Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda

Adrienne M. Lucas
Patrick J. McEwan
Moses Ngware
Moses Oketch

Abstract

Primary school enrollments have increased rapidly in sub-Saharan Africa, spurring concerns about low levels of learning. We analyze field experiments in Kenya and Uganda that assessed whether the Reading to Learn intervention, implemented by the Aga Khan Foundation in both countries, improved early-grade literacy as measured by common assessments. We find that Ugandan literacy (in Lango) increased by 0.2 standard deviations. We find a smaller effect (0.08) on a Swahili literacy test in Kenya. We find no evidence that differential effects are explained by baseline differences across countries in student test scores, classroom attributes, or implementation fidelity. A plausible explanation that cannot be directly tested is differential effective exposure to the literacy treatment in the tested languages. Students in Kenya were tested in Swahili, which is not necessarily the main language of instruction in primary schools, despite official policy. © *2014 by the Association for Public Policy Analysis and Management.*

INTRODUCTION

In the past two decades, African governments and international organizations have invested heavily in increasing the quantity of primary schooling. In Kenya and Uganda, the most notable influx of students occurred with the elimination of school fees in 2003 and 1997, respectively (Grogan, 2008; Lucas & Mbiti, 2012a). With increased enrollments, policy discussions have shifted to the very low levels of achievement among primary school students (Uwezo, 2011a, 2011b). Yet, these discussions are hampered by a lack of evidence on the interventions that might increase learning among primary school children, and whether evidence in one country can be fruitfully generalized to another.

A first generation of school-based experiments, mostly in Kenya, found that delivering textbooks and instructional materials to classrooms produced no achievement gains (Glewwe et al., 2004; Glewwe, Kremer, & Moulin, 2009; Kremer, Moulin, & Namunyu, 2003). Other experiments found zero or negligible effects of monetary grants (Blimpo & Evans, 2011; Das et al., 2013; Pradhan et al., 2011). Yet, a growing body of studies suggests that deliberate efforts to improve the quality of instruction—usually by combining materials and teacher training—can yield larger effect sizes of at least 20 percent of a standard deviation (Banerjee et al., 2007; Chay, McEwan, & Urquiola, 2005; Friedman, Gerard, & Ralaingita, 2010; He, Linden, & MacLeod, 2008, 2009; Jamison et al., 1981; Piper & Korda, 2011). We will review this literature in greater detail in the next section.

Journal of Policy Analysis and Management, Vol. 00, No. 00, 1–27 (2014)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21782

This paper presents complementary results from two randomized experiments conducted simultaneously in Kenya and Uganda.[1] In each country, clusters of primary schools in poor districts were randomly assigned to receive the Reading to Learn (RTL) intervention, implemented by the Aga Khan Foundation (AKF), in the official languages of reading instruction in the early primary grades (Lango in Uganda and Swahili in Kenya). RTL is a five-step *scaffolding* approach to literacy instruction, building from a conceptual understanding of stories, to the decoding of letter-sound relationships, to the eventual written production of new sentences and stories. AKF trained early-grade teachers, head teachers (i.e., principals), and school-management committees. They also provided classrooms with literacy materials in the official languages of instruction and English, and conducted ongoing mentoring of teachers. Students in treatment and control schools were given three examinations prior to the intervention, and at a predetermined end line. The numeracy and written literacy exams were written exams that assessed students' written responses to either written or spoken stimulation. The oral literacy exam assessed students' oral language capacity through a one-on-one interaction with the enumerator.

In Uganda, we find that the treatment effect of RTL on early-grade students is 0.18 standard deviations for written literacy and 0.20 for oral literacy. In contrast, the Kenyan effect on oral literacy is 0.08 standard deviations, with no effect on written literacy. Not unexpectedly, given the main focus of the intervention, there is no effect on the numeracy assessment in either country. The contrasting literacy findings are puzzling, since treated schools in both countries shared the same implementer, a nominally similar treatment, and common literacy assessments (except for the language). We assess several explanations for the smaller Kenyan effects.

One explanation is that Kenyan students already had higher test scores at baseline, on average, and their classrooms had higher initial levels of classroom inputs such as textbooks. The intervention might have been more successful for students starting with a lower baseline level of achievement or fewer available classroom resources. We do not find evidence consistent with these explanations, given the insignificance in both countries of interaction effects between the treatment indicator and baseline measures. We also find no evidence that Kenyan treatment schools received fewer classrooms inputs during the treatment, or that they implemented the instructional methodology with less fidelity. The most plausible explanation—but one that cannot be empirically tested—is that smaller effects result from heterogeneous exposure to the RTL treatment in the tested languages of Lango and Swahili. In Uganda, classroom teachers employed Lango in daily instruction. Despite the official policy, Kenyan teachers often used English. Swahili instruction occurred for as little as 30 minutes per day, insufficient to allow the entire scaffolding approach to be applied.

The paper makes two contributions to the growing literature on school improvement in developing countries. First, it adds to the mounting evidence that a coherent instructional model, aligned with materials and teacher training, can improve learning in poor primary schools. This is particularly germane given the sparse Ugandan evidence, and the many zero effects in earlier Kenyan experiments that focused on the delivery of instructional materials. It echoes the earlier optimism of Chay, McEwan, and Urquiola (2005) and Banerjee et al. (2007), who found that well-designed and targeted instructional interventions can increase student learning in poor settings.

Second, the parallel experiments allow us to consider the external validity of treatment effects, or whether causal relationships "[hold] over variations in persons,

[1] For background on the experiment, see the official evaluation report of the African Population and Health Research Center (APHRC; Oketch et al., 2012).

settings, treatment variables, and measurement variables" (Shadish, Cook, & Campbell, 2002, p. 507). We rule out variation in the treatment and its implementation as the principal explanation for different treatment effects. In contrast, other multisite field experiments have shown that heterogeneity in implementers—including their efficiency and other attributes—explains treatment effect heterogeneity across sites (Allcott & Mullainathan, 2012; Bold et al., 2013). We also find little evidence to support the idea that persons or classroom settings—though poorer in Uganda—can explain cross-country differences in effects.

The next section describes the literature on school improvement in more detail, focusing on recent studies with high internal validity, as well as the RTL intervention. The following section reviews the sample and evaluation design, as well as the data. The subsequent section assesses the internal validity of the experiment, including baseline balance in student and school variables, and the role of student attrition between baseline and follow-up. The penultimate section presents the main results within each country, and then considers various explanations for the relatively smaller Ugandan results. Finally, the last section summarizes and concludes.

IMPROVING PRIMARY SCHOOL LEARNING

Prior Research

A decade ago, Glewwe (2002) noted the scarcity of credible estimates of the causal links between school-based interventions in developing countries and student learning. Since then, the number of experiments with a learning outcome has grown quickly. They are broadly divided among evaluations of education inputs, health inputs, and incentive and accountability schemes (Kremer, Brannen, & Glennerster, 2013; McEwan, 2013). This paper focuses on the first category of interventions.[2]

Several randomized experiments in Kenya find that providing additional education inputs like textbooks and flipcharts did not improve learning, as measured by government exams (Glewwe et al., 2004; Glewwe, Kremer, & Moulin, 2009; Kremer, Moulin, & Namunyu, 2003). School libraries in India and the Philippines, even when accompanied by some staff training, led to no or small (0.13) effect sizes on language test scores, respectively (Abeberese, Kumler, & Linden, in press; Borkum, He, & Linden, 2012). Monetary grants to schools, unaccompanied by other interventions, had statistically insignificant effect sizes on achievement in Gambia and Indonesia, and less than 0.1 in India (Blimpo & Evans, 2011; Das et al., 2013; Pradhan et al., 2011). Large reductions in early-grade class size—from 82 to 44, on average—had no effect on test scores, unless classes were taught by contract employees rather than civil-servant teachers (Duflo, Dupas, & Kremer, 2012). These studies suggest that simply providing instructional inputs or unrestricted grants does little to improve student achievement.

More complex interventions, focusing on improving the quality of classroom instruction in language or mathematics, have delivered stronger results. An early experiment in Nicaragua found that math textbooks increased math scores by

[2] The second category includes the treatment of intestinal helminths (Baird et al., 2012; Miguel & Kremer, 2004), and the school-based provision of food and micronutrients (Adelman et al., 2008; Vermeersch & Kremer, 2004). The third category includes interventions to increase the quantity and quality of school performance information (Barr et al., 2012); to link teacher pay to performance measures (Glewwe, Nauman, & Kremer, 2010; Muralidharan & Sundararaman, 2011; Podgursky & Springer, 2007); to encourage school-based management and parent participation (Duflo, Dupas, & Kremer, 2012); and to hire teachers with flexible job contracts (Bold et al., 2013; Duflo, Dupas, & Kremer, 2012; Muralidharan & Sundararaman, 2012).

0.4 standard deviations when teachers received training aligned with the math curriculum (Jamison et al., 1981). In Chile and India, remedial tutoring of low-achieving students by trained classroom aides—rather than public school teachers—improved test scores in treated schools by at least 0.2 standard deviations (Banerjee et al., 2007; Chay, McEwan, & Urquiola, 2005). An Indian NGO that delivered scripted English lessons via flashcards and an interactive tutoring machine found effects of at least 0.25 standard deviations, with or without the machine-based instruction, and regardless of whether the lessons were implemented by regular teachers or NGO-hired staff (He, Linden, & MacLeod, 2008). The same NGO provided an early-literacy program in local languages to preschool and first-grade children, finding effect sizes on a literacy assessment of 0.12 to 0.7 (higher when supplementing rather than replacing existing instruction; He, Linden, & MacLeod, 2009). In Mali and Liberia, public school teachers received structured lesson plans and in-service training in separate models of literacy instruction (Friedman, Gerard, & Ralaingita, 2010; Piper & Korda, 2011). Using the same reading assessment, both experiments found effect sizes of at least 0.4. In summary, impact evaluations increasingly show that well-articulated approaches to improving instructional quality—using both materials and training—can improve student achievement in poor contexts.

The RTL Intervention

This paper evaluates a model for reading instruction implemented by AKF in grades 1 to 3 of poor public schools in Kenya and Uganda. The RTL intervention was developed in Australia more than a decade ago, with the goal of improving reading skills among those behind grade level, especially the Aboriginal population.[3] RTL emphasizes a five-step scaffolding approach to reading instruction (Aga Khan Foundation East Africa, 2013). First, the teacher prepares students to read by providing background knowledge of a story. Second, she reads the story aloud, until the children understand the story and can repeat its words. Third, she writes each sentence of the story on paper, and points to the words as she reads them. Fourth, the teacher and children cut up sentences into component words, and practice letter-sound relationships and spelling. Fifth, children practice rewriting the text, as well as the writing of new texts with their classmates.

AKF implemented the scaffolding approach via several activities (Aga Khan Foundation East Africa, 2013).[4] First, AKF and staff from the respective Ministries of Education co-facilitated training of lower primary teachers and head teachers in the use of the instructional method and local-language materials, management of large classes, and learning assessment. Second, the program endowed classrooms with reading and numeracy learning materials in the languages of instruction, including mini-libraries, book corners, and lockable storage facilities. Third, AKF explained RTL to School Management Committees (SMCs), and encouraged SMCs to prioritize lower primary literacy. Fourth, AKF staff in conjunction with AKF trained Ministry of Education personnel regularly monitored and mentored teachers during school visits and periodic meetings with teachers of geographically proximate schools.

[3] See http://www.readingtolearn.com.au/.

[4] The activities described in this paragraph constitute the RTL *core* model. In addition, half of the treatment schools in Kinango and all of the treatment schools in Amolatar received a *core plus* model that further included a parental involvement component. This component established mini-libraries in each community and encouraged parents to borrow books and read and tell stories to their children. Among the treatment schools, the core or core plus models were not randomized. Therefore, we are not able to separately identify the effects of the two approaches.

In both countries, the school year is between January and November. RTL started in October 2009, with teachers receiving 12 days of in-service training led by separate teams in Kenya and Uganda. Due to unforeseen delays, most schools did not receive the classroom mini-libraries until April 2010. As the intervention continued, AKF-trained teams visited treatment schools monthly to provide in-class mentoring support. Teachers were invited to quarterly meetings to share ideas with peers, observe model lessons, and receive refresher training from AKF staff. 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 locally trained.

Schools not selected for the treatment followed 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 Coast province, Dubeck, Jukes, and Okello (2012) found that typical literacy instruction emphasized word recognition, oral language, and choral repetition (but not reading) of sentences. In general, teachers were not comfortable using phonics-based instruction that emphasized letter-sound relationships, a feature of RTL's step 4.

EVALUATION DESIGN

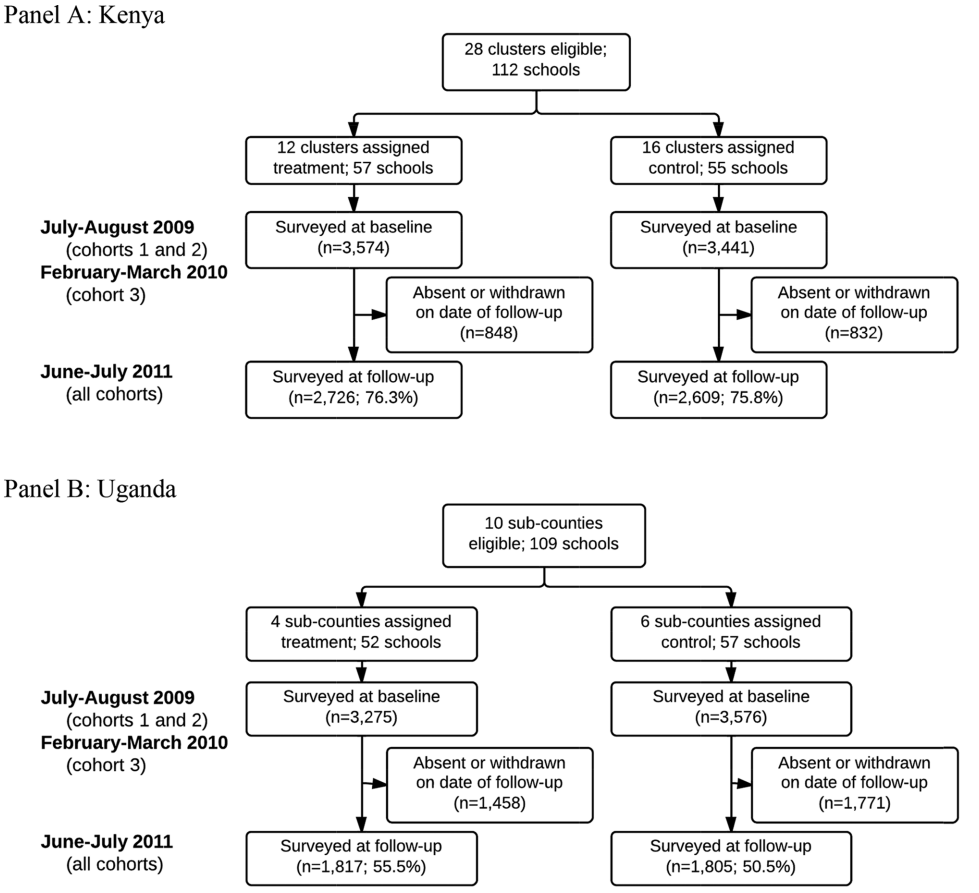
Sample and Randomization

In the Kenyan Coast province, the adjacent districts of Kwale and Kinango have among the lowest average scores on the primary school exit examinations and the highest poverty rates in the country (Dubeck, Jukes, & Okello, 2012). The Ministry of Education divides the schools in each district into geographically proximate clusters, with monitoring and support provided by the same officials to an entire cluster. In order to encourage official support and minimize cross-school contamination, the randomization occurred at the cluster level. The 28 clusters in the experimental sample each included one to eight schools, a total of 112 (see Figure 1, panel A).[5] The sample was divided into three strata: (1) Kwale school clusters, (2) Kinango clusters that participated in the Kenya School Improvement Project (KENSIP) intervention, and (3) non-KENSIP clusters in Kinango. (KENSIP is a separate school-improvement program implemented by AKF.) In Uganda, 10 clusters coincided with administrative subcounties and contained two to 16 schools each, a total of 109 schools (see Figure 1, panel B). They are located in two districts, Amolatar and Dokolo, with historically poor educational performance and a recent history of violence during the Lord's Resistance Army insurgency. The two Ugandan districts are also the strata.

Random assignment occurred at the cluster (or subcounty) level, within the five strata. In Kenya, 12 of 28 clusters were randomly assigned to the treatment group, while four of 10 Ugandan subcounties were assigned to the treatment group (see Figure 1, panels A and B). AKF treated all schools within treatment clusters. In addition, they treated two schools located in Ugandan control clusters. Nonetheless, we assign schools to their initial condition in subsequent analyses, so that our estimates can be interpreted as intention-to-treat effects.[6]

[5] Note that the official evaluation report includes 31 clusters in Kenya (Oketch et al., 2010). However, three of these were not randomized, but rather added to the treatment group ex post.

[6] In the school-level regression $TREATED_s = \beta_0 + \beta_1 ITT_s + \varepsilon_s$, where *TREATED* indicates whether school s was treated and *ITT* is the 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 an R-squared of 0.96.



Note: Cohort 1 students were in grade 2 in January 2009. Cohort 2 students were in grade 1 in January 2009. Cohort 3 students were in grade 1 in January 2010. The school year starts in January. In treated schools, the intervention started in October 2009 and continued during the 2010 and 2011 school years in grades 1, 2, and 3.

Figure 1. Experimental Samples in Kenya and Uganda.

Data Collection and Treatment Duration

APHRC administered exams and surveys in the treatment and control schools, independently of the implementing organization. The baseline data collection occurred in two phases, recalling that the Kenyan and Ugandan school year operate between January and November. The first phase was conducted in July and August 2009 and included questionnaires addressed to teachers and head teachers, visual classroom observations, and achievement exams for students in grades 1 and 2.[7] A second baseline occurred in February and March 2010 for newly enrolled grade 1 students.

[7] To reduce data collection costs, APHRC tested subsamples of students within each baseline cohort. If a school had multiple sections (or streams) in a given grade, a single section was randomly selected at the baseline. Within cohorts 1 and 2, 20 students per section were randomly sampled, stratifying by gender. If fewer than 20 students were in a section, all were selected. The same procedure was followed for cohort 3, although 25 students were randomly selected from each section.

Henceforth, we refer to the cohorts of second graders in 2009, first graders in 2009, and first graders in 2010 as cohorts 1, 2, and 3, respectively. All cohorts participated in follow-up data collection in June and July 2011.

The treatment began in October 2009, near the end of the school year, and continued during 2010 and 2011. Thus, cohort 1 was exposed to the treatment for part of 2009 and the entire 2010 school year (but not 2011 since RTL did not target grade 4). Cohort 2 was exposed for part of 2009, all of 2010, and approximately half of 2011. Finally, cohort 3 was exposed for all of 2010 and half of 2011.[8] We report estimates of treatment effects that pool the three cohorts, as well as separate estimates for each cohort.

Test Design

APHRC designed tests in numeracy, written literacy, and oral literacy in consultation with the implementing organization, national curriculum and assessment experts, and practitioners (Oketch et al., 2012). They developed test items in English, drawing on the primary school curriculum from Kenya and Uganda. The test forms were then translated into Swahili and Lango, the official early-grade languages of reading instruction in Kenya and Uganda, respectively. The Kenyan numeracy exam remained in English, following official policy on language of instruction in mathematics. After piloting, they compiled grade-specific test forms, such 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 students completed all items.[9]

At follow-up, baseline students in all cohorts were tested if they were present on the day of the survey.[10] Students completed baseline test items as well as new items specific to their current grade. The baseline and follow-up test forms varied depending on the grades in which they were applied. However, the use of repeated anchor items across forms facilitates the linking of tests to a common scale. Within each subject, we estimated a one-parameter (i.e., Rasch) item response theory (IRT) 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 (Wright & Stone, 1979). We then standardized the resulting logit scale in each exam by the mean and the standard deviation of the baseline exam. Hence, all subsequent effects—in Kenya and Uganda—can be interpreted as a proportion of the pooled standard deviation on the baseline test.

Estimation

The main regression specification in each country is

$$POST_{isj} = \alpha + \beta ITT_{sj} + \sum_{e=1}^{3} \gamma_e PRE_{e,isj} + X'_{isj}\gamma + \delta_j + \varepsilon_{isj} \quad (1)$$

[8] Cohort 3 was potentially subject to up to two months of the RTL treatment prior to the baseline survey.
[9] The numeracy exams were administered first, followed by the written literacy exam. The final exam was the oral literacy exam that involved one-on-one interaction between an enumerator and a student.
[10] Any absent baseline students were replaced by another randomly selected student of the same gender in their expected grade (e.g., absent students from the 2010 grade 1 baseline were replaced with students in grade 2 in 2011). A subsequent section describes 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 the following section includes the replacement students.

Table 1. Baseline student characteristics.

	Control (1)	Treatment (2)	Difference (3)
Numeracy score			
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			
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			
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: Total student baseline sample: 13,931. Kenya: 3,574 treatment and 3,441 control. Uganda: 3,275 treatment and 3,576 control. Columns 1 and 2: standard deviations appear in parenthesis. Column 3: standard errors clustered at the unit of randomization appear in parentheses.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.
Critical values adjusted for the number of clusters.

where the dependent variable is the score on a posttest—either written literacy, oral literacy, or numeracy—administered to student i enrolled in school s in stratum j. The variable ITT indicates initial assignment to the treatment group; the variables PRE_e are scores on pretests administered at the baseline in each subject e; the X are a vector of controls including dummy variables indicating students' gender and cohort; the δ_j are dummy variables indicating experimental strata; and ε_{isj} is the idiosyncratic error assumed to be independent between school clusters, but allowed to be correlated within them. β is the treatment effect on student test scores. Given randomized assignment, it is not strictly necessary to control for baseline scores, though we always do so to improve the precision of estimated treatment effects, and to adjust for imbalance in the baseline scores across treatment and control groups.

INTERNAL VALIDITY

Baseline Balance

Table 1 compares the baseline test scores of students in treatment and control groups. Students in control schools tended to have higher scores on all three tests. We find a statistically significant difference between the average treatment and control scores on the numeracy exam in the Kenyan sample, favoring control schools. Other differences are statistically insignificant. The immediate implication is that it is preferable to control for baseline test scores. Given the common test scale, we also conclude that students in Kenya had 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 scored 1.3 standard deviations higher, on average. Given that these differences exist even for the youngest cohorts

Table 2. Baseline classroom and teacher characteristics.

	Kenya			Uganda		
	Control (1)	Treatment (2)	Difference (3)	Control (4)	Treatment (5)	Difference (6)
Panel A: Classroom attributes						
Lesson plans or notes	0.67 (0.47)	0.68 (0.47)	−0.01 (0.08)	0.84 (0.37)	0.80 (0.40)	0.04 (0.07)
Recommended textbooks	0.71 (0.46)	0.77 (0.42)	−0.06 (0.08)	0.38 (0.49)	0.29 (0.45)	0.09 (0.05)
Other books or reading materials	0.40 (0.49)	0.55 (0.50)	−0.15** (0.07)	0.37 (0.48)	0.20 (0.40)	0.17** (0.07)
Student-made materials	0.34 (0.48)	0.34 (0.48)	0.00 (0.07)	0.10 (0.31)	0.09 (0.29)	0.01 (0.04)
Notebooks	0.79 (0.41)	0.90 (0.31)	−0.11* (0.06)	0.59 (0.49)	0.62 (0.49)	−0.03 (0.10)
Wall charts or visual teaching aids	0.71 (0.46)	0.77 (0.42)	−0.06 (0.09)	0.66 (0.47)	0.45 (0.50)	0.21* (0.11)
Chalkboard, eraser, and chalk	0.83 (0.38)	0.92 (0.27)	−0.09 (0.06)	0.86 (0.34)	0.86 (0.35)	0.01 (0.07)
Panel B: Teacher characteristics						
Education of at least a high school diploma	0.94 (0.24)	0.95 (0.22)	−0.01 (0.03)	0.90 (0.30)	0.97 (0.16)	−0.07** (0.03)
Lower primary preservice teacher training	0.23 (0.42)	0.24 (0.43)	−0.01 (0.05)	0.69 (0.46)	0.68 (0.47)	0.01 (0.03)
Number of years of teaching	12.60 (9.82)	12.02 (9.74)	0.58 (1.43)	13.08 (6.85)	11.95 (8.08)	1.13 (0.87)
Number of years teaching current subject	10.36 (9.07)	9.67 (9.22)	0.69 (1.11)	7.89 (6.20)	8.16 (7.38)	−0.27 (0.89)
Adequately prepared to teach the subject curriculum	0.79 (0.41)	0.81 (0.39)	−0.02 (0.05)	0.57 (0.50)	0.56 (0.50)	0.01 (0.06)

Notes: Panel A: classrooms are all grade 1 and 2 classrooms surveyed in the initial baseline. Presence of attribute coded as 1. Sample sizes: 423 Kenya, 365 Uganda. Panel B: presence of characteristic coded as 1. Sample sizes: 592 Kenya, 332 Uganda. Columns 1, 2, 4, and 5: standard deviations appear in parenthesis. Columns 3 and 6: standard errors clustered at the unit of randomization appear in parentheses.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.

who were tested in first grade, the cross-country differences cannot be explained by differential exposure to the quantity or quality of primary school instruction.[11]

During the baseline, enumerators visually inspected classrooms that served grades 1 and 2, noting the presence of classroom attributes. Table 2 (panel A) compares the visibility of these various elements across the treatment and the control group. Some

[11] The differences must instead be explained by differential exposure to family, peer, and school inputs from birth to baseline; unfortunately we have little data that might be used to credibly explain cross-country differences.

Table 3. Attrition at follow-up.

	Not present at follow-up					
	Kenya			Uganda		
	(1)	(2)	(3)	(4)	(5)	(6)
ITT	−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)
ITT × numeracy score			−0.033			0.010
			(0.021)			(0.019)
ITT × written literacy score			−0.009			−0.001
			(0.020)			(0.012)
ITT × 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	7,040	7,040	7,040	6,891	6,891	6,891
R-squared	0.001	0.01	0.01	0.01	0.02	0.02
Proportion of baseline absent		0.24			0.47	

Notes: Sample includes all students who completed at least one exam at baseline. All columns include strata dummy variables as controls. Standard errors clustered at the unit of randomization appear in parentheses.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.
Critical values adjusted for number of clusters.

imbalance is evident, and we later assess whether treatment effects are sensitive to controls for these variables. Further, Kenyan classrooms were better endowed with some inputs than Ugandan classrooms.

Teachers were also surveyed at the baseline (see panel B). Within each country, they were generally similar in their levels of experience and preservice training across treatment and control schools, although 90 percent of Ugandan control group teachers had at least a high school diploma, compared with 97 percent in the treatment group. As an indirect measure of teacher quality, Kenyan teachers were about 20 percentage points more likely than Ugandan teachers to state that they felt adequately prepared to the teach the subject.

Attrition

Twenty-four and 47 percent of the Kenyan and Ugandan baseline samples, respectively, were not present at the follow-up testing (see Table 3).[12] Because of data limitations, our measure of attrition is the sum of both temporary absenteeism on

[12] Attrition may be due to absence or dropout. Based on classroom rosters examined at the follow-up, the estimated absenteeism—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.

the day of testing and permanent dropout from the school. We test for differential attrition by treatment status for two reasons. First, an important effect of the treatment might have been to reduce the likelihood that students were absent or dropped out of school. Second, differential attrition could introduce bias in our achievement results. Table 3 reports whether attrition rates differed across treatment and control groups, and whether they were 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. Differential attrition did not occur in Kenya, but students in the Ugandan treatment group were 5 percentage points less likely to attrit (statistically significant at 10 percent).

Columns 2 and 5 further control for baseline test scores, and the evidence of differential test scores across attriting and nonattriting students is mixed. Even if attriting students were, on average, low achieving, we are mainly concerned about potential imbalance in baseline attributes across attriting students 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, nor are they jointly significant. Additionally, the variables included explain very little of the variation in the probability of attrition as demonstrated by the low R^2 across all models. The treatment may have slightly reduced the likelihood of being absent on the day of the follow-up exam in Uganda, but this effect does not appear to be differential by baseline test score. Our subsequent estimates always control for baseline test scores, and we further conduct a bounding exercise in which the higher attrition rate in each country's control group is applied to the same country's treatment group in order to bound the estimates, a methodology similar to Lee (2009).

RESULTS

Treatment Effects in Kenya and Uganda

Table 4 presents estimates of equation (1) for each country and test. The sample in each column includes students who took the indicated test and at least one baseline test.[13] The degrees of freedom for the critical values in all tables have been adjusted following Cameron, Gelbach, and Miller (2008). In Kenya (columns 1 to 3) we find no statistically significant effect of the program on numeracy or written literacy scores. However, the program increased oral literacy scores by 0.077, or 8 percent of the baseline standard deviation, and we reject that this coefficient is equal to zero at the 10-percent level. In contrast, for Uganda we find in columns 4 to 6 that the treatment increased written literacy scores by 0.2 standard deviations and oral literacy by 0.18.[14] RTL is primarily a literacy intervention, and so the lack of an effect in both countries for numeracy is not unexpected. However, RTL methods could be

[13] More than 96 percent of students who were tested in both the baseline and the follow-up completed all three exams in the baseline. When a student did not take a baseline exam, we recode the missing value to zero, and include a dummy variable indicating observations with missing values (Puma et al., 2009). The findings are almost identical if we simply replace a student's missing scores with the average of his or her non-missing scores. Table 10 (column 4) reports additional estimates that limit the sample to students with scores for all three baseline exams.

[14] These estimates rely on the intention to treat randomization. Because two schools randomized into the control status on Uganda were subsequently treated, we can instrument for *TREATED* with *ITT* in an IV framework. The resulting estimates are similar to the results in Table 4. The coefficients on *TREATED* in the second stage are 0.125 for numeracy, 0.208 for written literacy, and 0.196 for oral literacy. As with the results in Table 4, the numeracy result is not statistically different from zero, while the second two are statistically different from zero at 1 percent.

Table 4. Effect of treatment on achievement.

	Kenya			Uganda		
	Numeracy (1)	Written literacy (2)	Oral literacy (3)	Numeracy (4)	Written literacy (5)	Oral literacy (6)
ITT	−0.011	0.024	0.077*	0.117	0.199***	0.179***
	(0.059)	(0.032)	(0.042)	(0.087)	(0.054)	(0.047)
				[0.246]	[0.014]	[0.010]
Baseline numeracy score	0.308***	0.103***	0.122***	0.254***	0.218***	0.139***
	(0.021)	(0.023)	(0.026)	(0.014)	(0.018)	(0.020)
Baseline written literacy score	0.181***	0.150***	0.233***	0.106***	0.147***	0.125***
	(0.022)	(0.021)	(0.024)	(0.025)	(0.034)	(0.034)
Baseline oral literacy score	0.473***	0.548***	0.505***	0.228***	0.340***	0.266***
	(0.030)	(0.027)	(0.032)	(0.020)	(0.033)	(0.023)
Observations	5,323	5,302	5,305	3,604	3,596	3,575
R-squared	0.58	0.55	0.52	0.40	0.42	0.38
Average control group change	1.02	0.81	1.23	0.50	0.78	0.78

Notes: Samples include students who completed the specified follow-up test and at least one baseline test. All regressions include gender, cohort, and strata dummy variables, as well as dummy variables indicating missing values of each baseline test score. Standard errors clustered at the unit of randomization appear in parentheses.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.
Critical values adjusted for number of clusters. Columns 4 to 6: *P*-values associated with wild cluster bootstrap standard errors appear in brackets.

incorporated in mathematics instruction, and an increase in literacy might have had a spillover effect on student performance on a numeracy assessment. The magnitude of the numeracy point estimate is smaller (0.12) and statistically insignificant. Even so, the estimate is quite imprecise, which prevents strong conclusions.

Given the small number of clusters in Uganda, the typical formula for cluster-robust standard errors could result in inappropriately small standard errors. Therefore, we calculate *P*-values associated with the wild cluster bootstrap-T method described 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 5 percent in both cases, and that the numeracy results remain statistically insignificant.

Table 5 (panel A) reports treatment effects by the three cohorts in each country, since each cohort was treated in different grades and for a different duration.[15] We create three new treatment variables that are the interaction of *ITT* and each cohort dummy variable. The coefficients vary in their sign and statistical significance, especially in Kenya. But, in all cases we fail to reject the null hypothesis that the coefficients are equal. For the sake of parsimony and statistical power, we pool cohorts for the remainder of the estimates.

Table 5 (panel B) further assesses whether RTL treatment effects vary by gender. In Kenya, boys performed worse than girls on follow-up tests, but there is no evidence that RTL was relatively more effective in increasing the test scores of either gender (indicated by the small and statistically insignificant coefficients

[15] Unfortunately the duration of treatment and cohort are perfectly correlated, and we cannot separately identify the importance of duration versus the grade at the start of the treatment.

Table 5. Heterogeneous effect of treatment by cohort and gender.

	Kenya			Uganda		
	Numeracy (1)	Written literacy (2)	Oral literacy (3)	Numeracy (4)	Written literacy (5)	Oral literacy (6)
Panel A: Heterogeneous effect of treatment by cohort						
ITT × cohort 1	0.125	0.091*	0.135**	0.065	0.146*	0.179**
	(0.082)	(0.052)	(0.065)	(0.114)	(0.065)	(0.070)
ITT × cohort 2	−0.037	−0.008	0.080	0.162	0.165*	0.169**
	(0.079)	(0.054)	(0.071)	(0.092)	(0.076)	(0.053)
ITT × cohort 3	−0.099	−0.003	0.028	0.125	0.276**	0.187**
	(0.080)	(0.068)	(0.075)	(0.096)	(0.091)	(0.061)
F-test of equality of interaction coefficients						
F-statistic	2.12	0.85	0.38	0.86	0.91	0.03
P-value	0.14	0.44	0.69	0.45	0.44	0.97
Observations	5,323	5,302	5,305	3,604	3,596	3,575
R-squared	0.58	0.55	0.52	0.40	0.42	0.38
Panel B: Heterogeneous effect of treatment by gender						
ITT	−0.019	0.033	0.070	0.100	0.223***	0.218***
	(0.061)	(0.038)	(0.045)	(0.087)	(0.053)	(0.048)
ITT × male	0.017	−0.017	0.014	0.035	−0.048	−0.080*
	(0.033)	(0.031)	(0.046)	(0.041)	(0.040)	(0.043)
Male	−0.047**	−0.088***	−0.098***	0.091**	0.065*	0.065*
	(0.021)	(0.019)	(0.030)	(0.032)	(0.031)	(0.031)
F-test for joint significance of ITT coefficients						
F-statistic	0.15	0.39	1.65	1.18	9.22	10.76
P-value	0.86	0.68	0.21	0.35	0.01	0.00
Observations	5,323	5,302	5,305	3,604	3,596	3,575
R-squared	0.58	0.55	0.52	0.40	0.42	0.38

Notes: Cohort 1: grade 2 in 2009. Cohort 2: grade 1 in 2009. Cohort 3: grade 1 in 2010. Sample includes students who completed the specified follow-up test and at least one baseline test. All regressions include gender, cohort, and strata dummy variables, as well as dummy variables indicating missing values of each baseline test score. Standard errors clustered at the unit of randomization appear in parentheses.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.
Critical values adjusted for number of clusters.

on the interaction terms in columns 1 to 3).[16] In Uganda, boys performed better than girls on follow-up assessments (by less than 0.1 standard deviations). There is promising but weak evidence that RTL helps overcome this gender gap. RTL was more effective among girls, but only on the oral literacy assessment. The implied effects for girls and boys are 0.22 and 0.14, respectively, although the coefficient on the interaction term is only statistically significant at 10 percent.

Why Do Effects Vary Across Countries?

The treatments effects are larger in Uganda, despite the same implementing organization, a nominally similar treatment, and the same literacy assessments, except for the language. This section uses available data to test hypotheses on the sources of these differential effects. We empirically assess two explanations: heterogeneity in

[16] Earlier research in Kenya showed that girls had lower achievement than boys, by 0.25 standard deviations on the primary school leaving examination (Lucas & Mbiti, 2012b).

baseline attributes of students and classrooms across countries, and heterogeneity in implementation fidelity. We also discuss a third plausible explanation that cannot be empirically tested: heterogeneity in exposure to RTL in the tested languages.[17]

Ugandan students had much lower baseline test scores, on average, suggesting a wider scope for learning gains on the early-grade literacy skills emphasized in RTL. Thus, Table 6 (panel A) tests whether Kenyan treatment effects are larger among lower achieving Kenyan students (or whether they are smaller among higher achieving Ugandan students). In Kenya, we actually find that students with *higher* baseline scores have larger treatment effects on numeracy, and no evidence of interactions across other outcomes or in Uganda.[18]

One might be concerned that the Kenyan test score distribution is shifted so far to the right that it contains few low-scoring students (for whom the treatment might have been effective). Thus, we repeated the analysis in panel A of Table 6 after limiting the sample in each country to those students whose scores fell in a region of common support. First, we limited each country's sample to those students whose scores fell between the smaller of the two countries' maxima and the larger of the two countries' minima. Second, we used the first and 99th percentiles instead of the minima and maxima. In results not presented here, we found substantively similar results to those in Table 6.

A related hypothesis is that RTL was more appropriate for classrooms in which students had similar test scores, perhaps because teachers were better able to apply the instructional method in more homogeneous classrooms. Panel B of Table 6 tests for differential effects by within-classroom heterogeneity, as proxied by the standard deviation of the subject-specific baseline test scores in a classroom. Across all countries and exams, we do not find any statistically significant evidence of interactions with *ITT*.

Ugandan classrooms also had lower levels of some classroom and teacher attributes. A natural hypothesis is that the resources and teacher training provided by RTL are particularly helpful in underresourced settings. Table 6 (panel C) tests for heterogeneity by the baseline classroom attributes reported in Table 2. For ease of interpretation, we calculate the proportion of attributes visible in a classroom. The interactions between this index and *ITT* are insignificant.[19] In other results not reported here, we separately interacted each classroom attribute measure with *ITT*, finding only two marginally significant coefficients with signs that are inconsistent with larger effects among lower resourced classrooms. To test for heterogeneity

[17] We cannot empirically reject other potential hypotheses for the different magnitudes of the treatment effects in each country (perhaps relating to cross-country differences in culture and language use). For these to be plausible explanations, we emphasize that the differences cannot merely introduce gaps in the baseline assessments; rather, they must interact with the RTL treatment, thereby influencing magnitude of country-specific treatment effects that control for baseline test scores.

[18] As a further test of heterogeneity by baseline test score, we interacted the treatment indicator with dummy variables indicating students' within-country baseline score quartiles. In five of the six cases, we failed to reject the equality of the interaction coefficients (results not shown). We reject the equality of the interaction coefficients for numeracy in Kenya, where the coefficient on the interaction between students with the lowest quartile of test scores and the treatment is negative and statistically significant. This is consistent with the finding in Table 6 (panel A, column 1), and the opposite of what would be expected if baseline student score differences were driving the heterogeneous effects across countries.

[19] The point estimates for the coefficients on the interaction term represent the expected score change when comparing a classroom with no attributes (index value of 0) to one with all attributes (index value of 1). These coefficients are large in absolute magnitude for Uganda (columns 4 to 6), yet based on the standard deviation presented in the table, a one standard deviation change in the attribute index would only differentially change expected scores by 0.045 standard deviations in numeracy and written literacy and 0.013 standard deviations in oral literacy. Even moving between the 10th and 90th percentile values of the index would only differentially change the expected score by 0.12 for numeracy and written literacy and 0.03 for oral literacy.

Table 6. Heterogeneous effect of treatment by baseline student scores and classroom attribute.

	Kenya			Uganda		
	Numeracy (1)	Written literacy (2)	Oral literacy (3)	Numeracy (4)	Written literacy (5)	Oral literacy (6)
Panel A: Heterogeneous effect of treatment by baseline student scores						
ITT	−0.048	−0.003	0.077	0.119	0.152***	0.185***
	(0.064)	(0.051)	(0.046)	(0.085)	(0.036)	(0.045)
ITT × baseline subject score	0.104**	0.044	0.008	−0.002	−0.082	−0.015
	(0.040)	(0.040)	(0.038)	(0.029)	(0.046)	(0.050)
Baseline subject score	0.255***	0.149***	0.503***	0.254***	0.145***	0.269***
	(0.021)	(0.021)	(0.032)	(0.018)	(0.035)	(0.027)
F-test for joint significance of ITT coefficients						
F-statistic	3.58	1.96	2.02	0.99	8.93	8.71
P-value	0.04	0.16	0.15	0.41	0.01	0.01
Observations	5,305	5,286	5,285	3,585	3,572	3,495
R-squared	0.58	0.55	0.52	0.40	0.42	0.38
Panel B: Heterogeneous effect of treatment by baseline classroom heterogeneity						
ITT	0.104	0.018	−0.026	−0.008	0.233	0.243**
	(0.136)	(0.149)	(0.145)	(0.153)	(0.155)	(0.087)
ITT × baseline classroom score standard deviation	−0.190	0.014	0.158	0.160	−0.058	−0.115
	(0.175)	(0.259)	(0.186)	(0.152)	(0.228)	(0.171)
Baseline classroom score standard deviation	0.283**	0.313**	0.010	−0.265**	0.310	0.047
	(0.124)	(0.119)	(0.131)	(0.085)	(0.210)	(0.130)
F-test for joint significance of ITT coefficients						
F-statistic	0.73	0.39	3.03	1.77	14.71	10.07
P-value	0.49	0.68	0.07	0.23	0.00	0.01
Observations	5,323	5,302	5,305	3,604	3,596	3,575
R-squared	0.58	0.55	0.52	0.40	0.42	0.38
Average classroom score standard deviation	0.62	0.55	0.66	0.77	0.48	0.57
Panel C: Heterogeneous effect of treatment by baseline classroom attributes						
ITT	−0.150	0.013	0.071	0.286*	0.364*	0.227
	(0.240)	(0.151)	(0.162)	(0.146)	(0.183)	(0.156)
ITT × baseline classroom attribute index	0.193	−0.004	−0.018	−0.324	−0.325	−0.091
	(0.312)	(0.204)	(0.223)	(0.305)	(0.336)	(0.298)
Baseline classroom attribute index	0.042	0.216*	0.278*	0.429**	0.270	0.156
	(0.229)	(0.119)	(0.154)	(0.185)	(0.190)	(0.189)
F-test for joint significance of ITT coefficients						
F-statistic	0.20	0.06	1.06	2.50	6.94	7.78
P-value	0.82	0.94	0.36	0.14	0.02	0.01
Observations	5,323	5,302	5,305	3,604	3,596	3,575
R-squared	0.58	0.55	0.52	0.40	0.42	0.38
Average index value		0.67			0.49	
Average index standard deviation		0.18			0.14	

Notes: Sample includes students who completed the specified follow-up test and at least one baseline test. All regressions include gender, cohort, and strata dummy variables, as well as dummy variables indicating missing values of each baseline test score. Standard errors clustered at the unit of randomization appear in parentheses. Panel C: classroom attribute index measured from 0 to 1 as the proportion of the seven classroom attributes in Table 2 visible at baseline.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.
Critical values adjusted for number of clusters.

by baseline teaching characteristics (in results not presented here), we used three self-reported teacher characteristics: whether the teacher had received lower primary preservice training, the number of years of teaching, and whether the teacher felt adequately prepared to teach the subject curriculum. We assigned each student the average score across all lower primary teachers within a school. We included all three measures and their interactions with the treatment indicator in a single regression for each subject exam. We only found a heterogeneous effect on oral literacy in Uganda for the portion of teachers who received lower primary training: the higher the portion of teachers with this training, the smaller the score gains among the treated students. Additionally, we failed to reject the joint insignificance of the three teacher measures, with the exception of oral literacy in Uganda.

We next examine the role of heterogeneity in implementation fidelity across countries, first considering the issue of whether the treatment actually increased the amount of classroom inputs over baseline levels. Table 7 reports a variant of equation (1) in a sample of second-grade classrooms, the only grade that was visually assessed at both baseline and follow-up. The dependent variable is equal to 1 if a classroom attribute was visible at the follow-up. Control variables include dummy variables indicating the baseline values of all seven classroom inputs. Consistent with program goals (and with the presence of such attributes being endogenous to the program), we find that the likelihood of observing other reading materials, student-made materials, and wall charts and visual aids increased across both countries. There were also increases in both countries in recommended textbooks. AKF reported providing some textbooks in Uganda, and it is also possible that existing textbooks were simply distributed and used more frequently, given AKF-supplied lockable storage in classrooms. The increase in notebooks and chalkboards in Uganda, inputs not supplied by RTL, is more puzzling. However, they could have been provided by the treatment schools themselves because of decreased expenditures on other items and a new emphasis on lower primary grades.

To assess whether differential changes in these attributes can explain RTL treatment effects, we controlled for changes in classroom inputs between baseline and follow-up in the reestimation of equation (1).[20] 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 as before (0.07), 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 of 0.16 to 0.17, or at least 85 percent the size of the original coefficients. Therefore, differential changes in classroom inputs do not appear to be the primary cause of the differences between the treatment effects across the countries.

Even though the implementation was designed to be uniform, schools and teachers varied in how faithfully they applied the RTL instructional approach. AKF conducted an implementation assessment in each country, based on a monitoring checklist of student, teacher, and head teacher activities (Table 8), prior to knowing any results from APHRC data collection. The implementation scores across the two countries are approximately the same with 25 percent of schools with *high* scores of 7 to 11, 50 percent with *medium* scores of 5 or 6, and 25 percent with *low* scores of 4 of less. Assuming that implementation fidelity was graded similarly across countries—as AKF intended—differences in implementation do not appear to explain cross-country differences in treatment effects.

[20] The change in each classroom characteristic was measured as −1 (item was present in baseline but not follow-up), 0 (no change in presence or absence of item), or 1 (item was absent at baseline and present at follow-up).

Table 7. Effect of treatment on classroom attributes.

	Lesson plans or notes (1)	Recommended textbooks (2)	Other books or reading materials (3)	Student-made materials (4)	Notebooks (5)	Wall charts or visual teaching aids (6)	Chalkboard, eraser, and chalk (7)
Panel A: Kenya							
ITT	0.036	0.197*	0.414***	0.196*	0.063	0.243***	0.044
	(0.073)	(0.107)	(0.067)	(0.097)	(0.056)	(0.081)	(0.055)
Observations	192	192	192	192	192	192	192
R-squared	0.09	0.17	0.25	0.17	0.08	0.30	0.12
Baseline average	0.67	0.74	0.47	0.34	0.84	0.74	0.88
Panel B: Uganda							
ITT	0.016	0.315***	0.427***	0.183	0.148*	0.468***	0.159**
	(0.078)	(0.094)	(0.030)	(0.113)	(0.075)	(0.076)	(0.053)
Observations	167	167	167	167	167	167	167
R-squared	0.07	0.14	0.29	0.13	0.12	0.33	0.14
Baseline average	0.82	0.33	0.28	0.10	0.60	0.56	0.87

Notes: Sample includes second-grade classrooms from schools that were sampled in baseline and follow-up. Dependent variable equal to 1 if item was visible. All regressions include dummy variables for all baseline classroom characteristics and strata. Standard errors clustered at the unit of randomization appear in parentheses.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.
Critical values adjusted for number of clusters.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table 8. Implementation assessment.

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.

Notes: A school's implementation score was determined by the number of affirmative statements. Schools with scores 7 to 11 were considered high, scores of 5 to 6 medium, and 0 to 4 were low implementers relative to the ideal RTL model.

Table 9. Heterogeneity by the degree of implementation.

	Kenya			Uganda		
	Numeracy (1)	Written literacy (2)	Oral literacy (3)	Numeracy (4)	Written literacy (5)	Oral literacy (6)
ITT × high implementation	0.039	0.072**	0.123**	0.073	0.351***	0.267***
	(0.055)	(0.031)	(0.046)	(0.077)	(0.073)	(0.076)
ITT × medium implementation	−0.044	−0.015	0.051	0.176	0.216**	0.207**
	(0.063)	(0.031)	(0.052)	(0.121)	(0.081)	(0.076)
ITT × low implementation	−0.121	−0.059	0.002	0.148	0.121	0.145*
	(0.098)	(0.053)	(0.041)	(0.092)	(0.093)	(0.067)
Test for equality of implementation coefficients						
F-statistic	1.48	5.19	4.38	1.74	1.65	0.55
P-value	0.25	0.01	0.02	0.23	0.25	0.60
Observations	5,323	5,302	5,305	3,604	3,596	3,575
R-squared	0.51	0.52	0.49	0.37	0.37	0.37

Notes: Sample includes students who completed the specified follow-up test and at least one baseline test. All regressions include gender, cohort, and strata dummy variables, as well as dummy variables indicating missing values of each baseline test score. Standard errors clustered at the unit of randomization appear in parentheses.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.
Critical values adjusted for number of clusters.

Table 9 further tests whether schools judged to be better implementers had larger treatment effects by interacting *ITT* with three indicators of high, medium, or low degrees of implementation fidelity. In general, the effect of the treatment is monotonically decreasing in the quality of implementation. 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). These could reflect that high implementers are a nonrandom subset of schools in each country, and so

we consider these results merely suggestive.[21] For now, suppose that *grade inflation* led Kenyan raters to overstate the degree of implementation fidelity. It would have to be severe to completely explain the cross-country differences, since the literacy point estimates for Kenyan high implementers are smaller than even the Ugandan low implementers. In short, cross-country variation in implementation appears an unlikely explanation for the differential treatment effects.

A final explanation—but one that cannot be subjected to empirical tests—is that students in Kenya and Uganda received similar doses of the treatment overall, but that Kenyan students received proportionally less in the tested language. Formal school policy in both Kenya and Uganda specifies that the first three years of primary instruction should be taught in students' mother tongues (with the exception of numeracy in Kenya, which is taught in English). For this reason Ugandan literacy assessments were in Lango and Kenyan assessments in Swahili. In Uganda, instruction occurred in Lango. Moreover, one hour each day was allocated Lango instruction, a sufficient amount of time for all five steps of an RTL literacy lesson to occur, according to AKF.

Despite the official policy, the primary instructional language in Kenyan early grades was usually English (anecdotally, some treatment schools displayed posters that read "We Only Speak English Here"). This is consistent with the qualitative fieldwork of Dubeck, Jukes, and Okello (2012) in other Coast province schools, where some primary teachers punished the use of local languages, while others switched back and forth between English, Swahili, and a mix of other local languages. Swahili was an approximate mother tongue, but was typically taught as a second language for 30 minutes at the end of the day, and in some cases this time was used for additional English instruction.[22] This had two detrimental effects on the likelihood of detecting an RTL effect on Swahili literacy.

First, based on classroom observations, AKF believes that 30 minutes is not long enough for all five RTL steps to be completed, with the steps of practicing letter-sound relationships, spelling, and writing most likely to be omitted (N. Shekhova, personal communication, 2013). Second, AKF mentoring personnel typically observed the earliest classes in the day. In Kenya this was either numeracy or English. Therefore, RTL mentoring support during the school year was typically provided for English literacy. Teachers might have applied the suggestions to Swahili instruction, but the mismatch between the curricula in the two languages would have made this difficult. We conclude that relatively smaller treatment effects in Kenya plausibly result from a mismatch between the languages of instruction and assessment, although this cannot be empirically verified.

Robustness

Table 10 provides evidence that the main results are robust to alternative specifications. Each column contains the coefficient of interest from six separate regressions,

[21] Implementation fidelity is likely endogenous to observed or unobserved variables that would lead to higher test scores even in the absence of treatment. To partly assess this endogeneity, we tested for differences in baseline characteristics across the sample of treatment schools that are high, medium, and low implementers. We found 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. Even so, we find that the effects in Table 9 persist after controlling for baseline test scores and the other covariates just mentioned.

[22] Many Kenyan students spoke a mother tongue similar to Swahili, but not necessarily with a formally established orthography. The evaluation data do not include student-specific data on mother tongue.

Table 10. Robustness.

	Preferred specification (1)	Panel with student fixed effects (2)	Repeated cross-section (3)	Students with three baseline test scores (4)	Adjusted attrition	
					Lower bound (5)	Upper bound (6)
Panel A: Kenya						
Dependent variable: numeracy score						
ITT	−0.011	0.035	0.077	−0.007	−0.015	−0.005
	(0.059)	(0.079)	(0.077)	(0.059)	(0.059)	(0.059)
Observations	5,323	10,610	13,524	5,275	5,314	5,314
R-squared	0.58	0.54	0.37	0.58	0.58	0.58
Dependent variable: written literacy score						
ITT	0.024	0.033	0.049	0.026	0.022	0.032
	(0.032)	(0.031)	(0.035)	(0.032)	(0.033)	(0.032)
Observations	5,302	10,572	13,494	5,254	5,287	5,287
R-squared	0.55	0.63	0.41	0.55	0.55	0.55
Dependent variable: oral literacy score						
ITT	0.077*	0.044	0.066	0.080*	0.072	0.084*
	(0.042)	(0.053)	(0.053)	(0.042)	(0.042)	(0.042)
Observations	5,305	10,570	13,495	5,257	5,291	5,291
R-squared	0.52	0.68	0.49	0.52	0.52	0.52
Panel B: Uganda						
Dependent variable: numeracy score						
ITT	0.117	0.157	0.128	0.124	0.002	0.221**
	(0.087)	(0.139)	(0.139)	(0.085)	(0.077)	(0.084)
	[0.246]	[0.301]	[0.496]	[0.158]	[0.979]	[0.034]
Observations	3,604	7,170	13,751	3,494	3,433	3,433
R-squared	0.40	0.26	0.22	0.40	0.37	0.38

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table 10. Continued.

	Preferred specification (1)	Panel with student fixed effects (2)	Repeated cross-section (3)	Students with three baseline test scores (4)	Adjusted attrition	
					Lower bound (5)	Upper bound (6)
Dependent variable: written literacy score						
ITT	0.199***	0.227**	0.215**	0.204***	0.113*	0.334***
	(0.054)	(0.080)	(0.068)	(0.055)	(0.055)	(0.053)
	[0.014]	[0.017]	[0.025]	[0.018]	[0.073]	[0.000]
Observations	3,596	7,144	13,733	3,486	3,425	3,425
R-squared	0.42	0.52	0.41	0.42	0.39	0.42
Dependent variable: oral literacy score						
ITT	0.179***	0.240	0.254*	0.186***	0.046	0.284***
	(0.047)	(0.136)	(0.123)	(0.045)	(0.046)	(0.036)
	[0.010]	[0.130]	[0.025]	[0.006]	[0.351]	[0.002]
Observations	3,575	6,990	13,586	3,465	3,415	3,415
R-squared	0.38	0.49	0.43	0.38	0.35	0.38

Notes: Column 1 taken from Table 4. See text for details of samples and specifications in columns 2 to 6. Standard errors clustered at the unit of randomization included in parentheses.
*Significant at 10 percent, **significant at 5 percent, ***significant at 1 percent.
Critical values adjusted for number of clusters. *P*-values associated with wild cluster bootstrap standard errors appear in brackets.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

estimated for each country and test score. Column 1 repeats the preferred estimates from Table 4. In column 2 we reorganize the data as a panel, stacking observations from baseline and follow-up. This enables us to estimate a regression with student fixed effects in lieu of lagged test scores:

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

where $test_{ist}$ is the test score 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 treatment effect. The stacked sample of student observations is restricted to students who took both the pretest and posttest of the specified subject.

Unlike column 2, the specification in column 3 does not limit the sample to include students observed at both baseline and follow-up. Instead, it includes baseline observations even when students do not appear in the follow-up (because of attrition), and follow-up observations that do not appear in the baseline (because they were replacement students randomly sampled from the same cohort, as described in footnote 10). The regression, estimated in stacked, student-level data, is

$$test_{ist} = \alpha + \beta ITT_s * followup_t + \gamma followup_t + X'_{isj}\lambda + \delta_s + \varepsilon_{ist}$$

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

Column 4 limits the sample to students who were present at the follow-up and took all three baseline exams. (Recall that the preferred specification did not discard observations with missing baseline test scores, but instead dummies out missing values, as described in footnote 13.) Finally, columns 5 and 6 include a bounding exercise in the spirit of Lee (2009). The samples in each column remove students from each country's treatment group in order to match the higher level of attrition experienced by the control group. In column 5 we remove students from the top of the score distribution of the treatment group, creating a lower bound of the treatment effect. In column 6 we remove the lowest scoring students from the treatment group, creating an upper bound of the treatment effect.[23]

In general, the robustness checks do not overturn the main pattern of results.[24] The Kenyan treatment effects continue to be small and statistically insignificant in numeracy and written literacy. The original effect of 0.077 in oral literacy, statistically significant at 10 percent, varies from 0.04 to 0.08 in other columns, though it is often estimated with less precision, as in columns 2 and 3. The Ugandan results are generally consistent with the preferred specification. Across all columns, the large and statistically significant effects on written literacy are 0.11 to 0.29. For oral literacy, the range of point estimates is 0.05 to 0.25, though the lowest estimate is not statistically significant. That estimate, in column 5, uses a sample that removed the highest scoring students from the treatment group. It is a lower bound, but one that

[23] Because of concerns over baseline imbalance, the estimates in columns 5 and 6 still control for baseline test scores. As an additional test of differential attrition, we ran regressions of each baseline test score on an attrition dummy, *ITT*, and an interaction between the attrition dummy and *ITT*. None of the Ugandan coefficients on the interactions were statistically different from zero, suggesting that differential attrition is not the source of our statistically significant findings within each country.

[24] The results are substantively similar if we estimate models similar to the preferred specification with the percentage correct instead of the IRT test scores as the dependent variables, while including additional controls for the test form and cohort.

Table 11. Robustness, pooled sample.

	Single treatment effect across both countries (1)	Differential effect, pooled estimation (2)
Dependent variable: numeracy score		
ITT	0.042	−0.027
	(0.051)	(0.062)
ITT × Uganda		0.169
		(0.106)
Observations	8,927	8,927
R-squared	0.60	0.60
Dependent variable: written literacy score		
ITT	0.100***	0.025
	(0.034)	(0.034)
ITT × Uganda		0.186***
		(0.060)
Observations	8,898	8,898
R-squared	0.62	0.63
Dependent variable: oral literacy score		
ITT	0.124***	0.067
	(0.035)	(0.046)
ITT × Uganda		0.139**
		(0.066)
Observations	8,880	8,880
R-squared	0.54	0.54

Notes: Sample includes students in both countries who completed the specified follow-up test and at least one baseline test. All regressions include gender, cohort, and strata dummy variables, as well as dummy variables indicating missing values of each baseline test score. Standard errors clustered at the unit of randomization included in parentheses.
*Significant at 10 percent; **significant at 5 percent; ***significant at 1 percent.
Critical values adjusted for the number of clusters.

is unlikely to be correct given evidence from Table 3 that the correlation between test scores and the probability of attrition is similar across treatment and control groups.

Finally, the estimates in Table 11 use a pooled sample from both countries, facilitated by the commonly scaled tests. In column 1, across both countries, RTL 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. Column 2 contains the estimates over the same pooled sample, allowing for differential effects by country. The results confirm that the point estimates for Uganda are statistically different than the Kenyan point estimates for both written and oral literacy.

CONCLUSIONS

We used field experiments to evaluate the RTL method of increasing early-grade literacy across four districts in Kenya and Uganda. We find that the treatment increased written and oral literacy in Uganda by around 20 percent of a standard deviation. In Kenya, it had a smaller effect of 8 percent of a standard deviation on oral literacy. It did not affect numeracy test scores in either country. The findings are consistent with a growing literature showing that early-grade test scores can be influenced by school-based investments, at least if instructional inputs are

accompanied by well-aligned teacher training (Banerjee et al., 2007; Chay, McEwan, & Urquiola, 2005; Friedman, Gerard, & Ralaingita, 2010; He, Linden, & MacLeod, 2008, 2009; Piper & Korda, 2011).

The experiments also provided 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 setting—and perhaps mediated program effects—it was not markedly better or worse in either setting. An independent organization, APHRC, 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 hypotheses as to why program effects varied by country. We test for heterogeneous effects within each country by baseline student scores and classroom attributes, but find no evidence that treatment effects were larger among lower achieving students or in underresourced classrooms. We also find no evidence that differential implementation fidelity, as measured by presence of classroom inputs and AKF-observed classroom activities, can explain the differential effects. We tentatively conclude that the main driver of the differences was the interaction between the intervention and the country-specific instructional language. In Uganda students were tested in the instructional language, Lango. In Kenya, literacy exams were administered in Swahili, which was not the main instructional language and a subject taught for less time than English. It cannot be ascertained whether RTL effects might have been observed if English literacy assessments had been applied, but it is a plausible hypothesis for future research. The results highlight the vital importance of carefully adapting school curricula and outcome measures when treatments and impact evaluations are transferred from one context to another.

ADRIENNE M. LUCAS is an Assistant Professor of Economics, Department of Economics, Lerner College of Business and Economics, University of Delaware, 419 Purnell Hall, Newark, DE 19716 (e-mail: alucas@udel.edu).

PATRICK J. MCEWAN is a Professor of Economics, Department of Economics, Wellesley College, 106 Central Street, Wellesley, MA 02481 (e-mail: pmcewan@wellesley.edu).

MOSES NGWARE is the Head of Education Research Program, APHRC Campus, Manga Close Off Kirawa Road, 2nd Floor, 10787-00100 Nairobi, Kenya (e-mail: mngware@aphrc.org).

MOSES OKETCH is the Director of Research, APHRC, and Reader, Department of Humanities and Social Sciences, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, United Kingdom; APHRC Campus, Manga Close Off Kirawa Road, 2nd Floor, 10787-00100 Nairobi, Kenya (e-mail: moketch@aphrc.org).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of the William and Flora Hewlett Foundation and its Quality Education in Developing Countries (QEDC) initiative. We thank Maurice Mutisya and David Torres Irribarra for exceptional research assistance. For useful comments, we thank the Editor and three anonymous referees, as well as Elizabeth Adelman, Kristin Butcher, Alan de Brauw, Erick Gong, Ellen Green, Saul Hoffman, Paul Larson, Isaac

Mbiti, Lynn Murphy, Chloe O'Gara, Maria Perez, Dana Schmidt, Emiliana Vegas, and Andrew Zeitlin.

REFERENCES

Abeberese, A. B., Kumler, T. J., & Linden, L. L. (in press). Improving reading skills by encouraging children to read: A randomized evaluation of the Sa Aklat Sisikat reading program in the Philippines. Journal of Human Resources.

Adelman, S., Alderman, H., Gilligan, D. O., & Lehrer, K. (2008). The impact of alternative food for education programs on learning achievement and cognitive development in northern Uganda. Unpublished manuscript.

Aga Khan Foundation East Africa. (2013). The Reading to Learn (RTL) model. Unpublished manuscript.

Allcott, H., & Mullainathan, S. (2012). External validity and partner selection bias. NBER Working Paper No. 18373. Cambridge, MA: National Bureau of Economic Research.

Baird, S., Hicks, J. H., Kremer, M., & Miguel, E. (2012). Worms at work: Long-run impacts of child health gains. Unpublished manuscript.

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

Barr, A., Mugisha, F., Serneels, P., & Zeitlin, A. (2012). Information and collective action in the community monitoring of schools: Field and lab experimental evidence from Uganda. Unpublished manuscript.

Blimpo, M. P., & Evans, D. K. (2011). School-based management and educational outcomes: Lessons from a randomized field experiment. Unpublished manuscript.

Bold, T., Mwangi, Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). Scaling-up what works: Experimental evidence on external validity in Kenyan education. Unpublished manuscript.

Borkum, E., He, F., & Linden, L. L. (2012). School libraries and language skills in Indian primary schools: A randomized evaluation of the Akshara library program. NBER Working Paper No. 18183. Cambridge, MA: National Bureau of Economic Research.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. Review of Economics and Statistics, 90, 414–427.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. American Economic Review, 95, 1237–1258.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2013). School inputs, household substitution, and test scores. American Economic Journal: Applied Economics, 5, 29–57.

Dubeck, M. M., Jukes, M. C. H., & Okello, G. (2012). Early primary literacy instruction in Kenya. Comparative Education Review, 56, 48–68.

Duflo, E., Dupas, P., & Kremer, M. (2012). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. National Bureau of Economic Research Working Paper 17939. Cambridge, MA: National Bureau of Economic Research.

Friedman, W., Gerard, F., & Ralaingita, W. (2010). International independent evaluation of the effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) program in Mali, mid-term report. Research Triangle Park, NC: RTI International.

Glewwe, P. (2002). Schools and skills in developing countries: Education policies and socioeconomic outcomes. Journal of Economic Literature, 40, 436–482.

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: The case of flip charts in Kenya. Journal of Development Economics, 74, 251–268.

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. American Economic Journal: Applied Economics, 1, 112–135.

Glewwe, P., Nauman, I., & Kremer, M. (2010). Teacher incentives. American Economic Journal: Applied Economics, 2, 205–227.

Grogan, L. (2008). Universal primary education and school entry in Uganda. Journal of African Economies, 18, 183–211.

He, F., Linden, L. L., & MacLeod, M. (2008). How to teach English in India: Testing the relative productivity of instruction methods within the Pratham English language program. Unpublished manuscript. New York, NY: Columbia University.

He, F., Linden, L. L., & MacLeod, M. (2009). A better way to teach children to read? Evidence from a randomized controlled trial. Unpublished manuscript. New York, NY: Columbia University.

Jamison, D. T., Searle, B., Galda, K., & Heyneman, S. P. (1981). Improving elementary mathematics education in Nicaragua: An experimental study of the impact of textbooks and radio on achievement. Journal of Educational Psychology, 73, 556–567.

Kremer, M., Moulin, S., & Namunyu, R. (2003). Decentralization: A cautionary tale. Poverty Action Lab Paper No. 10. Cambridge, MA: Poverty Action Lab.

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

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

Lucas, A. M., & Mbiti, I. (2012a). Access, sorting, and achievement: The short-run effects of free primary education in Kenya. American Economic Journal: Applied Economics, 4, 226–253.

Lucas, A. M., & Mbiti, I. (2012b). Does free primary education narrow gender differences in schooling outcomes? Evidence from Kenya. Journal of African Economies, 21, 691–722.

McEwan, P. J. (2013). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. Unpublished manuscript. Wellesley, MA: Wellesley College.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72, 159–217.

Muralidharan, K., & Sundararaman, V. (2010). Contract teachers: Experimental evidence from India. Unpublished manuscript. San Diego: University of California.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. Journal of Political Economy, 119, 39–77.

Oketch, M., Ngware, M., Mutisya, M., Kassahun, A., Abuya, B., & Musyoka, P. (2012). East Africa Quality in Early Learning (EAQEL) impact evaluation report. Nairobi, Kenya: African Population and Health Research Center.

Piper, B., & Korda, M. (2011). EGRA Plus: Liberia, program evaluation report. Research Triangle Park, NC: RTI International.

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. Journal of Policy Analysis and Management, 26, 909–949.

Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Alishjabana, A., Gaduh, A., & Artha, R. P. (2011). Improving educational quality through enhancing community participation. Policy Research Working Paper 5795. Washington, DC: World Bank.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). What to do when data are missing in group randomized controlled trials. NCEE 2009–0049. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Uwezo. (2011a). Are our children learning? Annual learning assessment report. Nairobi, Kenya: Uwezo Kenya.

Uwezo. (2011b). Are our children learning? Annual learning assessment report. Kampala, Uganda: Uwezo Uganda.

Vermeersch, C., & Kremer, M. (2004). School meals, educational achievement, and school competition: Evidence from a randomized experiment. World Bank Policy Research Working Paper No. 3523. Washington, DC: World Bank.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago, IL: Mesa Press.