Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: http://www.tandfonline.com/loi/uree20

Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial

Matthew C. H. Jukes, Elizabeth L. Turner, Margaret M. Dubeck, Katherine E. Halliday, Hellen N. Inyega, Sharon Wolf, Stephanie Simmons Zuilkowski & Simon J. Brooker

To cite this article: Matthew C. H. Jukes, Elizabeth L. Turner, Margaret M. Dubeck, Katherine E. Halliday, Hellen N. Inyega, Sharon Wolf, Stephanie Simmons Zuilkowski & Simon J. Brooker (2016): Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2016.1221487

To link to this article: http://dx.doi.org/10.1080/19345747.2016.1221487

View supplementary material

Accepted author version posted online: 11 Aug 2016.
Published online: 11 Aug 2016.

Submit your article to this journal

Article views: 56

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=uree20

Download by: [Matthew Jukes] **Date:** 22 October 2016, At: 17:02

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
http://dx.doi.org/10.1080/19345747.2016.1221487

Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial

Matthew C. H. Jukes[a,b], Elizabeth L. Turner[c], Margaret M. Dubeck[a,b,d], Katherine E. Halliday[e], Hellen N. Inyega[f], Sharon Wolf[g], Stephanie Simmons Zuilkowski[h], and Simon J. Brooker[e]

ABSTRACT

We evaluated a program to improve literacy instruction on the Kenyan coast using training workshops, semiscripted lesson plans, and weekly text-message support for teachers to understand its impact on students' literacy outcomes and on the classroom practices leading to those outcomes. The evaluation ran from the beginning of Grade 1 to the end of Grade 2 in 51 government primary schools chosen at random, with 50 schools acting as controls. The intervention had an impact on classroom practices with effect sizes from 0.57 to 1.15. There was more instruction with written text and more focus on letters and sounds. There was a positive impact on three of four primary measures of children's literacy after two years, with effect sizes up to 0.64, and school dropout reduced from 5.3% to 2.1%. This approach to literacy instruction is sustainable, and affordable and a similar approach has subsequently been adopted nationally in Kenya.

KEYWORDS
early grade reading
literacy instruction
Africa
Kenya
RCT

Introduction

Literacy in Developing Countries

Despite recent improvements in access to schooling, literacy rates remain low in sub-Saharan Africa. Across the continent 63% of adults and 72% of youth aged 15–24 are literate (UNESCO, 2012). A number of studies (e.g., Gove & Cvelich, 2010; Uwezo, 2013) have found that in many countries in sub-Saharan Africa a large percentage of children fail to achieve functional literacy in the first three grades of school. Improving early-grade literacy

CONTACT Matthew C. H. Jukes ✉ mjukes@rti.org RTI International, 3040 East Cornwallis Road, PO Box 12194, Research Triangle Park, NC 27709-2194, USA.
[a]Harvard Graduate School of Education, Cambridge, Massachusetts, USA
[b]RTI International, Research Triangle Park, North Carolina, USA
[c]Duke University, Durham, North Carolina, USA
[d]University of Virginia, Charlottesville, Virginia, USA
[e]London School of Hygiene & Tropical Medicine, London, UK
[f]University of Nairobi, Nairobi, Kenya
[g]University of Pennsylvania, Philadelphia, Pennsylvania, USA
[h]Florida State University, Tallahassee, Florida, USA

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uree.
Supplemental data for this article can be accessed at http://dx.doi.org/10.1080/19345747.2016.1221487.

© 2016 Taylor & Francis Group, LLC

is therefore at the top of the global policy agenda, as evidenced by a series of recent initiatives and meetings. The World Literacy Summit led to the Oxford declaration (World Literacy Foundation, 2012), which called for action on five fronts, one of which was the need for "[a] strong evidence base for why universal literacy is fundamental to an individual's and country's success and evidence on strategies and best practices that are having the greatest effect." Between 2011 and 2015, the United States Agency for International Development's (USAID) Goal 1 was "improved reading skills for 100 million children in primary grades" (USAID, 2011), a goal that will continue through the next project cycle. A focus on early-grade reading is also explicit in the United Kingdom's Department for International Development (DFID) 2010–2015 strategy and is consistent with the World Bank's strategy to improve learning for all (World Bank, 2011).

Recent research on literacy in developing countries has focused on assessing and improving children's literacy in the early grades of primary school and the evidence base is slowly accumulating. We focus here on school-based strategies to improve reading. Although a complexity of contextual factors, including poverty, health, late enrollment, and limited access to print contribute to delayed reading acquisition (Badian, 1988; Heath, 1983; Hungi, Ngware, & Abuya, 2014; Jukes, Drake, & Bundy, 2008), government policy probably has the greatest influence over what happens in the classroom. One key factor that schools can influence is the method of instruction (Dubeck, Jukes, & Okello, 2012; Pressley, 2001; Stuhlman & Pianta, 2009), with evidence suggesting that students learn best when literacy skills are taught in an explicit, systematic, and appropriate way (Snow, Burns, & Griffin, 1998). Explicit means that the concept is directly taught and modeled so the student does not have to infer what the teacher means. Systematic instruction progresses in a sequence moving from easiest to more difficult. Learning to read any alphabetic system depends on understanding the relationship between sounds and the letters that represent them. Regardless of context, students who do not have this understanding are likely to struggle with reading. Despite the growing consensus on the need to develop literacy skills in an explicit and systematic manner, educators in some countries are only just beginning to teach skills that are known to improve literacy levels (Anderson–Levitt, 2004; Arnold, Bartlett, Gowani, & Merali, 2006).

Kenyan Context and Policy

Since Kenya abolished school fees in 2003, most children in Kenya now enroll in school. However, limited funding has led to increased class sizes (World Bank, 2014), student–textbook ratios of 3:1 (Piper & Mugenda, 2012), and shortages of classroom space and teaching materials (Sifuna, 2007; UNESCO, 2005). These resource constraints make it difficult for teachers to provide their students with a quality education. A national survey in Kenya found that more than half of students in Grade 3 are unable to infer meaning from short passages of text (Wasanga, Ogle, & Wambua, 2010). A number of other assessments found similar results—Kenyan children may have had access to school, but were not necessarily learning much there (Mugo, Kaburu, Limboro, & Kimutai, 2011; Onsomu, Nzomo, & Obiero, 2005; Piper, 2010; Piper & Mugenda, 2012).

At the time of the study reported in this article (2010–12) the Kenyan education policy did not mandate a specific method to teach reading. Instead, the policy suggested that teaching methods should meet the students' learning needs and the objective for the lesson (Ministry of Education, 2006). Generally, these methods could include teaching the relationships between the letters and their sound (i.e., phonics), teaching words as a whole (i.e., look–say), or a combination of these (Commeyras & Inyega, 2007). Our own analysis (Dubeck et al., 2012) in the study region found that attention to developing oral language skills was prioritized over teaching the relationships between sounds and symbols. In general, teachers use lecture and whole-class oral pedagogies in Kenya (Ackers & Hardman, 2001; Dubeck et al., 2012; Pontefract & Hardman, 2005).

The Kenyan national education policy specified the use of the mother tongue (i.e., the local language spoken in a student's home) as the language of instruction in Grades 1 through 3, transitioning to English in Grade 4 and thereafter (Kibui, 2014; Ministry of Education, 2006). However, in practice English was used widely in the early primary grades (Lewis, 2009; Piper & Miksic, 2009; Trudell & Piper, 2014). Both English and Swahili (referred to in the Kenyan education system as "Kiswahili") are taught as subjects to all students starting in Grade 1. In coastal Kenya, where our study took place, mother tongue languages are predominantly those from the Mijikenda family of nine related ethnic groups. Swahili is a lingua franca in the region.

The HALI Literacy Intervention

The aims of the HALI (Health and Literacy Intervention) project were to improve the literacy outcomes of schoolchildren, to reduce their burden of malaria and to investigate the interaction between these two interventions. However, there was no impact of the malaria screening and treatment program on educational achievement nor any interaction between malaria and literacy interventions (Halliday et al., 2014). In this article we report only the literacy intervention evaluation. In all aspects of design, the literacy intervention sought to build on effective instructional practices that were already in use locally. For example, we found (Dubeck et al., 2012) that during Swahili instruction some teachers were explicitly teaching the relationship between sounds and syllables. The HALI intervention sought to expand this practice in Swahili and encourage its use in English and in general, to help teachers use literacy skills in one language to aid literacy acquisition in another. We also aimed to increase children's engagement with print during common practices such as song and oral reading.

The HALI literacy intervention is described in greater detail elsewhere (Dubeck, Jukes, Brooker, Drake, & Inyega, 2015) and summarized here. The intervention supported teachers in developing the literacy skills of one cohort of children through the first two years of primary school and contained the following elements:

- 140 sequential, semiscripted lesson plans for literacy sessions, each one in either Swahili or English, which were given to all participating teachers;
- Training, including a three-day initial workshop that included guided opportunities to create new instructional materials, a problem-solving workshop four months after the commencement of the school year, and a refresher training the following school year;
- Ongoing support for teachers for two years through weekly text messages providing brief instructional tips and motivation to implement lesson plans. Teachers also received credit of $0.50—around 50 Kenyan shillings—each week for their mobile phones. A total of 200 Kenyan shillings over the course of a month represents about 1% of the 16,662 Kenyan shillings starting salary for primary school teachers (IEA, 2014).

There has been increasing interest in the use of mobile phone technology as an educational tool in Africa (South African Institute for Distance Education, 2008; Valk, Rashid, &

Elder, 2010), based on the high rate of ownership of the device on the continent. In Kenya, it was estimated in 2012 that 71% of the population owned a mobile phone (TNS, 2012). There is promising evidence elsewhere in sub-Saharan Africa that mobile phones can be used effectively to improve adult literacy (Aker, Ksoll, & Lybbert, 2012; Beltramo & Levine, 2012) and they are increasingly being used remotely to provide resources for teachers, both for their own professional development and for use in the classroom (Walsh et al., 2013). However, we are not aware of any previous published evaluation of a teacher professional development program using mobile phones to coach and support teachers in a low-income country. Subsequent to our intervention, evaluations have emerged showing a positive impact of a teacher training program involving text-message communication (Piper, Zuilkowski, & Mugenda, 2014) and also of a program of teacher support using tablets (Piper, Jepkemei, Kwayumba, & Kibukho, 2015; Piper, Zuilkowski, Kwayumba, & Strigel, 2016).

The aim of this study was to evaluate the effectiveness and cost-effectiveness of the HALI intervention in improving children's literacy outcomes. A key additional aim was to investigate the mechanisms by which the intervention improved outcomes. We argue, with others (Funnell & Rogers, 2011; Pawson & Tilley, 1997; Stern et al., 2012; White, 2009), that testing the theory underlying the intervention is essential if the lessons of the evaluation are to be applied in other contexts and to other programs. Despite a number of experimental studies investigating the impact of improved instruction on literacy outcomes in recent years, there is very little rigorous evidence on the changes seen in the classroom that lead to improved outcomes. Classroom observation studies in Africa have found improvements in group work and teacher–pupil interaction (Hardman et al., 2009) and in teacher perceptions (Sailors et al., 2014) in response to training. One observational study in Kenya found associations between teacher behavior and student achievement (Ngware, Oketch, & Mutisya, 2014). However, there remains a lack of evidence on classroom processes from experimental studies in general and from early grade reading interventions in particular.

Methods

The HALI literacy intervention was evaluated between January 2010 and March 2012 together with a program of screening and treatment for malaria. The evaluation was conducted by the same team that implemented the HALI program. To avoid any bias in findings that this might entail, a number of measures were put in place. An independent data-monitoring committee was set up to scrutinize data collection and analysis. An analysis plan was submitted to the committee before data were inspected in order, among other things, to ensure that choice and definition of outcome variables were fixed before data were analyzed. Impact analyses were conducted by a statistician working independently from the HALI implementation team. To avoid bias in data collection, the assessment team were blind to the intervention status of the school and were highly trained so that they responded to all students in the same scripted way, reducing the influence of assessor beliefs on student scores. Each of these measures is described in more detail in the relevant section below.

The evaluation involved a cluster randomized trial (Brooker et al., 2010), in which 101 public primary schools were randomly allocated to one of four arms receiving either: (a) the malaria intervention alone; (b) the literacy intervention alone; (c) both interventions combined; or (d) neither intervention. Children from Grade 1 were randomly selected and followed up for 24 months to assess the impact of the interventions.

Participants

The study was conducted in rural government primary schools in Kwale and Msambweni districts,[1] approximately 50 km south from Mombasa on the Kenyan coast (Figure 1). We excluded schools that were more than 70 km from the project office and those with another literacy project taking place, leaving 101 schools—21 and 80 schools in Kwale and Msambweni districts, respectively. Participants were students in Grade 1 at the start of study in each of the 101 schools.

Sample Size Estimation

A sample size of 100 schools with 25 children per class and one class per school was assumed, with 50 schools each randomized to control and to intervention, resulting in a total of 2,500 children. This is sufficient to detect an effect size of 0.19 standard deviation (*SD*) with 80% power at the 5% significance level (Hayes & Bennett, 1999), assuming an intraclass correlation (ICC) of 0.2 (ICC varied from 0.1 to 0.2 with mathematics and literacy tests in Grade 2 in a previous impact evaluation in western Kenya; Clarke et al., 2008) and a correlation between baseline and final outcomes of 0.7. In practice, because group tests were conducted in groups of 15, we randomly selected 30 children per class, where available, which resulted in a larger sample size than originally planned and accommodated for attrition.

Random Allocation of Schools

We invited all available 101 schools to participate in the study and all schools accepted. Random allocation of the 101 schools to study arm was conducted in two stages, each involving public randomization ceremonies to ensure the process was transparent to stakeholders. At the time of our study, schools in Kenya were divided into Teacher Advisory Center (TAC) tutor zones,[2] which meet to discuss curriculum and progress, supported by a TAC tutor. In our sample there were 26 TAC tutor zones with between three and six schools each. Inclusion of control schools and intervention schools from the same TAC tutor zone may have led to leakage of the intervention between schools. Such contamination was documented in a literacy instruction evaluation in a nearby district (Crouch, Korda, & Mumo, 2009). Therefore, the first stage of randomization involved random allocation by TAC tutor zone so that each of the 26 TAC tutor zones were randomly allocated to receive either the literacy intervention or to serve as a literacy control. This ensured that no TAC tutor zone contained both intervention and control schools. Randomization was stratified by (a) TAC tutor zone size and (b) average primary school leaving exam scores across the TAC tutor zone. Overall, 51 schools were randomized to literacy intervention and 50 schools to literacy control. The second stage involved random allocation of the health intervention by individual school rather than TAC tutor zone since there was no concern of contamination between schools in the same TAC tutor zone. This randomization was stratified by school achievement and

[1] Kenyan administrative regions were redrawn in 2013. Mswambeni district is now Msambweni subcounty and Kwale district is divided into Matuga and Kinango subcounties.

[2] Now called Curriculum Support Officers.

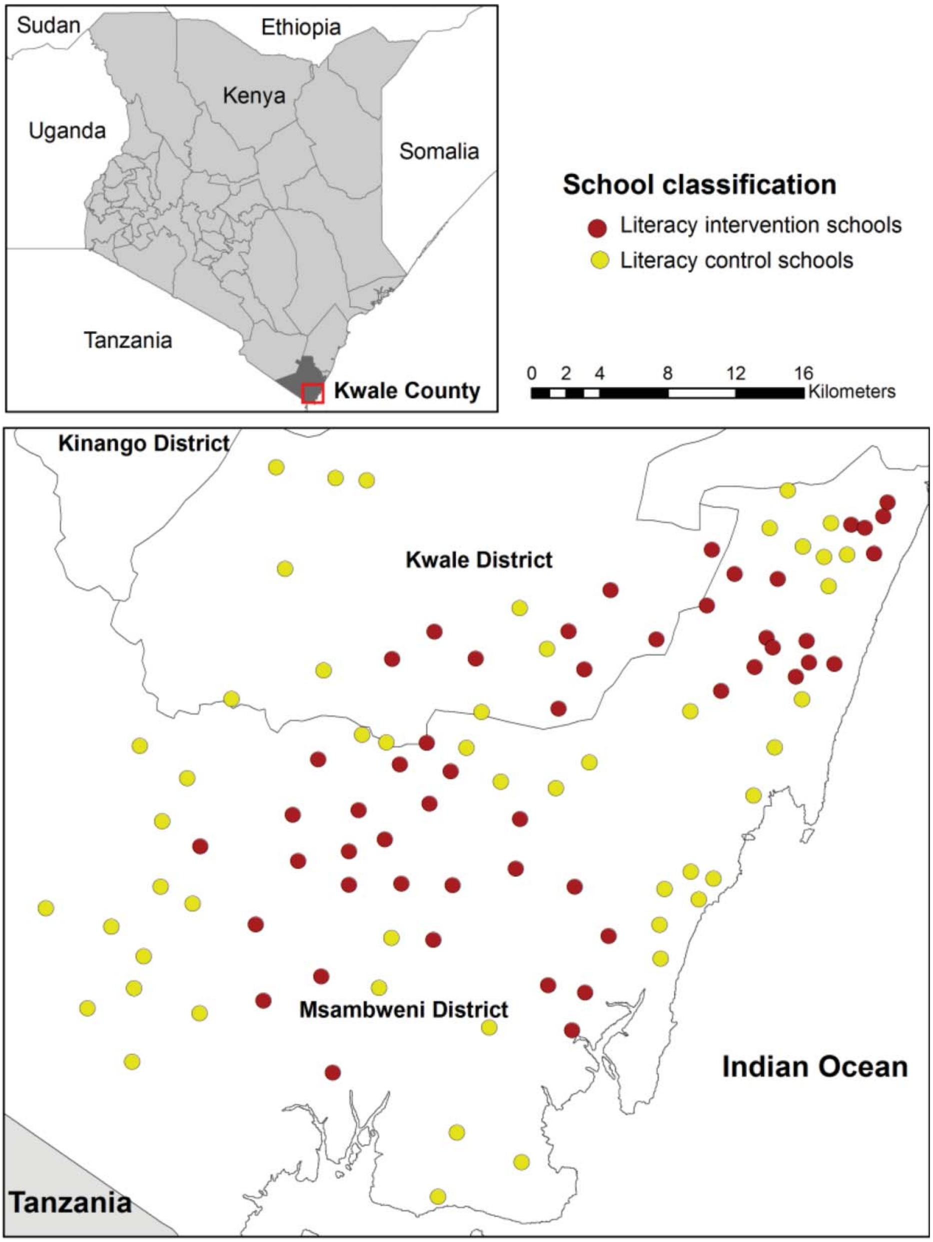

Figure 1. Map of study schools in Kwale and Msambweni districts, coastal Kenya. Inset: Map of Kenya with Kwale and Msambweni districts shaded in gray.

literacy group allocation, yielding 51 schools in health intervention and 50 schools in health control.

Sensitization and Recruitment

At the national level, the study was approved by the Division of Malaria Control; Ministry of Public Health and Sanitation; and Ministry of Education, Science and Technology. At provincial and district levels, meetings were held with the Provincial Medical Officer and the Provincial Director of Education in Mombasa, as well as district health and education officials in Kwale and Msambweni. Consent was also sought from school headteachers and TAC tutors. Prior to randomization, enumeration of children was carried out and school meetings were held with parents and guardians to explain the study and seek written informed consent. Children provided verbal informed assent prior to assessment in the school. The study was approved by the Kenya Medical Research Institute and National Ethics Review Committee (SSC number 1543), the London School of Hygiene & Tropical Medicine Ethics Committee (5503), and the Harvard University Committee on the Use of Human Subjects in Research (F17578-101).

Outcome and Measures

Children's educational outcomes were assessed at baseline, and at 9 months (FU1) and 24 months (FU2) follow-up (Table 1). The main aim of the outcome measures was to assess

Table 1. Test description, test–retest reliability over three days and test usage in three survey time points.

					Outcomes measured at each time point		
Domain	Test (score range or scale)	Individual (I) or group (G) assessment	Test–retest reliability	Reliability sample size	Baseline	Follow-up 1 (9 mos)	Follow-up 2 (24 mos)
Phonological awareness	Beginning sound identification (0–10)	G	0.89	33	●	●	
Oral vocabulary	Swahili receptive language (0–25)	G	0.85	35	●	●	
Letter reading	Swahili letter sounds (lpm)	I	0.83	30	●	●	●
	English letter knowledge (lpm)	I	0.62	30	●	●	●
Word reading	Swahili word identification (wpm)	I	0.84	30	●	●	●
	English word identification (wpm)	I	0.89	30	●	●	●
Passage reading	Swahili passage reading fluency (wpm)	I	0.92	30	●	●	●
	English passage reading fluency (wpm)	I	0.93	30	●	●	●
	Swahili passage comprehension (0–5)	I	—	30	●	●	●
	English passage comprehension (0–5)	I	—	30	●	●	●
Spelling	Spelling (0–20)	G	0.82	65	●	●	●
Sustained attention	Pencil Tap (0–20)	I	0.61	35	●		
	Code transmission (0–20)	G	0.85	35		●	●
Numeracy	Number Identification (0–20)	I	0.95	35	●	●	
	Quantity Discrimination (0–10)	I	0.71	35	●	●	
	Arithmetic (0–60)	I	0.85	35			●
Nonverbal reasoning	Ravens (0–22 at baseline; 0–12 at follow-up)	I	0.60	35	●	●	●

Notes. lpm=letters per minute; wpm=words per minute.

children's progress through phases of literacy acquisition. We also included assessments of early numeracy skills, sustained attention, and nonverbal reasoning to assess whether the literacy intervention had spillover effects in other domains. Some tests were conducted orally with individuals in both Swahili and English; other tests were in small groups and required written responses. Group tests included those of receptive language and beginning sounds phonemic awareness, which were assessed only in Swahili, and spelling, which was assessed only in English.

Educational Assessments

Educational assessments are available from corresponding author on request.

Literacy Assessments

Literacy tests were adapted from previous work in Kenya (Jukes, Vagh, & Kim, 2006), and the Early Grade Reading Assessment (EGRA; RTI International, 2009), both of which have been widely used in the least developed countries in sub-Saharan Africa and elsewhere and from the Phonological Awareness Literacy Screening (Invernizzi, Juel, Swank, & Meier, 2007). Tests were conducted at all three time points except for beginning sound awareness and receptive language, which develop earlier than other literacy skills and were assessed in Grade 1 (baseline and nine months) only (see Table 1).

Beginning Sound Awareness. A group test of Swahili phonological awareness. The assessor names a target picture and three additional ones. Children mark the picture that has the same beginning sound as the target one. Items incorporate age-appropriate vocabulary, consisting of words with concrete meanings. Ten items in total.

Receptive Language. A group test of Swahili oral language skills. For each item, a word is read out in Swahili and children are required to mark the corresponding picture from a choice of four on their answer sheet. Twenty-five items in total.

Spelling. A group test in English only, which assesses phonemic awareness and letter knowledge. Students spell five words with consonant-vowel-consonant syllable patterns (e.g., sad). Credit is given for phonetically acceptable choices for beginning and ending consonant sounds (e.g., "p" for "b") and for the middle vowel. Reversals (e.g., "b" for "d") are considered handwriting errors, not spelling errors. Students receive one correct mark for each phoneme that is ascribed the correct letter or an acceptable substitution and one bonus point for each word spelled correctly. The maximum score is four per word or 20 in total.

Letter Reading Fluency. Children are given 60 seconds to read 100 letters presented in random order. The English version of the test involves reading the names of the letters (*ay, bee, cee)* presented in uppercase. The Swahili alphabet is presented in lowercase and consists of 23 letters and seven digraphs (e.g., dh, sh), and children give the letter sounds. The score on both versions is rate of letters identified correctly per minute.

Word Reading Fluency. Two individual tests in English and Swahili. Students are required to read a list of 50 words typically found in beginning reading materials. The score is the rate of words read correctly per minute.

Passage Reading Fluency. Students read a Grade-2 level narrative passage in both English and Swahili. The score is the rate of words read correctly per 60 seconds.

Passage Reading Comprehension. After the passage-reading fluency assessment, students are asked questions that correspond to the text they read. It includes four explicit comprehension and one inferential question. The score is the total correct out of five.

Numeracy Assessments

Numeracy assessments were based on previous work in Kenya (Jukes, Vagh, & Kim, 2006) and the Early Grade Mathematics Assessment (Reubens, 2009). In Grade 1, the numeracy assessments consisted of a combined score from the number identification and quantity discrimination subtests.

Number Identification. Students are given a list of one-, two-, and three-digit numbers to read out in one minute. Maximum score is 20.

Quantity Discrimination. Students are presented with a pair of one-, two-, and three-digit numbers, which the administrator also reads out, and are asked which one is bigger. Maximum score is 10.

Written Numeracy. At the 24-month follow-up, numeracy was assessed with only an untimed written test involving 38 questions of basic arithmetic.

Sustained Attention and Reasoning

Sustained attention was assessed using the Pencil Tapping test at baseline and Code Transmission at FU1 and FU2.

Pencil Tapping. An individual test of sustained attention and executive control, adapted from existing measures (Bierman, Nix, Greenberg, Blair, & Domitrovich, 2008; Diamond & Taylor, 1996). Students are required to tap a pencil on the desk once when the assessor taps twice, twice when the assessor taps once and not at all when the assessor taps three times. A delay of up to 30 seconds is introduced between stimulus and response. As a distractor, the child is given a page of shapes to color while they wait for the taps. The maximum score is 20.

Code Transmission. A group test of sustained attention used elsewhere in Kenya (Clarke et al., 2008) and adapted from the TEA-Ch (Tests of everyday attention for children) battery (Manly, Robertson, Anderson, & Nimmo-Smith, 1999). A list of digits is read out loud at the speed of one every two seconds and children are required to listen for a "code"—two consecutive occurrences of the number five—and then record the number that preceded the code. Due to floor effects this test was not used at baseline. The maximum score is 20.

Raven's Progressive Matrices. Nonverbal reasoning was assessed by the Raven's Progressive Matrices task (Raven, Styles, & Raven, 1998).

Before the study began, all instruments were adapted to the Kenyan context over a period of five months to ensure face validity and appropriate stimuli. The provisional battery of tests was administered in five schools in the study area to assess test–retest reliability over a period of one week, relationships among tests (concurrent validity), and properties of individual test items.

Table 1 shows test–retest reliability data over three days for each of the assessments. Inter-rater reliability was assessed for the spelling test because of its relatively complex scoring procedures and was found to be high (0.99) among a sample of 20 assessors. A test of print concepts (e.g., concept of word in text) was dropped from the test battery because of low reliability.

Observations and Teacher Interviews

During the second school term of the evaluation, before the FU1 outcome measurements, two unannounced visits to each school were carried out to conduct teacher interviews and observations.

Classroom Observations. Adapted from the Classroom Language Arts Systematic Sampling and Instructional Coding (CLASSIC) classroom observation tool (Scanlon, Gelzheiser, Fanuele, Sweeney, & Newcomer, 2003). In all schools, an assessor observed English and Swahili lessons on two separate visits. In intervention schools only, the same assessor also observed a third lesson on each visit in which the HALI intervention materials were taught. Every 90 seconds a "snapshot" of the classroom was taken to provide detailed information about student and teacher behavior in five categories. The following description summarizes these categories with emphasis on behaviors highlighted in our results: (a) The *material* to which the teacher is referring or to which the pupils are attending. It could be a visible object such as written text or something spoken like a rhyme. The key classification used in our analyses was whether the material was written (e.g., on a book or chalkboard) or oral (e.g., a spoken rhyme). (b) The *language part* to which the teacher is referring or to which the pupils are attending. These were categorized as story or rhyme, sentence, word, word part, letter, or sound. (c) The teacher's specific *instructional focus* (e.g., the meaning of the word or taking the word apart). (d) The *teacher activity* (e.g., gaining student attention, reading to children). (e) The *student response* or activity. The key student response categories of interest were student reading and student writing.

Table 2 shows the classification of classroom behaviors in greater detail together with inter-rater reliability statistics. High inter-rater reliability was achieved for materials, language part, and student response. It was challenging to achieve high reliability with instructional focus and teacher activity categories, which were more subjective. These categories were removed from subsequent analysis.

Teacher Interviews. The teacher interview was a mixture of closed- and open-ended questions conducted at the end of the lesson being observed. It was based partly on a

Table 2. Inter-rater reliability for observation of behaviors in 10 classes by 10 raters.

Category	Subcategory	Example	Inter-rater reliability (Fleiss's kappa)
Materials	Mode	Written, oral	0.78
	Type	Book, choral	0.76
Language part	Language part	Word, sentence	0.79
Instructional focus	Focus	Meaning, taking the word apart	0.45
	Type	Prediction, rhyme	0.59
Teacher activity	Behavior	Give instructions, read aloud	0.59
	Teacher position	Front of class	1.00
	Teacher feedback	Positive response with explanation	0.69
Student response	Behavior	Reading, oral response, writing	0.88
	By whom	Individual girl, small group	0.84
	To whom	Teacher, classmate	1.00

questionnaire developed in previous work in Kenya (Jukes et al., 2006) and asked about the observed lesson as well as the teachers' background and experience.

Classroom Inventory. This observation schedule recorded the state of the classroom and the materials present, with a particular focus on those suggested in the teacher training.

Household and School Questionnaires. These were conducted to obtain background information about students' households and about school quality (Filmer & Prichett, 2001).

Data Analysis

The trial protocol was registered with the U. S. National Institute of Health Clinical Trials Registry (ClinicalTrials.gov; Identifier: NCT00878007). Primary and secondary outcomes were prespecified in a statistical analysis plan and approved by an independent data monitoring committee (DMC) separately for 9-month (FU1) and 24-month (FU2) analyses. Primary outcomes were those used as key indicators of whether the intervention was effective. Secondary outcomes were other indicators that were judged to be less important indicators or those anticipated to be less likely to be affected by the intervention. There were three prespecified primary outcomes at both FU1 and FU2: spelling score (0–20), Swahili letter sounds (letters per minute [lpm]) and English letter knowledge (lpm), with Swahili word identification (words per minute [wpm]) prespecified as a fourth primary outcome at FU2.

Analyses were conducted at the child level on an intention-to-treat basis. All outcomes were numerical test scores (e.g., spelling score) or continuous measures of scores per minute (e.g., letters read per minute). Linear mixed effects models were used to account for the hierarchical structure of the data. Specifically, independent zero-mean normally distributed random effects were included for: (a) TAC tutor zone (ν_j), (b) school (θ_{jk}), and (c) child (η_{jkl}), with an additional residual random error (ε_{jklm}). Intervention effects for FU1 and FU2 were estimated from the same model by including fixed effects for follow-up visit, literacy intervention arm, and their interaction. Prespecified primary analyses for each outcome included adjustment for the baseline measure of that outcome, age and sex, and adjustment for study design features by including the indicator terms for malaria intervention arm and for the school-cluster exam performance score (i.e., a proxy for the stratification factor).

The primary form of the model can be expressed as follows:

$$\begin{aligned}
&(\text{Primary Model})\ \mathrm{Y_{jklm}} \\
&\quad = \alpha_1 + \alpha_2\, FU2_{jklm} + \beta_1 T_j + \beta_2 T_j FU2_{jklm} + \gamma_{Baseline} Y_{jkl0} + \gamma_{Malaria\ int} I_{jk} \\
&\qquad + \gamma_{Age} A_{jkl} + \gamma_{Male} M_{jkl} + \nu_j + \theta_{jk} + \eta_{jkl} + \varepsilon_{jklm} \\
&\qquad \nu_j \sim N\left(0, \sigma_\upsilon^2\right); \theta_{jk} \sim N\left(0, \sigma_\theta^2\right); \eta_{jkl} \sim N\left(0, \sigma_\eta^2\right); \varepsilon_{jklm} \sim N\left(0, \sigma_\varepsilon^2\right)
\end{aligned}$$

for $j = 1, \ldots., 26$ TAC tutor zones, $k = 1, \ldots., S_j$ schools in the jth TAC tutor zone, $l = 1, \ldots.,$ n_{jk} child in the kth school in the jth TAC tutor zone and for $m = 1, 2$ follow-up visits corresponding to FU1 (9 months) and FU2 (24 months), respectively. The indicator terms $FU2_{jklm}$, T_j, and I_{jk} indicate whether an outcome is at FU2, whether the jth school cluster was allocated to literacy intervention and whether the jkth school was allocated to the malaria intervention, respectively. Other fixed terms are as follows: Y_{jkl0} is the baseline measure of the outcome (or of an appropriate proxy measure), A_{jkl} is baseline age, and M_{jkl} is an indicator of whether the child is male.

In order to compare effects on different outcomes, all results were standardized by dividing by the control arm *SD* at each of the follow-up time points (Glass, 1976). We refer to these estimates as standardized effects sizes and the former nonstandardized estimates as the adjusted mean difference. A second set of prespecified models included additional adjustment for SES (five categories: poorest—least poor), language spoken at home (three categories: Mijikenda language group, Swahili, and other) and preschool attendance, which were included as fixed effects in the primary model equation (see above). Restricted maximum likelihood estimation was used, and all analyses were performed in Stata 13.0 (StataCorp, 2013).

Outcome data for each child may be missing at one or both of the follow-up time points. This could arise because a child did not provide assent to participate at that time point or was not present at school when we conducted the outcome assessments or because the child had formerly withdrawn from the study. Such nonresponse could lead to bias of the estimated intervention effect if, for example, different types of children were lost to follow-up in the intervention arm compared to the control arm. To determine the effects of nonresponse, we performed sensitivity analyses that explicitly accounted for predictors of missing outcomes in order to try to correct for bias in the estimated intervention effect. The specific steps are as follows. We first determined whether treatment arm or any baseline child or home-environment characteristics were predictive of nonresponse at FU1 and separately at FU2 using logistic regression. We then refitted the primary impact analysis model shown above for each outcome and adjusted for all characteristics indicated to be independent predictors of nonresponse (with $p < 0.1$) at either FU1 or FU2. Results from such likelihood-based analyses provide unbiased estimates of the adjusted intervention effects (Fitzmaurice, Laird, & Ware, 2011), assuming that we had identified all covariates that were predictive of nonresponse (i.e., assuming that the nonresponse data were missing at random [MAR], conditional on the identified covariates). We used results from these models as a sensitivity analysis to evaluate whether our primary estimates of the intervention effect were sensitive to further adjustment.

Subgroup analyses were performed for all literacy outcomes by using the model indicated in the equation above, with an interaction between each follow-up indicator and the indicator (or indicators, if more than two levels) for the subgroup. Prespecified subgroups according to variables measured at study baseline included gender, preschool attendance, having books at home, reading aloud at school, getting help with schoolwork at home, missing breakfast, anemia, stunting, having repeated a grade, and baseline literacy levels.

School dropout was analyzed using a logistic regression model analogous to the primary model above: Random effects were included at the school and TAC tutor zone levels with age, sex, baseline SES, school-cluster exam performance score, and an indicator for malaria intervention as covariates.

Results

Study Profile

In total, 3,753 children were randomly selected for the consent process before visiting schools in January 2010. At initial school meetings, parents of 2,838 (75.6%) children were present and gave consent. Of these we selected a maximum of 30 children in each school to take part in the baseline survey, totaling 2,539. On average (*SD*), 25.3 (1.7) and 24.6 (3.3) children were selected in control and intervention schools, respectively. Due to challenges in conducting surveys in one of the intervention schools, only nine children in that school were finally enrolled. Of the 2,539 study children, 2,516 (99.1%) took part in baseline assessments, 2,238 (88.1%) in the first follow-up at 9 months and 2,030 (80.0%) in the second follow-up at 24 months. Participation in each follow-up was comparable for both intervention and control arms. By the end of data collection, 69 (2.7%) children had exited the study because they were deceased or had withdrawn. The trial profile is detailed in Figure 2.

Baseline Participant Characteristics

Table 3 shows that children had broadly similar characteristics in each of the literacy intervention study arms. School-level factors were similar across study arms in terms of exam scores and despite the differences in variability of school sizes by study arm, the median school size was similar. However, a higher proportion of schools in the control arm had school feeding programs when surveyed in January 2010. The school experiences of the children were highly comparable across study arms. Overall, most children (95.5%) had attended preschool, a third (32.1%) had previously failed Grade 1, and most (86.9%) read aloud in class. Overall home characteristics included Digo as the most common language used at home, which was more common in intervention than control (51.7% vs. 40.9%) and a third (34.3%) of children's parents never read to them. There were some other imbalances in terms of socioeconomic status (lower in the control arm) and schooling of household head (lower in the control arm). Table 4 shows that baseline measures of the educational outcomes were broadly comparable between arms. Overall, as expected at the start of Grade 1, most scores were low except for English letter knowledge, with an overall mean of 16.4 lpm and Swahili receptive language with a mean of 18.1 lpm.

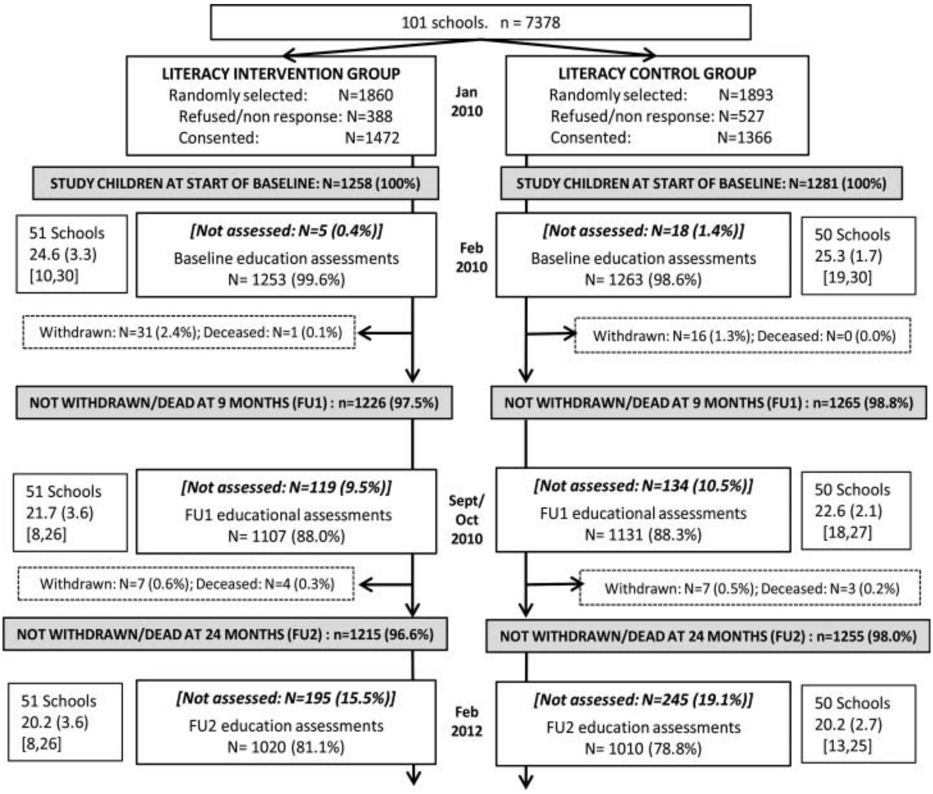
101 schools. n = 7378
LITERACY INTERVENTION GROUP
Randomly selected: N=1860
Refused/non response: N=388
Consented: N=1472
Jan 2010
LITERACY CONTROL GROUP
Randomly selected: N=1893
Refused/non response: N=527
Consented: N=1366
STUDY CHILDREN AT START OF BASELINE: N=1258 (100%)
STUDY CHILDREN AT START OF BASELINE: N=1281 (100%)
51 Schools
24.6 (3.3)
[10,30]
[Not assessed: N=5 (0.4%)]
Baseline education assessments
N= 1253 (99.6%)
Feb 2010
[Not assessed: N=18 (1.4%)]
Baseline education assessments
N= 1263 (98.6%)
50 Schools
25.3 (1.7)
[19,30]
Withdrawn: N=31 (2.4%); Deceased: N=1 (0.1%)
Withdrawn: N=16 (1.3%); Deceased: N=0 (0.0%)
NOT WITHDRAWN/DEAD AT 9 MONTHS (FU1) : n=1226 (97.5%)
NOT WITHDRAWN/DEAD AT 9 MONTHS (FU1) : n=1265 (98.8%)
51 Schools
21.7 (3.6)
[8,26]
[Not assessed: N=119 (9.5%)]
FU1 educational assessments
N= 1107 (88.0%)
Sept/ Oct 2010
[Not assessed: N=134 (10.5%)]
FU1 educational assessments
N= 1131 (88.3%)
50 Schools
22.6 (2.1)
[18,27]
Withdrawn: N=7 (0.6%); Deceased: N=4 (0.3%)
Withdrawn: N=7 (0.5%); Deceased: N=3 (0.2%)
NOT WITHDRAWN/DEAD AT 24 MONTHS (FU2) : n=1215 (96.6%)
NOT WITHDRAWN/DEAD AT 24 MONTHS (FU2) : n=1255 (98.0%)
51 Schools
20.2 (3.6)
[8,26]
[Not assessed: N=195 (15.5%)]
FU2 education assessments
N= 1020 (81.1%)
Feb 2012
[Not assessed: N=245 (19.1%)]
FU2 education assessments
N= 1010 (78.8%)
50 Schools
20.2 (2.7)
[13,25]

Figure 2. Trial profile: Flowchart of random allocation and study design. The percentages refer to the percentage of children who were invited to participate in the study who provided informed consent and enrolled in the trial. Data for the schools are: Number of schools; Mean (*SD*) number of children [min, max].

Teacher Compliance With Literacy Intervention

A total of 62 Grade-1 teachers were initially trained in the 51 schools in February 2010, more than one per school because some schools had multiple streams. However, we collected and analyzed data from only one stream of children per school in order to minimize costs. Teachers who transferred into intervention schools during the first term were given a one-day intensive training in their school. At the start of the second year (February 2011), 59 teachers were trained, 38 of whom taught Grade 1 the previous year and moved to Grade 2 with their class, so received refresher training; and 21 of whom were new to the project and so were provided with the initial and refresher training. Teacher attendance rates at the three training workshops were 95.2%, 98.4% and 96.3%, respectively.

The response rate of teachers to the weekly text messages containing a question averaged 87% over 37 weeks in Year 1 and 84% in Year 2. Two text messages did not contain a question and had a lower response rate of less than 50%.

An additional assessment of intervention fidelity came from weekly summary sheets completed by teachers with descriptions of and reflections on lessons taught. They were collected during unannounced health screening visits to reduce social desirability bias. Completion of sheets was time-consuming for teachers so we limited their collection to the first year (26

Table 3. Baseline school, child, and household characteristics of 2,539 Kenyan Grade-1 children enrolled in the trial in the 51 intervention and 50 control schools.

Characteristic; *n* (%)[a]		Intervention 51 schools	Control 50 schools
School characteristics[b]			
Exam score	Mean (*SD*)	221.9 (29.0)	227.3 (27.6)
School size	Median (IQR) [min, max]	511 (343, 686) [85, 4,891]	600 (371, 900) [199, 1,439]
School programs	Feeding	20 (39.2)	29 (58.0)
	Deworming	50 (98.0)	49 (98.0)
	Malaria control	6 (12.0)	15 (30.0)
School facilities	Water and sanitation	11 (22.4)	14 (28.0)
	Gender-separated toilets	49 (96.1)	49 (98.0)
	Hand-washing facilities	17 (33.3)	9 (18.0)
Child characteristics[b]		1,258 children	1,281 children
Age	Mean (*SD*)	7.7 (1.7)	7.9 (1.7)
	5–6	305 (24.2)	287 (22.4)
	7–8	573 (45.6)	525 (41.0)
	9–10	322 (25.6)	397 (31.0)
	11–15	58 (4.6)	72 (5.6)
Sex	Male	656 (52.2)	637 (49.7)
Nutritional status	Underweight	235 (24.3)	261 (27.7)
	Stunted	270 (23.0)	314 (27.0)
	Thin	225 (19.2)	238 (20.5)
School experience	Attended school before Grade 1	945 (95.5)	936 (95.8)
	Failed a grade	297 (30.5)	295 (30.9)
	Reads aloud in class	869 (85.9)	861 (87.1)
Household characteristics[b]			
Parental education	No schooling	363 (29.1)	435 (34.4)
	Primary schooling	692 (55.5)	667 (52.7)
	Secondary schooling	145 (11.6)	133 (10.5)
	Higher education	47 (3.8)	30 (2.4)
Socioeconomic status	Poorest	240 (19.2)	338 (26.5)
	Poor	249 (19.8)	268 (21.0)
	Median	266 (21.2)	222 (17.4)
	Less poor	250 (19.9)	235 (18.4)
	Least poor	250 (19.9)	213 (16.7)
Household size	1–5	365 (29.5)	370 (29.0)
	6–9	730 (59.0)	735 (57.6)
	10–31	142 (11.5)	170 (13.3)
Language spoken at home	Digo	651 (51.7)	524 (40.9)
	Duruma	170 (13.5)	380 (29.7)
	Other Mijikenda	28 (2.2)	28 (2.2)
	Swahili	197 (15.7)	171 (13.4)
	Other (non-Mijikenda)	212 (16.9)	177 (13.8)
No. times parent read to child last week	0	280 (33.0)	281 (35.7)
	1–3	400 (47.1)	338 (43.0)
	4–6	79 (9.3)	97 (12.3)
	7 and above	90 (10.6)	71 (9.0)

[a] Percentage of nonmissing children in each study arm presented for categorized data. For continuous data mean (*SD*) [min, max] is presented.

[b] All characteristics have less than 2% missing data with the exception of the following indicators: stunted, thin, and underweight.

weeks) of the intervention. During this time the mean number of HALI lessons taught by the 62 teachers was 54.6 lessons, on average two per week. These lessons were in addition to the regular language lessons the teacher may have taught.

As a measure of compliance with the intervention we recorded the presence of two instructional aids that teachers were requested to use in the classroom: a “washing line” for

Table 4. Baseline educational outcomes of 2,539 Kenyan Grade-1 children enrolled in the trial in the 51 intervention and 50 control schools.

	Mean (*SD*)/median (25th percentile, 75th percentile)	
Outcome (score range or scale)	Intervention *N*=1,258	Control *N*=1,281
Primary outcomes		
Spelling (0–20)	8.4 (4.6) /8.0 (5.0, 12.0)	7.8 (4.3) /7.0 (5.0, 11.0)
English letter knowledge (lpm)	16.2 (15.0) /13.0 (2.0, 28.0)	16.6 (15.1) /13.5 (3.0, 28.0)
Swahili letter sounds (lpm)	7.5 (11.6) /0.0 (0.0, 11.5)	5.2 (9.0) /0.0 (0.0, 8.0)
Swahili word identification (wpm)	1.9 (4.4) /0.0 (0.0, 1.0)	1.5 (4.0) /0.0 (0.0, 1.0)
Secondary outcomes		
Literacy		
English word identification (wpm)	1.5 (3.6) /0.0 (0.0, 1.0)	1.0 (3.0) /0.0 (0.0, 0.0)
English passage reading		
Fluency (wpm)	0.6 (3.7) /0.0 (0.0, 0.0)	0.4 (2.8) /0.0 (0.0, 0.0)
Comprehension (0–5)	0.0 (0.3) /0.0 (0.0, 0.0)	0.0 (0.3) /0.0 (0.0, 0.0)
Swahili passage reading		
Fluency (wpm)	0.6 (3.0) /0.0 (0.0, 0.0)	0.5 (2.7) /0.0 (0.0, 0.0)
Comprehension (0–5)	0.1 (0.3) /0.0 (0.0, 0.0)	0.0 (0.3) /0.0 (0.0, 0.0)
Beginning sounds (0–10)	5.4 (2.4) /5.0 (4.0, 7.0)	5.2 (2.4) /5.0 (3.0, 7.0)
Receptive language (0–25)	18.3 (3.9) /19.0 (16.0, 21.0)	17.8 (4.2) /19.0 (16.0, 21.0)
Nonliteracy		
Nonverbal reasoning (0–22)$^{\Delta}$	7.2 (2.4) /7.0 (6.0, 8.0)	7.7 (2.7) /7.0 (6.0, 9.0)
Numeracy (0–30) $^{\wedge}$	5.8 (4.6) /4.0 (3.0, 9.0)	5.3 (4.3) /4.0 (2.0, 8.0)
Sustained attention (0–20)*	12.1 (6.5) /14.0 (7.0, 18.0)	11.9 (6.7) /14.0 (7.0, 18.0)

Notes. SD: standard deviation; lpm: letters per minute; wpm: words per minute; $^{\Delta}$Raven's matrices test scored 0–22 at baseline and 9 months, scored 0–12 at 24 months; $^{\wedge}$sum of number identification (0–20) and quantity discrimination (0–10) at baseline and 9 months; written numeracy at 24 months; *pencil tap at baseline; single-digit code transmission at 9 and 24 months. Missing data at most 2.7% for each variable in both arms, except for pencil tap with 4.0% and 2.5% missing in control and intervention arms, respectively.

letters and words and a pocket chart for letters. These were selected because these aids were foundational to many of the lesson plans and their use implies minimal compliance.

We observed the washing line for letters in 46 of 50 intervention-school classrooms and in 18 out of 49 control-school classrooms. The pocket chart was present in 46 of 50 intervention classrooms and 4 of 49 control classrooms. Observation data were missing from one intervention and one control school.

Contamination

In order to interpret evaluations of education interventions, it is important to document contamination carefully (Keogh-Brown et al., 2007). To this aim, we recorded the level of contact between teachers in neighboring schools. Each teacher was asked to list up to four teachers in other schools with whom they had had contact in the previous month. The most common kind of contact was between one intervention school and another. Twenty-seven of the 51 intervention schools had teachers who said they had spoken with teachers from other intervention schools compared with just ten whose teachers had spoken with teachers from control schools. Control-school teachers were less likely than intervention-school teachers to have contact with other schools (25 control schools compared with 37 intervention schools), but where they did, they were also more likely to contact intervention-school teachers than control-school teachers (teachers in 15 control schools contacted intervention-school teachers compared with 10 control schools who contacted other control schools). Around two thirds of the conversations between teachers from different schools were about teaching methods, with a slightly higher percentage of conversations between two intervention-school teachers being about teaching methods (78%) compared to conversations involving control schools (65%). When asked to list local education projects, teachers in only six of the 50 control schools mentioned HALI and none said they had participated in the program. On balance, contamination in this study is likely to be relatively limited. Only 10 control schools reported discussing teaching methods with a teacher from an intervention school and only six mentioned the HALI project by name.

Effect of Literacy Intervention on Literacy Outcomes

Primary Outcomes

Impact analyses are presented in Table 5. At the 9-month follow-up, children in the literacy intervention arm had significantly higher mean-adjusted scores for the spelling task than children from control schools (Adjusted mean difference [Adj. MD]: 1.52, 95% CI: 0.98, 2.07, $p < 0.001$), with a moderate effect size (0.36, 95% CI: 0.23, 0.48). This gain was not sustained at 24-month follow-up (Adj. MD: 0.49, 95% CI: −0.06, 1.04, $p = 0.079$) with a small effect size (0.14, 95% CI: −0.02, 0.30).

At the 9-month follow-up, children in the literacy intervention arm scored significantly higher on assessment of Swahili letter sounds, with a greater than five-point mean difference between the intervention and control arm (Adj. MD: 5.59, 95% CI: 2.61, 8.57, $p < 0.001$). A similar, slightly smaller benefit was observed at 24 months (Adj. MD: 4.95, 95% CI: 1.95, 7.95, $p = 0.001$). Corresponding adjusted effect sizes were large at 0.64 (95% CI: 0.30, 0.97) and 0.38 (95% CI: 0.15, 0.61). However, at both 9- and 24-month follow-ups, no statistical difference in mean score was observed for English letter knowledge.

The impact on Swahili word identification was reasonable after 9 months (prespecified as a secondary outcome at this time point; Adj. MD: 1.88 wpm, 95% CI: −0.01, 3.76, $p = 0.051$) and of a larger magnitude at 24 months (prespecified as a primary outcome at this time point; Adj. MD: 2.44, 95% CI: 0.53, 4.35, $p = 0.012$). By contrast, the corresponding magnitude of adjusted effect sizes reduced over time being 0.23 (95% CI: −0.00, 0.47) at 9 months and 0.13 at 24 months (95% CI: 0.03, 0.24).

The magnitude of clustering was comparable in all four primary outcome models, therefore we focus on one outcome—spelling. We chose the spelling outcome because it covers literacy skills from basic to advanced and does not have ceiling or floor effects. With this outcome, most clustering was due to the two repeated measures on the child (ICC = 0.445), followed by clustering by school (ICC = 0.104), with negligible clustering by TAC tutor zone (ICC = 0.008), which was the unit of randomization.

Secondary Outcomes

For English word reading, a similar pattern was observed as for Swahili word reading with no evidence of an effect after 9 months and a benefit at 24 months (Adj. MD: 2.44 wpm, 95% CI: 0.53, 4.35, $p = 0.039$). For passage reading, the intervention led to a small improvement in oral reading fluency in both Swahili and English, which was statistically significant after 24 months (Adj. MD: 1.70 wpm, 95% CI: 0.08, 3.31, $p = 0.022$, for English, and, Adj. MD: 0.14 wpm, 95% CI: 0.03, 0.25, $p = 0.011$ for Swahili). In both cases, effect sizes were smaller at 24 months, as a result of larger control *SD* at the later time point, which was used to standardize the adjusted mean difference. There was a suggestion of some improvement

Table 5. Impact of HALI literacy intervention on literacy and nonliteracy outcomes for 2,491 and 2,470 Grade-1 HALI children at 9-month and 24-month follow-up, respectively.

	Intervention		Control					ICC[~]		
Follow-up time point	*n*	Mean (*SD*)	*n*	Mean (*SD*)	Adjusted mean difference[§] (95% CI)	Standardized adjusted mean difference[§§] (95% CI)	*p* value[∨]	TAC tutor zone	School	Child
Primary outcomes										
Spelling (score: 0–20)										
9 months	1,089	11.97 (4.77)	1,104	10.19 (4.29)	1.52 (0.98, 2.07)	0.36 (0.23, 0.48)	<0.001	0.008	0.104	0.445
24 months	1,006	11.89 (3.15)	984	11.14 (3.46)	0.49 (−0.06, 1.04)	0.14 (−0.02, 0.30)	0.079			
English letter knowledge (lpm)										
9 months	1,096	22.61 (16.58)	1,112	22.60 (16.64)	0.47 (−1.92, 2.86)	0.03 (−0.12, 0.17)	0.698	0.012	0.093	0.406
24 months	1,005	33.28 (18.99)	987	33.58 (19.21)	0.11 (−2.31, 2.53)	0.01 (−0.12, 0.13)	0.929			
Swahili letter sounds (lpm)										
9 months	1,095	10.40 (13.12)	1,112	4.83 (8.84)	5.59 (2.61, 8.57)	0.64 (0.30, 0.97)	<0.001	0.043	0.201	0.326
24 months	998	11.46 (15.92)	966	6.64 (13.17)	4.95 (1.95, 7.95)	0.38 (0.15, 0.61)	0.001			
Swahili word identification (wpm)[#]										
9 months	1,095	7.14 (10.45)	1,113	5.00 (8.07)	1.88 (−0.01, 3.76)	0.23 (−0.00, 0.47)	0.051	0.011	0.067	0.372
24 months	996	20.74 (18.43)	968	18.04 (18.41)	2.44 (0.53, 4.35)	0.13 (0.03, 0.24)	0.012			
Secondary outcomes										
Literacy outcomes										
English word identification (wpm)										
9 months	1,094	5.48 (8.04)	1,113	3.51 (6.45)	1.26 (−0.33, 2.86)	0.20 (−0.05, 0.44)	0.121	0.017	0.08	0.376
24 months	1,007	14.13 (15.11)	985	11.70 (14.14)	1.70 (0.08, 3.31)	0.12 (0.01, 0.23)	0.039			
English passage reading fluency (wpm)										
9 months	1,096	4.57 (10.99)	1,113	2.64 (8.04)	1.95 (−0.40, 4.29)	0.24 (−0.05, 0.54)	0.104	0.013	0.08	0.372
24 months	1,010	17.01 (20.91)	987	14.23 (19.91)	2.78 (0.40, 5.15)	0.14 (0.02, 0.26)	0.022			
English passage reading comprehension (score: 0–5)										
9 months	1,096	0.31 (0.82)	1,113	0.20 (0.69)	0.11 (0.01, 0.21)	0.16 (0.02, 0.30)	0.025	0.003	0.062	0.313
24 months	1,010	0.36 (0.82)	992	0.28 (0.74)	0.09 (−0.01, 0.19)	0.12 (−0.01, 0.25)	0.079			
Swahili passage reading fluency (wpm)										
9 months	1,096	4.43 (9.20)	1,112	2.97 (7.22)	1.58 (−0.24, 3.41)	0.22 (−0.03, 0.47)	0.089	0.011	0.065	0.37
24 months	990	16.82 (17.57)	967	14.50 (16.99)	2.42 (0.56, 4.27)	0.14 (0.03, 0.25)	0.011			

Swahili passage reading comprehension (score: 0–5)										
9 months	1,097	0.40 (0.94)	1,112	0.29 (0.80)	0.12 (−0.01, 0.26)	0.15 (−0.02, 0.33)	0.079	0.013	0.062	0.401
24 months	992	0.82 (1.12)	968	0.70 (1.07)	0.14 (−0.01, 0.28)	0.13 (0.00, 0.26)	0.059			
Beginning sounds (score: 0–10)										
9 months	1,098	6.66 (2.38)	1,113	6.32 (2.44)	0.34 (−0.08, 0.76)	0.14 (−0.03, 0.31)	0.116	0.019	0.122	N/A‡
24 months†	N/A	—	N/A	—	—	—	—	—	—	—
Receptive language (score: 0–25)										
9 months	1,093	20.35 (2.91)	1,107	19.75 (3.16)	0.62 (0.07, 1.17)	0.19 (0.02, 0.37)	0.028	0.028	0.145	N/A‡
24 months†	N/A	—	N/A	—	—	—	—	—	—	—
Nonliteracy outcomes										
Nonverbal reasoning—Raven's matrices test										
9 months (score: 0–22)	1,097	7.80 (2.41)	1,116	7.76 (2.43)	0.13 (−0.1, 0.35)	0.05 (−0.04, 0.15)	0.273	<0.001	0.036	0.099
24 months (score: 0–12)	1,008	4.13 (1.57)	994	4.11 (1.61)	0.09 (−0.15, 0.32)	0.05 (−0.09, 0.2)	0.473			
Numeracy (score: 0–20)ˆ										
9 months	1,096	9.14 (6.04)	1,113	8.59 (5.78)	0.43 (−0.03, 0.89)	0.07 (−0.01, 0.15)	0.069	<0.001	0.036	0.226
24 months	1,008	5.54 (2.83)	990	5.79 (3.16)	−0.46 (−0.93, 0.01)	−0.15 (−0.30, 0.00)	0.058			
Sustained attention (score: 0–20)*										
9 months	1,081	8.53 (3.47)	1,093	8.41 (3.94)	0.13 (−0.35, 0.62)	0.03 (−0.09, 0.16)	0.588	0.003	0.032	0.141
24 months	942	13.43 (4.95)	893	13.37 (4.99)	0.06 (−0.45, 0.57)	0.01 (−0.09, 0.11)	0.819			

Notes. CI: confidence interval; *SD*: standard deviation; lpm: letters per minute; wpm: words per minute; N/A: not applicable

§model-estimated mean difference using mixed-effects regression model with indicator terms for follow-up time (24 months vs. 9 months), study arm (intervention vs. control), and their interaction, adjusted for baseline measure of outcome (or proxy where not available, and as noted), age and sex, and for design features (malaria intervention arm, school mean exam score, with random effects for TAC tutor zone, for school and for child to account for the repeated measures on each child).

§§Effect size, that is, the standardized mean difference, is the model-estimated mean difference from the adjacent column divided by the *SD* in the control group at the follow-up time point;

∨*p* value for intervention effect at each follow-up time point;

∼ICC: intra-cluster correlation coefficient estimated from the model;

#prespecified as a primary outcome only at the 24-month follow-up time point;

†not measured at 24 months;

‡no random effect for child as outcome measured only at 9 months;

Δscored 0–22 at 9 months and scored 0–12 at 24 months;

ˆnumeracy score as sum of number identification (0–10) and quantity discrimination (0–20) at baseline and 9 months and as written numeracy at 24 months;

*pencil tap at baseline, single-digit code transmission at 9 months and 24 months.

in passage reading comprehension (scored 0 to 5) at each time point and in both languages, although only significant in English after 9 months (Adj. MD: 0.11, 95% CI: 0.01, 0.21, $p = 0.025$).

For the two remaining literacy outcomes of beginning sounds and receptive language, both measured only at baseline and 9 months, there was evidence of a benefit of the intervention on receptive language only (Adj. MD: 0.62, 95% CI: 0.07, 1.17, $p = 0.028$). There was no evidence of benefit of the intervention on any of the nonliteracy outcomes of nonverbal reasoning, numeracy, and sustained attention at either follow-up time point. In fact, the direction of effect on numeracy score at 24 months was negative (Adj. MD: −0.46, 95% CI: −0.93, 0.01, $p = 0.058$).

In summary, of the primary endpoints specified we found a significant impact of the intervention on spelling, Swahili letter sounds, and Swahili word reading (a primary endpoint at 24 months only). There was no effect on English letter knowledge.

Baseline Predictors of Nonresponse by Follow-up Time Point

There was no significant relationship between study arm and nonresponse to the educational survey at either follow-up time point. At 9 months, 50.2% of nonresponders and 49.5% of responders were in the intervention arm ($p = 0.58$). At 24 months, 46.8% of nonresponders and 50.2% of responders were in the intervention arm ($p = 0.29$). Baseline predictors of nonresponse were slightly different at each follow-up. Children unavailable at 9-month follow-up were more likely to be children whose parents had no schooling (36.4% vs. 31.2%, $p = 0.04$), with smaller household size (36.2% vs. 28.3% with 1–5 household members, $p = 0.007$) and with lower baseline receptive language ability (mean (*SD*) of 17.6 (4.2) vs. 18.1 (4.1), $p = 0.053$). Children unavailable at 24-month follow-up were slightly older at baseline (mean (*SD*) age of 7.9 (1.8) vs. 7.8 (1.7) years, $p = 0.072$), from a higher SES household (with 20.6% vs. 17.7% in the highest SES quintile and 20.8% vs. 23.4% in the lowest SES quintile, $p = 0.05$), and of a household less commonly of the predominant Digo language (41.8% vs. 47.4%, $p = 0.004$) (see Tables S1 and S2 in the online supplemental material). In summary, the following six baseline predictors were predictive of nonresponse at either FU1 or FU2: parental education, household size, baseline literacy ability, age, SES, and home language. In sensitivity analyses that accounted for nonresponse, we adjusted for all six predictors in the joint model for FU1 and FU2 (see results below).

Sensitivity Analyses

As prespecified in the analysis plan, we perform two sets of sensitivity analyses to understand how our results would change if some of our assumptions were not valid. First we adjusted the primary model for additional prespecified baseline variables (SES, language spoken at home, and preschool attendance) that we anticipated to be predictive of outcomes. Overall, attenuation of effects was observed, but few marked differences in magnitude or significance of effects and no marked differences for the primary outcomes were observed (see Table S3 in the online supplemental material). In our second set of sensitivity analyses we assessed whether our results altered when we accounted for six baseline predictors of nonresponse. To do this, we added the five additional baseline predictors of nonresponse to the primary model (age was already included in the primary model) to account for the 11.9%

nonresponse at 9 months and the 20% nonresponse at 24 months (Tables S1 and S2). This adjustment did not markedly alter results nor affect our conclusions on the primary outcomes (Table S3). For example, for spelling scores, Adj. MD of 1.52 (95% CI: 0.98, 2.07, $p <$ 0.001), 1.47 (95% CI: 0.95, 1.99, $p < 0.001$) and 1.39 (95% CI: 0.90, 1.88, $p < 0.001$) were estimated at 9 months for the primary model, for the fully adjusted model (with additional adjustment for SES, language spoken at home, and preschool attendance) and for the nonresponse model (with additional adjustment for parental education, household size, baseline literacy ability, SES, and home language), respectively. For secondary outcomes, marked changes were observed in reported significance due to adjustment for nonresponse for English words at 24 months (Adj. MD: 1.39, 95% CI: −0.04, 2.82, $p = 0.057$) and for English comprehension at 24 months (Adj. MD: 0.09, 95% CI: −0.01, 0.25, $p = 0.053$).

Subgroup Analyses

There was no strong evidence of heterogeneity of intervention effect on literacy outcomes for the prespecified subgroups except for gender. As a result, the only subgroup analyses we report are by gender. We include all literacy outcomes, even those with no evidence of a main effect (Table 6). For the two primary outcomes with evidence of a main effect at 9 months (spelling and Swahili letter sounds), there was a gender subgroup effect for spelling only ($p = 0.003$) with a larger beneficial effect of the literacy intervention in girls (Adj. MD: 1.93, 95% CI: 1.32, 2.54, $p < 0.001$) than in boys (Adj. MD: 1.12, 95% CI: 0.52, 1.72, $p <$ 0.001). There were no gender subgroup effects for any of the other primary outcomes at 9 months or at 24 months, although the estimated effect was consistently larger in females than in males for all four outcomes at 24 months. No subgroup effects were observed for secondary literacy outcomes at 9 months. At 24 months, greater beneficial effects were observed for girls in English word identification ($p < 0.001$), English passage-reading fluency ($p =$ 0.023) and Swahili passage-reading fluency ($p = 0.032$).

Impact on Teaching and Class Behaviors: Classroom Observations

Table 7 shows the percentage of time each type of classroom observation was recorded over the two observed classes. The table also shows model-based effect sizes reported in terms of standardized coefficients (standard deviations). These estimates were modeled controlling for teacher characteristics—teacher language, years of experience teaching, and education level—and were modeled at the school level.

As Table 7 shows, the intervention led to much more teaching with written text (effect size=1.15) at the expense of oral presentation (effect size = −0.77) or illustrations and visual materials that contain no visual text (effect size = −0.62). There was a concomitant increase in reading among students (effect size = 0.69) and a reduction in oral response (effect size = −0.75). Students also spent less time writing (effect size = −0.54), which largely meant less time copying whole words from the board, and more time manipulating objects (effect size = 0.67), presumably as a result of the letter and word cards developed as part of the project. In terms of the language part that was the focus of instruction, more time was spend focusing on letters (effect size = 1.08) and sounds (effect size = 0.57) at the expense of sentences (effect size = −0.62).

Table 6. Subgroup analyses: Impact of HALI literacy intervention on literacy outcomes for 2,491 and 2,470 Grade-1 HALI children at 9-month and 24-month follow-up, respectively by gender.

		Baseline	9 months					24 months				
	n	Mean (*SD*)	Mean (*SD*)		Standardized adjusted mean difference§§	Within-subgroup *p* value†	Between subgroup *p* value‡	Mean (*SD*)		Standardized adjusted mean difference§§	Within-subgroup *p* value†	Between subgroup *p* value‡
Gender			Int	Control	(95% CI)			Int	Control	(95% CI)		
Primary outcomes												
Spelling (score: 0–20)												
Female	1,205	8.08 (4.51)	12.18 (4.59)	10.02 (4.25)	0.45 (0.31, 0.60)	<0.001	0.003	12.24 (2.89)	11.20 (3.25)	0.23 (0.04, 0.42)	0.017	0.32
Male	1,262	8.13 (4.46)	11.77 (4.92)	10.37 (4.32)	0.26 (0.12, 0.40)	<0.001		11.58 (3.33)	11.08 (3.67)	0.07 (−0.10, 0.24)	0.408	
English letter knowledge (lpm)												
Female	1,214	16.46 (15.07)	23.12 (16.14)	23.10 (16.95)	−0.01 (−0.17, 0.15)	0.918	0.329	35.08 (18.58)	34.19 (18.62)	0.04 (−0.11, 0.18)	0.612	0.108
Male	1,255	16.35 (15.00)	22.14 (16.98)	22.10 (16.31)	0.06 (−0.10, 0.23)	0.446		31.67 (19.23)	32.97 (19.78)	−0.02 (−0.16, 0.12)	0.771	
Swahili letter sounds (lpm)												
Female	1,213	6.37 (10.59)	10.50 (13.53)	4.94 (9.12)	0.60 (0.26, 0.95)	0.001	0.911	12.80 (16.78)	7.32 (14.11)	0.38 (0.16, 0.61)	0.001	0.455
Male	1,256	6.33 (10.32)	10.31 (12.74)	4.72 (8.55)	0.66 (0.29, 1.02)	<0.001		10.26 (15.01)	5.97 (12.15)	0.38 (0.12, 0.63)	0.005	
Swahili word identification (wpm)												
Female	1,212	1.26 (3.36)	5.39 (8.02)	3.57 (6.69)	0.24 (−0.02, 0.50)	0.076	0.853	15.37 (15.75)	11.38 (12.80)	0.19 (0.08, 0.31)	0.001	0.111
Male	1,254	1.31 (3.28)	5.56 (8.06)	3.44 (6.21)	0.22 (−0.05, 0.50)	0.107		13.03 (14.44)	12.03 (15.37)	0.08 (−0.04, 0.20)	0.206	
Secondary literacy outcomes												
English word identification (wpm)												
Female	1,213	1.57 (3.94)	7.32 (10.88)	5.12 (8.29)	0.14 (−0.13, 0.40)	0.319	0.397	22.70 (18.99)	18.80 (18.53)	0.23 (0.09, 0.37)	0.002	0.001
Male	1,256	1.78 (4.47)	6.98 (10.06)	4.87 (7.85)	0.26 (−0.03, 0.54)	0.081		19.01 (17.75)	17.27 (18.27)	0.04 (−0.08, 0.15)	0.544	
English passage reading fluency (wpm)												
Female	1,213	0.46 (3.43)	4.53 (11.20)	2.59 (8.05)	0.23 (−0.10, 0.56)	0.179	0.854	18.54 (21.40)	14.02 (19.21)	0.23 (0.09, 0.37)	0.001	0.023
Male	1,256	0.47 (3.18)	4.61 (10.81)	2.70 (8.05)	0.25 (−0.07, 0.58)	0.128		15.66 (20.39)	14.44 (20.61)	0.06 (−0.07, 0.19)	0.343	
English passage reading comprehension (score: 0–5)												
Female	1,213	0.04 (0.33)	0.27 (0.74)	0.19 (0.66)	0.15 (−0.02, 0.32)	0.093	0.698	0.36 (0.81)	0.28 (0.74)	0.12 (−0.04, 0.28)	0.142	0.779
Male	1,256	0.03 (0.25)	0.34 (0.88)	0.22 (0.73)	0.17 (0.01, 0.32)	0.035		0.35 (0.83)	0.27 (0.74)	0.12 (−0.03, 0.28)	0.125	
Swahili passage reading fluency (wpm)												
Female	1,213	0.57 (3.04)	4.47 (9.56)	3.03 (7.44)	0.19 (−0.10, 0.47)	0.195	0.745	18.78 (18.70)	15.10 (17.45)	0.21 (0.08, 0.33)	0.001	0.032

Male	1,256	0.57 (2.66)	4.40 (8.87)	2.91 (7.01)	0.25 (−0.05, 0.54)	0.104		15.08 (16.33)	13.90 (16.51)	0.08 (−0.05, 0.21)	0.22	
Swahili passage reading comprehension (score: 0–5)												
Female	1,213	0.04 (0.32)	0.39 (0.95)	0.28 (0.80)	0.17 (−0.03, 0.37)	0.097	0.763	0.91 (1.21)	0.71 (1.08)	0.21 (0.06, 0.36)	0.007	0.11
Male	1,256	0.05 (0.32)	0.40 (0.92)	0.30 (0.81)	0.14 (−0.06, 0.33)	0.165		0.74 (1.04)	0.68 (1.05)	0.05 (−0.10, 0.21)	0.493	

Notes. CI: confidence interval; lpm: letters per minute; wpm: words per minute;§model-estimated mean difference using mixed-effects regression model with indicator terms for follow-up time (24 months vs. 9 months), study arm (intervention vs. control), gender (male vs. female) and all three pairwise and single three-way interaction terms, adjusted for baseline measure of outcome (or proxy where not available, and as noted), age and sex, and for design features (malaria intervention group, school mean exam score, with random effects for TAC tutor zone, for school and for child to account for the repeated measures on each child).

§§Effect size, that is, the standardized mean difference, is the model-estimated mean difference from the adjacent column divided by the *SD* in the control arm at the follow-up time point;

#prespecified as a primary outcome only at the 24-month follow-up time point;

†*p* value for intervention effect within each level of the subgroup at each follow-up time point (i.e., separately for males and for females);

‡heterogeneity *p* value to test for any subgroup effect of gender at the specific follow-up time point.

Table 7. Impact of the literacy intervention on observed classroom behavior.

Materials of instruction	Literacy intervention (n=51)	Control (n=50)	Difference	Standardized adjusted effect size
Written	67.5%	53.0%	14.5%	1.15***
Oral	47.8%	52.7%	−5.0%	−0.77***
Illustration	6.8%	13.5%	−6.8%	−0.62**
Management	5.3%	5.3%	0.0%	0.06
Student behavior				
Read	39.1%	28.7%	10.4%	0.75***
Write	22.4%	28.9%	−6.5%	−0.54**
Listen	27.8%	29.0%	−1.2%	−0.04
Manipulate	13.3%	8.7%	4.5%	0.67***
Oral	20.5%	27.6%	−7.1%	−0.75***
Physical response	9.8%	11.5%	−1.6%	−0.21
Management	14.5%	15.6%	−1.1%	−0.13
Initiate	8.1%	8.5%	−0.4%	−0.08
Not engaged	8.7%	9.3%	−0.6%	−0.08
Language part				
Sounds	1.1%	0.2%	0.9%	0.57**
Letters	4.3%	1.3%	3.0%	1.08***
Word part	4.9%	3.9%	1.1%	0.23
Word	40.7%	40.5%	0.2%	0.09
Sentence	17.4%	28.4%	−11.0%	−0.62**
Story, song, poem, passage	9.4%	5.0%	4.4%	0.33†
Sample size=103				

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

School Dropout

We followed up with headteachers to find out what had happened to children not present in the school during the assessments. We derived a measure of dropout from their responses with a conservative definition that excluded children who were chronic absentees, were sick, had moved to another school, or were absent because of nonpayment of fees. In the literacy intervention arm, 27 students (2.1%) dropped out in their first two years of school compared with 68 students (5.3%) in the control arm. The results of logistic regression analysis indicate that the intervention had a significant impact on school dropout (Odds Ratio [OR] = 0.43, 95% CI: 0.25, 0.74, $p = .002$). Thus, the fitted odds that an intervention arm member would drop out were less than one half the odds that a control arm member would do the same during the study period.

Cost of the Literacy Intervention

We estimated the cost of the literacy intervention for a typical Kenya district with 62 teachers and reaching 3,844 children (including children who received the intervention but were not included in the evaluation), based on empirical costs collected in the study. (For additional information on costs, see Dubeck et al., 2015.) The total cost of the modeled district-level program was US$32,940 (Table 8) or US$531 per teacher and US$8.57 per child. Direct financial costs comprised 76% of the total cost.

Table 9 presents the cost breakdown by intervention component and resource type. We can see that three main intervention component contributors to cost were (a) the initial

Table 8. Summary of total, direct and indirect cost (US$2010).

	Total cost	Direct cost[a]	Indirect cost[b]
District-level program	32,940	25,049	7,907
Per teacher	531	404	128
Per child	8.57	6.52	2.06
%		76	24

[a] Direct costs includes all financial expenditure.
[b] Indirect costs include the opportunity cost of teacher and ministry officials during training and program support.

training (32.4%), (b) the teacher materials (28.6%), and (c) the SMS support (20.4%). Consumables were the greatest driver of costs (53.7%).

Acceptability of the Literacy Intervention From the Teachers' Perspective

During the follow-up trainings, facilitators led a combination of small focus group discussions of 6 to 12 people and individual interviews. We found that classroom materials were well received and teachers found the weekly text messages to be a good source of new teaching ideas. Teacher perceptions are reported in more detail elsewhere (Dubeck et al., 2015).

Discussion

The main goal of the literacy intervention was to develop teachers' capacity to improve their students' reading achievement. The intervention had a moderate short-term impact on spelling at 9 months. The impact on Swahili letter sounds was large at 9 months and moderate after 24 months. The lack of effect on English letter knowledge may be because English letters, unlike Swahili letters, were already emphasized in schools. There was a moderate improvement in Swahili word reading after 24 months. By the end of the second year there was also a significant impact on English word reading and oral reading fluency in both Swahili and English. The effect sizes for the positive impact on these outcomes range from 0 to 0.64. The larger effect sizes compare favorably to a recent review (McEwan, 2014) of teacher training interventions in developing countries, the majority of which failed to have a significant impact on student outcomes, with only 2 of 77 interventions demonstrating an effect size greater than 0.5. However, more recent studies in Kenya (Piper et al., 2014) and elsewhere (Crouch, 2015, Crouch & DeStefano, 2015; Room to Read, 2016) have consistently

Table 9. Program costs by resource type and intervention component (US$2010).

	Intervention component							
Resource type	Manual	Teaching materials	SMS support	Initial training	Follow-up training	District admin	Total	%
Consumables	1,454	8,911	5,596	1,735	—	—	17,695	53.7
Personnel	35	195	1,005	3,785	1,954	1,078	8,052	24.4
Transport	—	330	—	942	942	—	2,215	6.7
Facility	16	—	107	4,206	349	300	4,979	15.1
TOTAL	1,504	9,437	6,707	10,668	3,246	1,378	32,940	
%	4.6	28.6	20.4	32.4	9.9	4.2		

found effect sizes at or above the 0.5 level. These studies have also recorded a greater impact on oral reading fluency than the 2–3 wpm improvement found in the current study.

The intervention had a larger impact on girls than boys for several literacy outcomes. Other evaluations have found similar advantages for girls (Brombacher, Stern, Nordstrum, Cummiskey, & Mulcahy-Dunn, 2015; Piper & Korda, 2011), and the finding is consistent with evidence of better literacy outcomes for girls across low- and middle-income countries (Chiu & McBride-Chang, 2006; Grant & Behrman, 2010).

The study found large effects on instructional focus and on the behavior of teachers and students in class. Teachers were more likely to use written material and to focus on letters and sounds rather than whole words or sentences. Students spent more time reading text. The effect sizes for these improvements were large, ranging from 0.57 to 1.15. This result is important because detailed classroom observations of this kind are rare and data are particularly lacking from experimental studies. The findings demonstrate that significant change in teacher behavior is possible with relatively little in-person support. The findings are also important for understanding the mechanism underlying the intervention. This is critical for assessing the external validity of our findings—for applying the lessons of this evaluation in other contexts (White, 2009). We argue that our approach should be applied where teachers currently place little emphasis on students interacting with written text and on breaking down words into component parts (Dubeck et al., 2012) and that changing these aspects of teacher behavior are critical for success.

The intervention also reduced school dropout from 5.3% to 2.1%. Related research carried out in the region (Zuilkowski, Jukes, & Dubeck, 2016) found that school quality was a primary driver of decisions to drop out of school. It seems likely that more children (and their families) in the intervention arm decided to continue with their schooling because they were making greater progress and saw greater potential to prosper in the future. This finding suggests that the problem of school dropout in Kenya could be tackled in part by improving education quality.

The results of the study broadly support several aspects of our approach to designing an effective literacy intervention. Our first consideration was to design the intervention around teaching strategies with a rigorous evidence base showing effectiveness in other contexts. Our preparatory analysis (Dubeck et al., 2012) showed that several evidence-based strategies were underutilized by teachers on the Kenya coast. There was insufficient explicit use of text in teaching and little focus on breaking words into components to support reading or spelling. Our classroom observations indicated that it was these aspects of pedagogy that were improved the most by the intervention. Interestingly, although the focus on oral language decreased as a result of the intervention, students' oral language skills, as measured by the Swahili vocabulary test, were significantly improved by the new teaching methods nevertheless.

The second consideration was to implement new instructional methods that built from teachers' prior experiences and made connections between current and proposed methods explicit in all materials. By contrast, attempts to introduce entirely new pedagogical methods do not gain acceptance by teachers and often fail to change behavior (O'Sullivan, 2004; Schweisfurth, 2011; Vavrus, 2009). The provision of materials was also critical to successful adoption of the new approaches. By and large, teachers did not emphasize the use of text in their instruction before the intervention because they were working with textbooks that also failed to emphasize the use of text.

A third consideration involved designing an intervention that could be replicated, scaled up, and adopted by the government. This was achieved through sustainable methods such as sourcing the intervention materials locally and including instruction on how to use those materials to improve beginning reading skills. A crucial component of a scalable intervention is cost, and this intervention's cost per child of US$8.57 appears relatively inexpensive compared with a range of other educational interventions (Evans & Ghosh, 2008). This compares with a cost of $4.42 per pupil ($2.21 per subject) for teaching of English and Swahili in the Primary Math and Reading (PRIMR) initiative in Kenya (Piper et al., 2014). The lower cost in the PRIMR intervention likely resulted from its operation at a larger scale than HALI in terms of number of classes, subjects, and schools and because PRIMR did not account for indirect costs, which represented about a quarter of costs in the HALI project. A key strategy to ensure effectiveness at low cost was the use of text messages to support teachers. On the whole, this strategy was successful. The response rate to text messages was high (an average of 87%) and teachers told us that they valued the support provided by these messages. They were successful in creating a sense of community, making teachers feel valued and listened to, and being an important mechanism for feedback and improvement of the intervention. This was possible because teachers were given the opportunity to respond to text messages, and responses were selected for communication to the rest of the group. This contrasts with many other programs that offer only one-way text communication.

It is perhaps remarkable that such a low-cost intervention should have had a significant impact on learning. Other research points to the importance of in-person coaching (Piper & Zuilkowski, 2015), and similar interventions (Piper et al., 2014) typically involve providing pupil books. The HALI intervention was able to improve student literacy by supplementing the existing government approach and by using text-message support and locally made materials rather than in-person coaching and printed books.

Teachers' perception of the intervention is also important for its scalability. In general, perceptions were positive. The high response rate to the weekly text message and feedback through self-report methods such as summary sheets and focus group discussions provided good insight into successful aspects of the intervention as well as aspects to improve on for the future. A key concern was the increased time taken to prepare and conduct the intervention lessons compared with the standard curriculum, but it was broadly recognized that the lessons were popular with the students in terms of increasing engagement and improving their literacy acquisition.

Teachers' concerns about the time involved in implementing the provided lessons perhaps helped explain another result of the study, that there was a borderline effect of reduced scores on tests of numeracy due to the intervention. It seems plausible that teachers spent more time teaching literacy at the expense of numeracy because of the intervention. The HALI intervention was designed as an academic study to evaluate the effectiveness of an evidence-based approach to literacy instruction in Kenya. To develop the intervention into a program to be implemented by the government would require more integration with and consideration of other subjects in the government curriculum.

Subsequent to the evaluation of the HALI project the Kenya government evaluated the PRIMR initiative—a pilot of a comprehensive approach to early-grade instruction in literacy and mathematics. The pilot adopted the approach of supporting teachers with text messages and also used instructional principles similar to the HALI project. On the basis of this successful pilot, the government has introduced a national literacy program called Tusome ("let's read") to approximately 21,600 public primary schools and 1,000 low-cost private schools in nonformal settlements with the aim to improve English and Swahili reading outcomes for 5.4 million Kenyan children in Class 1 and 2 by 2018.

Limitations

One potential limitation of this study is that the evaluation was conducted by the same team of people that designed the intervention. In order to promote objectivity of the evaluation we submitted the trial protocol for publication before the study began and submitted statistical analysis plans with prespecified primary outcomes before data were inspected, and these plans were scrutinized by an independent data-monitoring committee. As with all new projects, it is likely that the intervention was managed and implemented with greater expertise and enthusiasm by the project team than could be expected in subsequent iterations and scaling up of the intervention.

The study included four primary outcomes and ten secondary outcomes increasing the likelihood of Type I error, particularly for effects of borderline significance. The effects of the intervention on spelling at 9 months and Swahili word reading at both 9 and 24 months were highly significant ($p <=.001$) and less vulnerable to this concern.

The child response rate was 88.1% and 80.0% at the 9- and 24-month follow-up, respectively (Figure 2). Although this compares favorably with other studies in developing countries, it leaves open the possibility of bias introduced by differential nonresponse. Fortunately, there was no differential nonresponse by study arm, and sensitivity analyses that adjusted for baseline predictors of nonresponse suggest that nonresponse only minimally affected intervention effect estimates.

Contamination is a concern in school-based experiments. We tried to limit contamination by including neighboring schools from the same TAC tutor zone in the same intervention arm. We also assessed contamination through teacher interviews. Although a few control school teachers discussed teaching methods with HALI school teachers, the rate was less than 15% of teachers, and our estimation is that the effect on results was minimal. If the effect had been substantial, it would have resulted in reduced estimates of treatment effects.

Our method of observing classrooms through unannounced visits provides a limited account of teachers' behavior. Our conclusions about which teacher behaviors were affected by the intervention are still valid, but the effect sizes may not be representative of everyday life in the school. If observation made the teachers more conscientious, then the results we found may have overstated genuine effect sizes. On the other hand, if the unannounced visits took place when teachers were unprepared or the visits made them anxious or antagonistic, our findings could have understated genuine effect sizes.

In conclusion, this study demonstrates that evidence-based approaches to instruction can make a significant impact on the literacy development of Kenyan children in the early grades of primary school and reduced dropout from school. In particular, shifting instructional focus from oral language to text and from words and sentences to letters and sounds seems to be important for improving early literacy development. These approaches are already being implemented in Kenya. Their cost-effectiveness could perhaps be improved further by using text messages to support teachers in implementing new instructional approaches and through the use of locally made materials. Together, these approaches could form the basis of sustainable and affordable approaches to addressing the literacy crisis in sub-Saharan Africa.

Acknowledgments

We are grateful to all the teachers, children, and parents who participated in this study and to all the members of the survey teams for their tireless effort and excellent work. We acknowledge the support provided by the ministries of education and health at national and district levels and TAC tutors at the district level, and thank Carlos Mcharo, Juddy Kengo, Caroline Hagelskamp, Geetha Mathews, Chandana Jasti, Morgan Sheppard, Elizabeth Adelman, and Elizabeth Allen for their contributions and support. We would also like to extend our sincere thanks to the members of the World Bank's Development Impact Evaluation Initiative, the study advisory group (David Schellenberg, Greg Fegan, Don Bundy, and Diana Elbourne), and the data-monitoring committee (Feiko ter Kuile, Penelope Phillips-Howard, and Carole Torgerson).

Funding

This work was supported by grants from the William and Flora Hewlett Foundation, the International Initiative for Impact Evaluation, the Partnership for Child Development, and the Development Impact Evaluation Initiative as part of the Malaria Impact Evaluation Program of the World Bank. Simon J. Brooker is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (098045).

ARTICLE HISTORY

Received 24 October 2015
Revised 1 July 2016
Accepted 3 August 2016

References

Ackers, J., & Hardman, F. (2001). Classroom interaction in Kenyan primary schools. *Compare: A Journal of Comparative Education*, *31*(2), 245–261. doi:10.1080/03057920120053238

Aker, J. C., Ksoll, C., & Lybbert, T. J. (2012). Can mobile phones improve learning? Evidence from a field experiment in Niger. *American Economic Journal: Applied Economics*, *4*(4): 94–120.

Anderson−Levitt, K. M. (2004). Reading lessons in Guinea, France, and the United States: Local meanings or global culture? *Comparative Education Review*, *48*(3), 229–252.

Arnold, C., Bartlett, K., Gowani, S., & Merali, R. (2006). *Is everybody ready? Readiness, transition and continuity: Lessons, reflections and moving forward* [Paper commissioned for the EFA Global Monitoring Report 2007, Strong Foundations: Early Childhood Care and Education]. Paris: UNESCO.

Badian, N. A. (1988). The prediction of good and poor reading before kindergarten entry: A nine-year follow-up. *Journal of Learning Disabilities*, *21*(2), 98–103.

Beltramo, T., & Levine, D. (2012). *Do SMS text messaging and SMS community forums improve outcomes of adult and adolescent literacy programs?* Retrieved from http://escholarship.org/uc/item/1c31c2m4

Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C., & Domitrovich, C. E. (2008). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Development and Psychopathology*, *20*(3), 821–843.

Brombacher, A., Stern, J., Nordstrum, L., Cummiskey, C., & Mulcahy-Dunn, A. (2015). *National Early Grade Literacy and Numeracy Survey–Jordan: Intervention impact analysis report*. Retrieved from https://www.eddataglobal.org/documents/index.cfm/01_Jordan_Intervention_Final_Report_07_April_2015_English1.pdf?fuseaction=throwpub&ID=911

Brooker, S., Okello, G., Njagi, K., Dubeck, M. M., Halliday, K. E., Inyega, H., & Jukes, M. C. H. (2010). Improving educational achievement and anaemia of school children: Design of a cluster randomised trial of school-based malaria prevention and enhanced literacy instruction in Kenya. *Trials*, *11*, 93. doi:10.1186/1745-6215-11-93

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. *Scientific Studies of Reading*, *10*(4), 331–362.

Clarke, S. E., Jukes, M. C., Njagi, J. K., Khasakhala, L., Cundill, B., Otido, J., … Brooker, S. (2008). Effect of intermittent preventive treatment of malaria on health and education in schoolchildren: A cluster-randomised, double-blind, placebo-controlled trial. *Lancet*, *372*(9633), 127–138. doi:10.1016/S0140-6736(08)61034-X

Commeyras, M., & Inyega, H. N. (2007). An integrative review of teaching reading in Kenyan primary schools. *Reading Research Quarterly*, *42*(2), 258–281.

Crouch, L., & DeStefano, J. (2015). *A practical approach to in-country systems research* (RISE Working Paper). Washington, DC: RTI International. Retrieved from http://www.riseprogramme.org/sites/www.rise.ox.ac.uk/files/14_Crouch-DeStefano.pdf

Crouch, L., Korda, M., & Mumo, D. (2009). *Improvements in reading skills in Kenya: An experiment in the Malindi District*. Retrieved from http://datatopics.worldbank.org/hnp/files/edstats/KENdprep09.pdf

DFID. (2014). *Learning for all: DFID's education strategy 2010–2015*. London, England: Author.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do." *Developmental Psychobiology*, *29*, 315–334.

Dubeck, M. M., Jukes, M. C. H., Brooker, S. J., Drake, T. L., & Inyega, H. N. (2015). Designing a program of teacher professional development to support beginning reading acquisition in coastal Kenya. *International Journal of Educational Development*, *41*, 88–96.

Dubeck, M. M., Jukes, M. C. H., & Okello, G. (2012). Early primary literacy instruction in Kenya. *Comparative Education Review*, *56*(1), 48–68.

Evans, D. K., & Ghosh, A. (2008). Prioritizing educational investments in children in the developing world. Santa Monica, CA: RAND Corporation.

Filmer, D., & Prichett, L. (2001). Estimating wealth effects without expenditure data—or tears: An application to educational enrollment in states of India. *Demography*, *38*, 115–132.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.) Hoboken, NJ: John Wiley & Sons.

Funnell, S., & Rogers, P. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: Jossey-Bass.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Gove, A., & Cvelich, P. (2010). *Early reading: Igniting education for all*. Research Triangle Park, NC: Research Triangle International.

Grant, M. J., & Behrman, J. R. (2010). Gender gaps in educational attainment in less developed countries. *Population and Development Review*, *36*(1), 71–89.

Halliday, K. E., Okello, G., Turner, E. L., Njagi, K., Mcharo, C., Kengo, J., … Brooker, S. J. (2014). Impact of intermittent screening and treatment for malaria among school children in Kenya: A cluster randomised trial. *PLOS Medicine*, *11*(1). doi:10.1371/journal.pmed.1001594

Hardman, F., Abd-Kadir, J., Agg, C., Migwi, J., Ndambuku, J., & Smith, F. (2009). Changing pedagogical practice in Kenyan primary schools: The impact of school-based training. *Comparative Education*, *45*(1), 65–86.

Hayes, R. J., & Bennett, S. (1999). Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, *28*, 319–326.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. Cambridge, England: Cambridge University Press.

Hungi, N., Ngware, M., & Abuya, B. (2014). Examining the impact of age on literacy achievement among Grade 6 primary school pupils in Kenya. *International Journal of Educational Development*, *39*, 247–259.

IEA. (2014). *Number of the week (33.07%): Distribution of teachers across job groups 2014*. Retrieved from http://www.ieakenya.or.ke/blog/number-of-the-week-33-07-distribution-of-teachers-across-job-groups-2014

Invernizzi, M., Juel, C., Swank, L., & Meier, J. (2007). *Phonological awareness literacy screening (PALS): Kindergarten*. Charlottesville, VA: University of Virginia.

Jukes, M. C. H., Drake, L. J., & Bundy, D. A. P. (2008). *School health, nutrition and education for all: Levelling the playing field*. Wallingford, England: CABI Publishing.

Jukes, M. C. H., Vagh, S. B., & Kim, Y. S. (2006). *Development of assessments of reading ability and classroom behavior*. Washington, DC: World Bank.

Keogh-Brown, M., Bachmann, M., Shepstone, L., Hewitt, C., Howe, A., Ramsay, C., … Campbell, M. (2007). Contamination in trials of educational interventions. *Health Technology Assessment*, *11* (43), iii, ix–107.

Kibui, A. W. (2014). Language policy in Kenya and the New Constitution for Vision 2030. *International Journal of Educational Science and Research*, *4*(5), 89–98.

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (16th ed.). Dallas, TX: SIL International.

Manly, T., Robertson, I. H., Anderson, V., & Nimmo-Smith, I. (1999). *Test of everyday attention for children: TEA-Ch*. Bury St. Edmunds, England: Thames Valley Test Company.

McEwan, P. J. (2014). Improving learning in primary school of developing countries. *Review of Educational Research*. doi:10.3102/0034654314553127

Ministry of Education. (2006). *Primary education English handbook*. Nairobi: Kenya Institute of Education.

Mugo, J., Kaburu, A., Limboro, C., & Kimutai, A. (2011). *Are our children learning? Annual learning assessment report*. Nairobi, Kenya: Uwezo.

Ngware, M., Oketch, M., & Mutisya, M. (2014). Does teaching style explain differences in learner achievement in low and high performing schools in Kenya? *International Journal of Educational Development*, 36, 3–12.

O'Sullivan, M. (2004). The reconceptualisation of learner-centred approaches: A Namibian case study. *International Journal of Educational Development*, 24(6), 586–602.

Onsomu, E., Nzomo, J., & Obiero, C. (2005). The SACMEQ II Project in Kenya: A study of the conditions of schooling and the quality of education. Harare and Nairobi: SACMEQ and Kenya Ministry of Education, Science and Technology.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London, England: Sage.

Piper, B. (2010). *Kenya Early Grade Reading Assessment findings report*. Research Triangle Park, NC: RTI International and East African Development Consultants.

Piper, B., Jepkemei, E., Kwayumba, D., & Kibukho, K. (2015). Kenya's ICT policy in practice: The effectiveness of tablets and e-readers in improving student outcomes. *FIRE: Forum for International Research in Education*, *2*(1). Retrieved from http://preserve.lehigh.edu/cgi/viewcontent.cgi?article=1025&context=fire

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

Piper, B., & Miksic, E. (2011). Mother tongue and reading: Using early grade reading assessments to investigate language-of-instruction policy in East Africa. In A. Gove & A. Wetterberg (Eds.), *The early grade reading assessment: Application and intervention to improve basic literacy* (pp. 139–182). Research Triangle Park, NC: RTI Press.

Piper, B., & Mugenda, A. (2012). *The Primary Math and Reading (PRIMR) Initiative baseline report*. Research Triangle Park, NC: RTI International.

Piper, B., & Zuilkowski, S. S. (2015). Teacher coaching in Kenya: Examining instructional support in public and nonformal school. *Teaching and Teacher Education*, *47*, 173–183.

Piper, B., Zuilkowski, S. S., Kwayumba, D., & Strigel, C. (2016). Does technology improve classroom reading outcomes? Comparing the effectiveness and cost-effectiveness of ICT interventions for early grade literacy in Kenya. *International Journal of Educational Development*, *49*, 204–214.

Piper, B., Zuilkowski, S. S., & Mugenda, A. (2014). Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. *International Journal of Educational Development*, *37*, 11–21.

Pontefract, C., & Hardman, F. (2005). The discourse of classroom interaction in Kenyan primary schools. *Comparative Education*, *41*(1), 87–106.

Pressley, M. (2001). *Learning to read: Lessons from exemplary first-grade classrooms*. New York, NY: Guilford Press.

Raven, J. C., Styles, I., & Raven, M. A. (1998). *Raven's progressive matrices: CPM parallel test booklet*. Oxford, England: Oxford Psychologists Press.

Reubens, A. (2009). *Early Grade Mathematics Assessment (EGMA): A conceptual framework based on mathematics skills development in children*. Research Triangle Park, NC: RTI International.

Room to Read. (2016). *Literacy program reading skills results summary*. San Francisco, CA: Author.

RTI International. (2009). *Early Grade Reading Assessment toolkit*. Prepared for the World Bank, Office of Human Development. Research Triangle Park, NC: Author. Retrieved September 12, 2016, from https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=149

Sailors, M., Hoffman, J. V., Pearson, P. D., McClung, N., Shin, J., Phiri, L. M., & Saka, T. (2014). Supporting change in literacy instruction in Malawi. *Reading Research Quarterly*, *49*(2), 209–231.

Scanlon, D. M., Gelzheiser, L., Fanuele, D., Sweeney, J., & Newcomer, L. (2003). *Classroom Language Arts Systematic Sampling and Instructional Coding (CLASSIC)*. Albany, NY: Child Research and Study Center, The University at Albany.

Schweisfurth, M. (2011). Learner-centred education in developing country contexts: From solution to problem? *International Journal of Educational Development*, *31*, 425–432.

Sifuna, D. N. (2007). The challenge of increasing access and improving quality: An analysis of universal primary education interventions in Kenya and Tanzania since the 1970s. *International Review of Education*, *53*(5–6), 687–699.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

South African Institute for Distance Education. (2008). *Using mobile technology for learner support in open schooling* [Report for the Commonwealth of Learning]. Retrieved from http://www.paddle.usp.ac.fj/collect/paddle/index/assoc/col008.dir/doc.pdf

StataCorp. (2013). *Stata Statistical Software: Release 13*. College Station, TX: Author.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations* [DFID Working Paper 38]. London, England: DFID.

Stuhlman, M. W., & Pianta, R. C. (2009). Profiles of educational quality in first grade. *The Elementary School Journal*, *109*(4), 323–342.

TNS. (2012). *Navigating growth in Africa*. London, England: Author.

Trudell, B., & Piper, B. (2014). Whatever the law says: Language policy implementation and early grade literacy achievement in Kenya. *Current Issues in Language Planning*, *15*(1), 4–21. doi:10.1080/14664208.2013.856985

UNESCO. (2005). *Challenges of implementing free primary education in Kenya*. Nairobi, Kenya: Author.

UNESCO. (2012). EFA global monitoring report: Youth and skills, putting education to work. Paris, France: Author.

USAID. (2011). *USAID Education Strategy 2011–2015*. Washington, DC: Author.

Uwezo. (2013). *Are our children learning? Literacy and numeracy across East Africa*. Retrieved from http://www.uwezo.net/wp-content/uploads/2012/08/2013-Annual-Report-Final-Web-version.pdf

Valk, J., Rashid, A., & Elder, L. (2010). Using mobile phones to improve educational outcomes: An analysis of evidence from Asia. *The International Review of Research in Open and Distance Learning*, *11*(1), 117–140.

Vavrus, F. (2009). The cultural politics of constructivist pedagogies: Teacher education reform in the United Republic of Tanzania. *International Journal of Educational Development*, *29*(3), 303–311.

Walsh, C. S., Power, T. P., Khatoon, M., Biswas, S. K., Paul, A. K., Sarkar, B. C., & Griffiths, M. (2013). The "trainer in your pocket": Mobile phones within a teacher continuing professional development program in Bangladesh. *Professional Development in Education*, *39*(2), 186–200.

Wasanga, P. M., Ogle, A. M., & Wambua, R. M. (2010). *Report on monitoring of learner achievement for class 3 in literacy and numeracy*. Nairobi, Kenya: Kenya National Examinations Council.

White, H. (2009). *Theory-based impact evaluation: Principles and practice* (International Initiative for Impact Evaluation Working Paper 3). New Delhi, India: International Initiative for Impact Evaluation.

World Bank. (2011). *Learning for all: Investing in people's knowledge and skills to promote development*. Washington, DC: Author.

World Bank. (2014). *EdStats*. Washington, DC: Author.

World Literacy Foundation. (2012). *The Oxford Declaration*. Retrieved from http://www.oxforddeclaration.org/

Zuilkowski, S. S., Jukes, M. C. H., & Dubeck, M. M. (2016). "I failed, no matter how hard I tried": A mixed-methods study of the role of achievement in primary school dropout in rural Kenya. *International Journal of Educational Development*, *50*, 100–107.