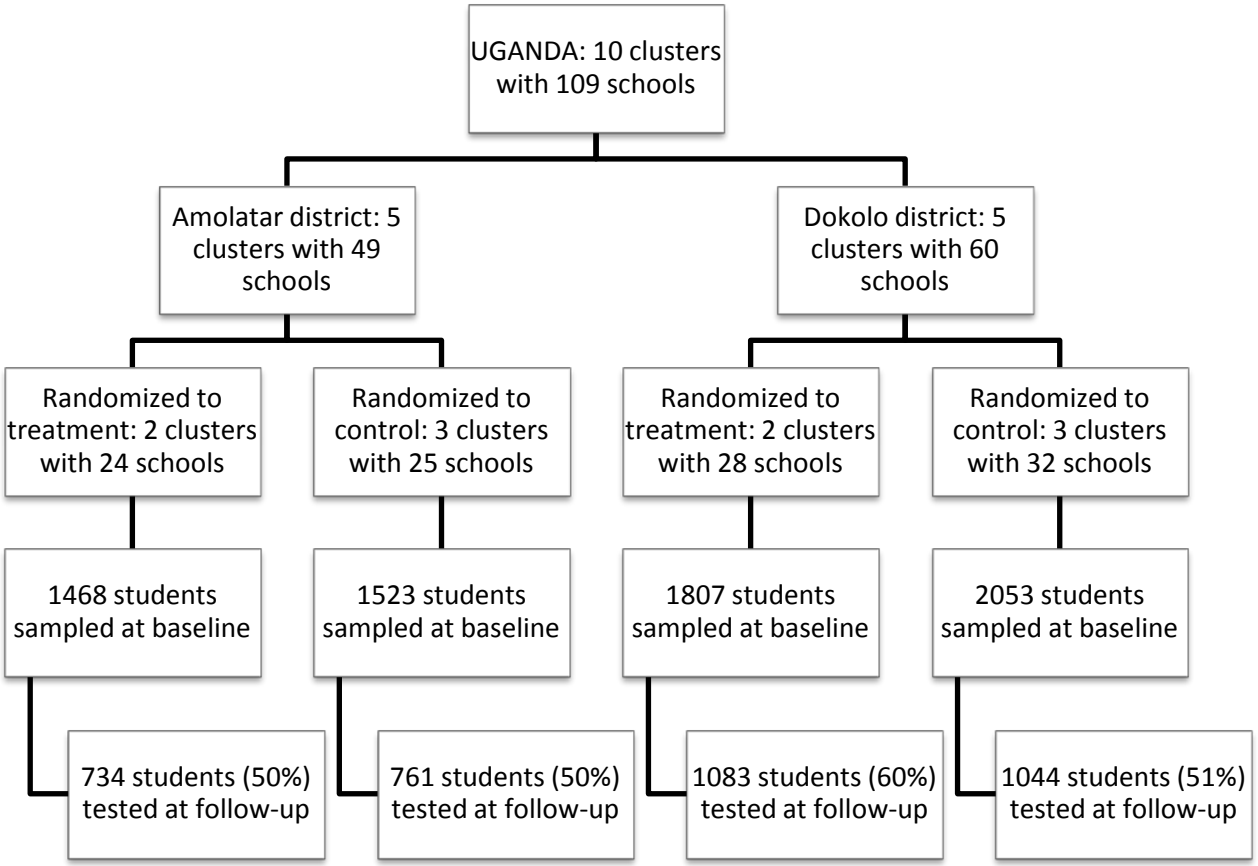

Figure 2: Project Timeline

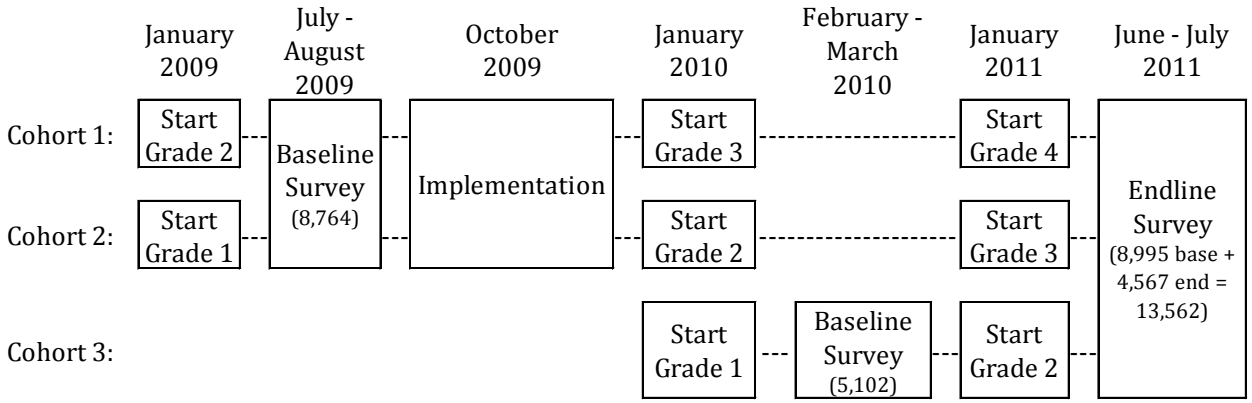

Table 1: Baseline School Characteristics

	Control	Treatment	Difference
Panel A: Head Teacher Survey (sample size: 213)			
Teacher Absenteeism is a Problem	0.39	0.32	0.07
	(0.49)	(0.47)	(0.10)
Average Total School Fees (US$)	1.54	2.10	-0.56
	(1.93)	(3.39)	(0.49)
Percent of School Families Head Teacher Knows	0.48	0.48	0.01
	(0.26)	(0.21)	(0.04)
Number of Instructional Hours	4.32	4.23	0.09
	(0.83)	(0.90)	(0.11)
School is a KCPE (Kenya) or PLE (Uganda) Center	0.85	0.82	0.04
	(0.35)	(0.39)	(0.05)
Number of Teachers Departing in Last 12 Months	1.71	1.37	0.34
	(1.85)	(1.43)	(0.29)
Panel B: Observed Classroom Characteristics (1=visible, sample size=649)			
Chalkboard	0.87	0.91	-0.04
	(0.33)	(0.29)	(0.03)
Exercise Books	0.67	0.79	-0.11 *
	(0.47)	(0.41)	(0.06)
Lesson Notes	0.51	0.49	0.02
	(0.50)	(0.50)	(0.07)
Lesson Plan	0.73	0.71	0.02
	(0.45)	(0.45)	(0.06)
Other Materials	0.19	0.25	-0.05
	(0.40)	(0.43)	(0.05)
Picture Books	0.20	0.18	0.02
	(0.40)	(0.38)	(0.04)
Text Books	0.52	0.54	-0.01
	(0.50)	(0.50)	(0.09)
Story Books	0.19	0.23	-0.04
	(0.39)	(0.42)	(0.05)
Student Made Materials	0.21	0.22	-0.01
	(0.41)	(0.41)	(0.05)
Visual Teaching Aids	0.51	0.53	-0.02
	(0.50)	(0.50)	(0.07)
Wall Charts	0.52	0.56	-0.04
	(0.50)	(0.50)	(0.07)
Panel C: Teacher Survey (sample size: 556)			
Education of at Least a High School Diploma	0.92	0.96	-0.03 *
	(0.27)	(0.20)	(0.02)
Number of Years of Teaching	12.79	12.00	0.79
	(8.78)	(9.17)	(0.95)
Number of Years at Current School	5.50	4.92	0.59
	(5.76)	(5.59)	(0.63)
Number of Years Teaching Current Subject	9.40	9.13	0.26
	(8.16)	(8.63)	(0.84)
Pre-Service Teacher Training	0.84	0.82	0.02
	(0.37)	(0.38)	(0.06)

Notes: Treatment and control group as originally defined by APHRC. * significant at 10%, ** significant at 5%, *** significant at 1% based on standard errors clustered at the unit of randomization (cluster in Kenya, sub-county in Uganda).

Table 2: Baseline Student Characteristics

	Control	Treatment	Difference	
Proportion Male	0.490	0.500	-0.010	
	(0.50)	(0.50)	(0.006)	
Numeracy Score	0.053	-0.054	0.107	
	(1.03)	(0.96)	(0.152)	
Kenya	0.463	0.257	0.206	**
	(0.822)	(0.821)	(0.087)	
Uganda	-0.342	-0.394	0.052	
	(1.061)	(0.990)	(0.089)	
Written Literacy Score	0.027	-0.028	0.055	
	(1.02)	(0.98)	(0.259)	
Kenya	0.686	0.573	0.114	
	(0.741)	(0.722)	(0.097)	
Uganda	-0.609	-0.683	0.074	
	(0.828)	(0.784)	(0.085)	
Oral Literacy Score	0.036	-0.037	0.072	
	(1.023)	(0.975)	(0.122)	
Kenya	0.291	0.209	0.082	
	(0.971)	(0.934)	(0.122)	
Uganda	-0.215	-0.309	0.094	
	(1.010)	(0.947)	(0.076)	

Notes: Treatment and control group as originally defined by randomization. Total student baseline sample: 13,931. Kenya: 3574 treatment and 3441 control. Uganda: 3275 treatment and 3576 control. * significant at 10%, ** significant at 5%, *** significant at 1% based on standard errors clustered at the unit of randomization (cluster in Kenya, sub-county in Uganda).

Table 3: Implementation Questions

Teachers	
1	Teachers are effectively using the five RtL steps in the correct sequence.
2	Teaching is done procedurally and with logical understanding and is not mechanical.
3	Teachers are innovative and committed to implementing the approach.
4	Teachers are motivated to support learners in numeracy and literacy outside teaching time.
Classroom Learning Environments	
5	Appropriate learning materials are used.
6	The classroom library is utilized.
7	Children are reading age appropriate texts.
8	There is enhanced peer support among learners.
School Leadership	
9	Head teachers provide technical support.
10	School and parents have a supportive relationship.
11	Functional school development plans prioritize lower grades.

Table 4: Absenteeism at Endline

	Absent at Endline					
	Kenya			Uganda		
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	-0.003	-0.007	-0.007	-0.051*	-0.053**	-0.044
	(0.015)	(0.015)	(0.020)	(0.015)	(0.020)	(0.025)
Numeracy Score		-0.018	-0.001		-0.039***	-0.043***
		(0.011)	(0.012)		(0.009)	(0.012)
Written Literacy Score		0.020*	0.024		-0.005	-0.004
		(0.010)	(0.015)		(0.006)	(0.012)
Oral Literacy Score		-0.025*	-0.037**		-0.014	-0.019
		(0.013)	(0.018)		(0.014)	(0.024)
Treated X Numeracy Score			-0.033			0.010
			(0.021)			(0.019)
Treated X Written Literacy Score			-0.009			-0.001
			(0.020)			(0.012)
Treated X Oral Literacy Score			0.022			0.003
			(0.024)			(0.027)
F-test of Joint Significance of Interaction Effects						
F-Statistic			0.90			0.48
P-Value			0.45			0.71
Observations	7040	7,040	7,040	6,891	6,891	6,891
R-squared	0.001	0.01	0.01	0.01	0.02	0.02
Portion of Baseline Absent		0.24			0.47	

Notes: Treatment and control group as originally defined by APHRC. Sample of all students who completed at least one exam at baseline. All columns include strata dummy variables as additional controls. * significant at 10%, ** significant at 5%, *** significant at 1%. Standard errors clustered at the unit of randomization (cluster in Kenya, sub-county in Uganda) appear in parenthesis. Critical values adjusted for number of clusters.

Table 5: Effect of Treatment on Achievement

	Numeracy	Written Literacy	Oral Literacy
	(1)	(2)	(3)
Panel A: Kenya			
Treated	-0.011	0.024	0.077*
	(0.059)	(0.032)	(0.042)
Baseline Numeracy Score	0.308***	0.103***	0.122***
	(0.021)	(0.023)	(0.026)
Baseline Written Literacy Score	0.181***	0.150***	0.233***
	(0.022)	(0.021)	(0.024)
Baseline Oral Literacy Score	0.473***	0.548***	0.505***
	(0.030)	(0.027)	(0.032)
Observations	5,323	5,302	5,305
R-Squared	0.58	0.55	0.52
Average Control Group Change	1.02	0.81	1.23
Panel B: Uganda			
Treated	0.117	0.199***	0.179***
	(0.087)	(0.054)	(0.047)
	[0.260]	[0.009]	[0.011]
Baseline Numeracy Score	0.254***	0.218***	0.139***
	(0.014)	(0.018)	(0.020)
Baseline Written Literacy Score	0.106***	0.147***	0.125***
	(0.025)	(0.034)	(0.034)
Baseline Oral Literacy Score	0.228***	0.340***	0.266***
	(0.020)	(0.033)	(0.023)
Observations	3,604	3,596	3,575
R-Squared	0.40	0.42	0.38
Average Control Group Change	0.50	0.78	0.78

Notes: Sample of students who completed specified endline test and at least one baseline test. Test scores measured in standard deviations. Treatment defined through original APHRC randomization. Students who did not take a particular baseline test are given a score of 0 and a dummy variable is included for each missing baseline test score. All regressions include gender, cohort, and strata dummy variables. Standard errors clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda) appear in parentehsis. * significant at 10%, ** significant at 5%, *** significant at 1%. Critical values adjusted for number of clusters. Panel B: P-values associated with wild cluster bootstrap standard errors appear in sqaure brackets.

Table 6: Heterogeneous Effect of Treatment by Cohort

	Numeracy	Written Literacy	Oral Literacy
	(1)	(2)	(3)
Panel A: Kenya			
Treated X Cohort 1	0.125	0.091*	0.135**
	(0.082)	(0.052)	(0.065)
Treated X Cohort 2	-0.037	-0.008	0.08
	(0.079)	(0.054)	(0.071)
Treated X Cohort 3	-0.099	-0.003	0.028
	(0.080)	(0.068)	(0.075)
F-Test of Equality of Interaction Coefficients			
F-Statistic	2.12	0.85	0.38
P-Value	0.14	0.44	0.69
Observations	5,323	5,302	5,305
R-Squared	0.58	0.55	0.52
Panel B: Uganda			
Treated X Cohort 1	0.065	0.146*	0.179**
	(0.114)	(0.065)	(0.070)
Treated X Cohort 2	0.162	0.165*	0.169**
	(0.092)	(0.076)	(0.053)
Treated X Cohort 3	0.125	0.276**	0.187**
	(0.096)	(0.091)	(0.061)
F-Test of Equality of Interaction Coefficients			
F-Statistic	0.86	0.91	0.03
P-Value	0.45	0.44	0.97
Observations	3,604	3,596	3,575
R-Squared	0.40	0.42	0.38

Notes: Sample of students who completed specified endline test and at least one baseline test. Cohort 1: Grade 2 in 2009. Cohort 2: Grade 1 in 2009. Cohort 3: Grade 1 in 2010. Test scores measured in standard deviations. Treatment defined through original APHRC randomization. Students who did not take a particular baseline test are given a score of 0 and a dummy variable is included for each missing baseline test score. All regressions include baseline test scores and gender, cohort, and strata dummy variables. Standard errors clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda) appear in parenthesis. * significant at 10%, ** significant at 5%, *** significant at 1%. Critical values adjusted for number of clusters.

Table 7: Heterogeneous Effect of Treatment by Baseline Characteristics

	Numeracy	Written Literacy	Oral Literacy	Numeracy	Written Literacy	Oral Literacy
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Kenya						
Treated	-0.048	-0.003	0.077	-0.019	0.033	0.070
	(0.064)	(0.051)	(0.046)	(0.061)	(0.038)	(0.045)
Treated X Baseline Score	0.104**	0.044	0.008			
	(0.040)	(0.040)	(0.038)			
Treated X Male				0.017	-0.017	0.014
				(0.033)	(0.031)	(0.046)
Baseline Numeracy Score	0.255***	0.104***	0.122***	0.308***	0.103***	0.122***
	(0.021)	(0.023)	(0.026)	(0.021)	(0.023)	(0.026)
Baseline Written Literacy Score	0.179***	0.149***	0.230***	0.181***	0.149***	0.233***
	(0.022)	(0.021)	(0.032)	(0.022)	(0.021)	(0.024)
Baseline Oral Literacy Score	0.475***	0.526***	0.503***	0.473***	0.548***	0.505***
	(0.030)	(0.040)	(0.032)	(0.030)	(0.027)	(0.032)
Male	-0.038**	-0.098***	-0.090***	-0.047**	-0.088***	-0.098***
	(0.017)	(0.016)	(0.023)	(0.021)	(0.019)	(0.030)
F-Test for Joint Significance of Treatment Coefficients						
F-Statistic	3.58	1.96	2.02	0.15	0.39	1.65
p-Value	0.04	0.16	0.15	0.86	0.68	0.21
Observations	5,305	5,286	5,285	5,323	5,302	5,305
R-Squared	0.58	0.55	0.52	0.58	0.55	0.52
Panel B: Uganda						
Treated	0.119	0.152***	0.185***	0.100	0.223***	0.218***
	(0.085)	(0.036)	(0.045)	(0.087)	(0.053)	(0.048)
Treated X Baseline Score	-0.002	-0.082	-0.015			
	(0.029)	(0.046)	(0.050)			
Treated X Male				0.035	-0.048	-0.080*
				(0.041)	(0.040)	(0.043)
Baseline Numeracy Score	0.254***	0.217***	0.136***	0.254***	0.218***	0.139***
	(0.018)	(0.018)	(0.020)	(0.014)	(0.018)	(0.020)
Baseline Written Literacy Score	0.108***	0.145***	0.135***	0.106***	0.147***	0.125***
	(0.025)	(0.035)	(0.025)	(0.025)	(0.034)	(0.034)
Baseline Oral Literacy Score	0.229***	0.379***	0.269***	0.228***	0.340***	0.267***
	(0.020)	(0.047)	(0.027)	(0.020)	(0.033)	(0.023)
Male	0.105***	0.044*	0.026	0.091**	0.065*	0.065*
	(0.022)	(0.023)	(0.025)	(0.032)	(0.031)	(0.031)
F-Test for Joint Significance of Treatment Coefficients						
F-Statistic	0.99	8.93	8.71	1.18	9.22	10.76
p-Value	0.41	0.01	0.01	0.35	0.01	0.00
Observations	3,585	3,572	3,495	3,604	3,596	3,575
R-Squared	0.40	0.42	0.38	0.40	0.42	0.38

Notes: Sample of students who completed specified endline test and at least one baseline test. Test scores measured in standard deviations. Treatment defined through original APHRC randomization. Students who did not take a particular baseline test are given a score of 0 and a dummy variable is included for each missing baseline test score. All regressions include cohort and strata dummy variables. Standard errors clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda) appear in parenthesis. * significant at 10%, ** significant at 5%, *** significant at 1%. Critical values adjusted for number of clusters.

Table 8: Effect of Treatment on Classroom Characteristics

	Lesson Plans	Lesson Notes	Visual Teaching Aids	Exercise Books	Text Books	Picture Books	Story Books	Student Made Materials	Other Reading Materials	Wall Charts	Chalkboard
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Panel A: Kenya											
Treated	0.028	0.093	0.355***	0.041	0.166	0.207*	0.527***	0.230**	0.173*	0.206**	-0.014
	(0.074)	(0.085)	(0.089)	(0.057)	(0.101)	(0.115)	(0.062)	(0.109)	(0.096)	(0.096)	(0.073)
Observations	192	192	192	192	192	192	192	192	192	192	192
R-Squared	0.20	0.14	0.33	0.11	0.20	0.18	0.33	0.18	0.16	0.30	0.17
Baseline Average	0.68	0.41	0.67	0.85	0.73	0.21	0.29	0.39	0.26	0.65	0.88
Panel B: Uganda											
Treated	0.090	0.143*	0.501***	0.124	0.298**	0.283***	0.399***	0.167	0.402***	0.513***	0.165**
	(0.074)	(0.069)	(0.072)	(0.081)	(0.114)	(0.057)	(0.058)	(0.115)	(0.080)	(0.062)	(0.057)
Observations	167	167	167	167	167	167	167	167	167	167	167
R-Squared	0.16	0.14	0.30	0.14	0.17	0.29	0.29	0.17	0.33	0.37	0.15
Baseline Average	0.79	0.72	0.37	0.59	0.39	0.20	0.17	0.11	0.14	0.40	0.87

Notes: Sample of second grade classrooms from schools that were sampled in baseline and endline. Dependent variable equal to 1 if item was visible. Linear probability models. All regressions include dummy variables for all baseline classroom characteristics and strata. Standard errors clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda) appear in parenthesis. * significant at 10%, ** significant at 5%, *** significant at 1%. Critical values adjusted for number of clusters.

Table 9: Effect of Treatment Net of Classroom Characteristics

	Kenya			Uganda		
	Numeracy	Written Literacy	Oral Literacy	Numeracy	Written Literacy	Oral Literacy
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	0.040	0.040	0.070	0.132	0.171**	0.155**
	(0.068)	(0.032)	(0.042)	(0.093)	(0.068)	(0.064)
Changes in Classroom Characteristics						
Lesson Plans	0.051	0.03	0.04	-0.026	-0.147*	-0.187**
	(0.067)	(0.040)	(0.051)	(0.091)	(0.079)	(0.079)
Lesson Notes	-0.029	-0.029	-0.043	-0.069	0.088	0.061
	(0.041)	(0.028)	(0.032)	(0.042)	(0.062)	(0.054)
Visual Teaching Aids	0.042	0.022	0.011	-0.031	-0.031	-0.079
	(0.065)	(0.045)	(0.061)	(0.048)	(0.032)	(0.051)
Exercise Books	-0.001	-0.032	-0.082	0.039	0.004	0.015
	(0.081)	(0.052)	(0.059)	(0.044)	(0.027)	(0.035)
Text Books	-0.039	-0.037	-0.018	-0.005	0.092*	0.054
	(0.068)	(0.043)	(0.041)	(0.043)	(0.045)	(0.055)
Picture Books	-0.056	-0.064*	-0.023	-0.079	-0.068	-0.025
	(0.060)	(0.031)	(0.051)	(0.046)	(0.060)	(0.053)
Story Books	-0.033	0.008	0.031	0.033	0.086	0.114**
	(0.079)	(0.044)	(0.049)	(0.067)	(0.064)	(0.041)
Study Materials	-0.055	0.004	0.016	0.055	0.029	0.053
	(0.050)	(0.037)	(0.033)	(0.065)	(0.047)	(0.046)
Other Materials	0.043	0.025	0.034	0.021	0.025	-0.034
	(0.073)	(0.051)	(0.061)	(0.075)	(0.086)	(0.046)
Wall Charts	-0.015	-0.006	0.01	-0.007	-0.069	-0.024
	(0.077)	(0.044)	(0.057)	(0.067)	(0.077)	(0.053)
Chalkboard	0.103	0.025	0.029	-0.043	-0.007	0.049
	(0.066)	(0.049)	(0.068)	(0.091)	(0.046)	(0.061)
Test for Joint Significance of Coefficients on Classroom Characteristics						
F-Statistic	1.35	1.45	1.33	2.46	4.75	8.06
P-Value	0.25	0.21	0.27	0.10	0.02	0.00
Observations	4,666	4,646	4,646	3,126	3,120	3,099
R-Squared	0.58	0.55	0.52	0.41	0.44	0.40

Notes: Sample of students who completed specified endline test and at least one baseline test. Test scores measured in standard deviations. Treatment defined through original APHRC randomization. Students who did not take a particular baseline test are given a score of 0 and a dummy variable is included for each missing baseline test score. All regressions include baseline test score and gender, cohort, and strata dummy variables. Changes in classroom characteristics measured as -1 (item was present in baseline but not endline), 0 (no change in presence or absence of item), or 1 (item was absent at baseline and present at endline). Standard errors clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda) appear in parenthesis. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 10: Heterogeneity by the Degree of Implementation

				With Additional Baseline Controls		
	Numeracy	Written Literacy	Oral Literacy	Numeracy	Written Literacy	Oral Literacy
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Kenya						
Treated X High Implementation	0.039	0.072**	0.123**	0.141**	0.123***	0.193***
	(0.055)	(0.031)	(0.046)	(0.061)	(0.024)	(0.034)
Treated X Medium Implementation	-0.044	-0.015	0.051	-0.036	-0.030	0.068
	(0.063)	(0.031)	(0.052)	(0.074)	(0.050)	(0.068)
Treated X Low Implementation	-0.121	-0.059	0.002	-0.022	-0.054	0.025
	(0.098)	(0.053)	(0.041)	(0.109)	(0.052)	(0.053)
Test for Equality of Implementation Coefficients						
F-Statistic	1.48	5.19	4.38	2.88	8.37	3.95
P-Value	0.25	0.01	0.02	0.07	0.00	0.03
Test for Joint Significance of Baseline Characteristics						
F-Statistic				17.63	12.67	19.32
P-Value				0.00	0.00	0.00
Observations	5,323	5,302	5,305	5,026	5,008	5,008
R-Squared	0.51	0.52	0.49	0.53	0.53	0.50
Panel B: Uganda						
Treated X High Implementation	0.073	0.351***	0.267***	0.036	0.425***	0.302***
	(0.077)	(0.073)	(0.076)	(0.074)	(0.056)	(0.087)
Treated X Medium Implementation	0.176	0.216**	0.207**	0.211	0.253***	0.261***
	(0.121)	(0.081)	(0.076)	(0.121)	(0.067)	(0.058)
Treated X Low Implementation	0.148	0.121	0.145*	0.170**	0.212**	0.171*
	(0.092)	(0.093)	(0.067)	(0.065)	(0.081)	(0.081)
Test for Equality of Implementation Coefficients						
F-Statistic	1.7	1.65	0.55	7.54	20.92	0.58
P-Value	0.23	0.25	0.60	0.01	0.00	0.58
Test for Joint Significance of Baseline Characteristics						
F-Statistic				7.54	2.44	73.71
P-Value				0.00	0.10	0.00
Observations	3,604	3,596	3,575	3,150	3,146	3,123
R-Squared	0.37	0.37	0.37	0.38	0.39	0.36

Notes: Sample of students who completed specified endline test and at least one baseline test. Test scores measured in standard deviations. Treatment defined through original APHRC randomization. All regressions include controls for all three baseline tests (students who did not take a particular test are given a score of 0) and dummy variables for each missing baseline test score, gender, and strata. Standard errors clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda) apear in parenthesis. * significant at 10%, ** significant at 5%, *** significant at 1%. Critical values adjusted for number of clusters.

Table 11: Baseline Correlates of Implementation

	Average at Baseline by Implementation Level			P-value of Test of Equality
	High	Medium	Low	
	(1)	(2)	(3)	(4)
Standardized Aggregate Score	0.16	-0.13	-0.14	0.03
Lower Primary Student Attendance Rate	0.83	0.82	0.82	0.78
Total Enrolled Lower Primary Students	88.9	95.1	86.0	0.43
Teacher Absenteeism is a Problem	0.19	0.31	0.41	0.04
Fraction of Teachers with a High School Diploma	0.96	0.97	0.97	0.97
Fraction of Teachers who Received Pre-Service Training	0.88	0.83	0.84	0.63
Required Yearly National Curriculum is Not Completed	0.16	0.31	0.48	0.01
Number of English Advisor Visits in the Last 18 Months	1.11	1.83	1.22	0.30
Number of Math Advisor Visits in the Last 18 Months	1.07	1.81	1.48	0.27
Number of School Inspector Visits in the Last 12 Months	2.07	2.10	2.22	0.39
Textbooks are Provided by the School to Pupils	0.83	0.68	0.65	0.46
Pupils are Allowed to Take Textbooks Home	0.29	0.10	0.22	0.34
Number of Learner Hours Per Day	3.93	4.46	4.13	0.15
Number of Discipline Events Per Week	3.09	3.01	3.46	0.90
Fraction of Students Benefiting from Feeding Program	0.57	0.45	0.42	0.52
Adequate Drinking Water is Available Throughout the Year	0.35	0.31	0.33	0.74
Fraction of Classroom With at Least One Repeating Student	0.90	0.90	0.85	0.62
Excess Demand for School Places	0.21	0.23	0.09	0.00
More Teachers Have Been Hired Than Left in the Last 12 Months	0.52	0.38	0.52	0.16
Number of Teachers Who Have Been Replaced in the Last 12 Months	1.38	1.08	0.65	0.34

Notes: Sample includes treatment schools with a head teacher questionnaire and AKF determined implementation score. Columns (1)-(3): mean values of stated variable. Column (4): P-value associated F statistic from test of equality of implementation coefficients from regression including strata dummy variables and standard errors clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda).

Table 12: Robustness

	Preferred Specification	Panel with Student Fixed Effects	Repeated Cross Section	Students with Three Baseline Test Scores	Adjusted Absenteeism		Single Treatment Effect Across Both Countries
					Remove Pupils with Highest Baseline Score	Remove Pupils with Lowest Baseline Score	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Kenya							
Dependent Variable: Numeracy Score							
Treatment Effect	-0.011	0.035	0.057	-0.007	-0.011	-0.011	0.042
	(0.059)	(0.079)	(0.083)	(0.059)	(0.059)	(0.059)	(0.051)
Observations	5,323	10,610	12,338	5,275	5,310	5,310	8,927
R-Squared	0.58	0.54	0.34	0.58	0.57	0.58	0.60
Dependent Variable: Written Literacy Score							
Treatment Effect	0.024	0.033	0.039	0.026	0.026	0.024	0.100***
	0.024	(0.031)	(0.033)	(0.032)	(0.033)	(0.033)	(0.034)
Observations	5,302	10,572	12,310	5,254	5,289	5,289	8,898
R-Squared	0.55	0.63	0.40	0.55	0.55	0.55	0.62
Dependent Variable: Oral Literacy Score							
Treatment Effect	0.077*	0.044	0.060	0.080*	0.078*	0.077*	0.124***
	(0.042)	(0.053)	(0.051)	(0.042)	(0.043)	(0.043)	(0.035)
Observations	5,305	10,570	12,311	5,257	5,293	5,292	8,880
R-Squared	0.52	0.68	0.49	0.52	0.52	0.52	0.54
Panel B: Uganda							
Dependent Variable: Numeracy Score							
Treatment Effect	0.117	0.157	0.131	0.124	0.117	0.12	
	(0.087)	(0.139)	(0.144)	(0.085)	(0.086)	(0.086)	
Observations	3,604	7,170	10,455	3,494	3,503	3,530	
R-Squared	0.40	0.26	0.23	0.40	0.38	0.381	
Dependent Variable: Written Literacy Score							
Treatment Effect	0.199***	0.227**	0.232**	0.204***	0.201***	0.201***	
	(0.054)	(0.080)	(0.076)	(0.055)	(0.056)	(0.054)	
Observations	3,596	7,144	10,425	3,486	3,496	3,522	
R-Squared	0.42	0.52	0.49	0.42	0.40	0.41	
Dependent Variable: Oral Literacy Score							
Treatment Effect	0.179***	0.240	0.239	0.186***	0.157***	0.179***	
	(0.047)	(0.136)	(0.132)	(0.045)	(0.048)	(0.046)	
Observations	3,575	6,990	10,289	3,465	3,479	3,501	
R-Squared	0.38	0.49	0.51	0.38	0.35	0.37	

Notes: Coefficients estimated separately for each Panel with the exception of Column 7. Column 1: from Table 4. Column 2: Data transformed into two observations per student. Sample limited to students with both baseline and follow-up scores for a specified exam. Includes student fixed effects and a post dummy variable. Displayed coefficient is from the interaction of post times treated. Column 3: Sample includes all surveyed students, even those who took only baseline or follow-up exams. Includes school fixed effects and sex, cohort, and post dummy variables. Displayed coefficient is from the interaction of post times treated. Columns 4-7: same controls and specification as Column 1. Column 4: Sample limited to students who completed all three baseline exams. Columns 5 and 6: Treatment sample adjusted to mirror level of attrition in the control group. Column 5: Students with the highest average baseline scores removed. Column 6: Students with the lowest average baseline scores removed. Column 7: Single sample with data from both countries. Standard errors clustered at the unit of randomization (cluster for Kenya, sub-county for Uganda) included in parenthesis. Crititcal values adjusted for the number of clusters. * significant at 10%, ** significant at 5%, *** significant at 1%.