World Development 106 (2018) 324–336

Contents lists available at ScienceDirect

World Development

journal homepage: www.elsevier.com/locate/worlddev

Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides

Benjamin Piper [a,*], Stephanie Simmons Zuilkowski [b], Margaret Dubeck [c], Evelyn Jepkemei [a], Simon J. King [d]

[a] RTI International, Misha Tower, 3rd Floor, 47 Westlands Road, P.O. Box 1181-00621, Village Market, Nairobi, Kenya
[b] Learning Systems Institute, Florida State University, University Center C 4600, Tallahassee, FL 32306, USA
[c] RTI International, 701 13th Street, NW, Suite 750, Washington, DC 20005-3967, USA
[d] RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709-2194, USA

ARTICLE INFO

Article history:
Accepted 24 January 2018

Keywords:
Literacy
Numeracy
Africa
Program evaluation
Randomized controlled trial
Reading

ABSTRACT

Several rigorously evaluated programs have recently shown positive effects on early literacy and numeracy outcomes in developing countries. However, these programs have not been designed to evaluate which ingredients of the interventions are most essential to improve literacy outcomes. Policy makers therefore lack evidence as to whether program ingredients such as teacher professional development (PD), instructional coaching, learner materials, teachers' guides, community support, or technology are required for program impact. The Kenya Primary Math and Reading Initiative was a randomized controlled trial that compared three treatment groups with specific ingredients and a control group. Using literacy and numeracy outcome measures for grades 1 and 2, we evaluated the benefits of the following ingredients: (1) teacher PD and teacher instructional support and coaching; (2) revised student books in literacy and numeracy, at a 1:1 ratio, added to PD and instructional support; and (3) structured teacher lesson plans added to student books, PD, and instructional support. We found that two of the three combinations of ingredients had statistically significant positive impacts on learning outcomes. The results showed that the third combination—PD, teacher instructional support and coaching, 1:1 student books, and structured teacher lesson plans—was most effective. A cost-effectiveness analysis on the ingredients showed that the option of PD and instructional support, 1:1 revised books, and teachers' guides was the most expensive, but that the additional impact on learning made this the most cost-effective intervention. This study rigorously analyzes which ingredients for literacy and numeracy improvement would be most effective for overall impact, and suggests to policy makers that careful decisions regarding program ingredients will lead to more effectively designed and implemented interventions to improve learning in developing countries.

1. Introduction

In recent years, national and international education policy makers increasingly have focused on quality—what do children learn in school, and how valuable are those skills when they leave school? This emphasis is evident in the new Sustainable Development Goals, adopted in 2015. Goal 4 is "Ensure inclusive and equitable quality education and promote lifelong learning," and more specifically, target 4.6 states: "by 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy" (United Nations, 2015, p. 21). Improving learning outcomes is, therefore, now a critical interest of donors and national governments alike.

Although stating that educational quality is important is a necessary first step for national policy makers, they must now begin the difficult work of determining *how* to improve their systems in order to achieve the goals of literacy and numeracy for all. Educational policy making can be difficult in any country, but it is especially challenging in developing countries, where both resources and guiding evidence are particularly scarce. Regarding the evidence (or lack thereof), whereas ever more early grade literacy programs are being rigorously evaluated, such research typically has not shown the effects of specific elements of the programs' technical design. In other words, a researcher might group a literacy program into the broad category of "teacher

* Corresponding author.
E-mail addresses: bpiper@rti.org (B. Piper), szuilkowski@lsi.fsu.edu (S. Simmons Zuilkowski), dubeck@rti.org (M. Dubeck), ejepkemei@tusome.rti.org (E. Jepkemei), sjking@rti.org (S.J. King).

https://doi.org/10.1016/j.worlddev.2018.01.018
0305-750X/

professional development" (PD), but then not comprehensively evaluate the impact of program components such as teacher in-service training, teacher pre-service training, a 1:1 pupil-to-textbook ratio, additional supplementary readers, or teacher lesson plans. Recent research in developing-country contexts has shown that effective programs can have a combination of possible intervention components—teacher professional development, student books, teachers' guides with daily lesson plans, community interventions, information and communication technology (ICT) interventions, and teacher coaching. Given the wide range of program components that may exist within a single study, meta-analyses of these types of studies do not result in a proven set of ingredients that education planners can choose from. Although a larger body of research has examined specific components of literacy programs tightly focused on quality improvement, such as textbook provision and structured teachers' guides, the literature is mixed and controversial. Given the few rigorous studies in Southern countries, there is even less evidence regarding the specific components required to improve numeracy outcomes.

Faced with this lack of clear evidence, it is important to compare various distinct program ingredients against different ingredient combinations, rather than solely comparing each treatment with a control group. In this paper, we aim to provide evidence on the effectiveness of three components of a literacy and numeracy intervention in the Kenyan context. This is important in Kenya, as in other countries focused on instructional improvement, due to the explicit attention paid to improved learning outcomes (the National Education Sector Plan: Ministry of Education [MoE], 2014) and the existence of ongoing, large-scale interventions that aim to rapidly improve learning outcomes. Given scarce resources in the Kenyan context, policy makers need to have evidence not only about the effect of programs, but also about the cost-effectiveness of the options available. This area of research is remarkably nascent, with few large-scale initiatives providing policy makers with cost-effectiveness comparisons (Piper, Zuilkowski, & Mugenda, 2014). Below, we examine how the different components of one intervention worked together to improve primary students' literacy and numeracy outcomes. This evidence may assist policy makers in deciding how to most effectively allocate educational funding in resource-scarce settings.

2. Background and context

2.1. Improving literacy and numeracy in developing countries—what works?

Several recent meta-analyses and systematic reviews have attempted to assess the effects of a range of educational intervention types that have been used in developing countries (Conn, 2014; Ganimian & Murnane, 2016; Glewwe, Hanushek, Humpage, & Ravina, 2013; Kremer, Brannen, & Glennerster, 2013; Kremer & Holla, 2009; Krishnaratne, White, & Carpenter, 2013; McEwan, 2015). As noted by Evans and Popova (2015), these reviews have little overlap in the studies included, and they sometimes categorize studies differently, possibly explaining their divergent findings. Additionally, such approaches compress wide categories of interventions—those using ICT, for example—into one averaged effect size. While this is a useful first step from a policy-making perspective, it does not go far enough in giving education policy makers the specific information they need. Early grade literacy and numeracy programs in developing countries may include a wide range of component parts. In some cases, all of these parts may be complementary and necessary in order to produce positive student results. In other cases, the effect may be driven by one or two of those components, with other parts of the program increasing the cost but not producing additional benefits. In this section, we briefly review the evidence for three components that are typically part of such programs—teacher professional development, student textbooks, and teachers' guides that include daily lesson plans.

The most basic version of the intervention discussed in this study (described further below) involved teacher professional development. Given that student–teacher interactions are at the core of learning to read, and that much of the teaching pedagogy used in Kenya is a didactic type not supported by current research (Dubeck, Jukes, & Okello, 2012), with limited teacher and student interactions (Ackers & Hardman, 2001), it is critical that teachers already in the classroom be retrained in literacy and numeracy pedagogical techniques. Piper and Spratt (2017) reviewed potential options for teacher professional development for Cambodia, describing a range of options that could be effective in the current educational context, many of which could be applicable in Kenya.

While not all teacher professional development programs are successful, studies in Kenya have found that teacher professional development programs can improve teachers' literacy knowledge (Dubeck, Jukes, Brooker, Drake, & Inyega, 2015). In a meta-analysis of 17 rigorous studies of primary-level teacher professional development programs in developing countries, McEwan (2015) found a mean effect size on student outcomes of 0.12 standard deviations (SD). Conn (2014), examining studies in sub-Saharan Africa in her meta-analysis, found far larger effects for pedagogical interventions, a broader category that includes teacher coaching: 0.918 SD for all studies, and 0.228 SD for those she categorized as "high quality."

In addition to teacher professional development, all three variations of the program studied here included teacher coaching as a base element. In the United States, studies in a variety of contexts have shown that coaching approaches can impact teacher pedagogy and student outcomes (Biancarosa, Bryk, & Dexter, 2010; Carlisle, Cortina, & Katz, 2011; Matsumura, Garnier, Correnti, Junker, & Bickel, 2010; Teemant, 2014). While evidence on coaching-based professional development is limited in sub-Saharan Africa, our work in Kenya found that this approach was also effective in this context (Piper & Zuilkowski, 2015; Piper et al., 2014; Zuilkowski & Piper, 2017). Given that Kenyan teachers generally have two or three years of postsecondary education, compared to four or more in wealthier countries, such one-on-one support is critical in developing teacher skills.

The role of text access in learning to read is central: While students can learn letters and words from a blackboard or slate, children cannot truly become fluent readers without exposure to a variety of reading material (Kim, Boyle, Zuilkowski, & Nakamura, 2017). Teachers in resource-poor contexts greatly value textbooks as pedagogic tools (Lee & Zuilkowski, 2015). The empirical evidence on the relationship between textbooks and student achievement is mixed, however. A number of studies, many conducted in the 1980s and 1990s, have found positive associations between textbook availability and outcomes in developing countries (Fehrler, Michaelowa, & Wechtler, 2009; Fuller, 1987; Harbison & Hanushek, 1992; Heyneman & Jamison, 1980; Heyneman, Jamison, & Montenegro, 1984; Lockheed, Vail, & Fuller, 1986; Riddell & Nyagura, 1991; Yara & Otieno, 2010). Many of these studies used multiple regression analytic methods that were unable to isolate the causal effect of the textbooks. More recently, other studies have shown that the mere presence of books is not enough to ensure improved student performance; other factors interact with the availability of textbooks to produce effects (Mohammed & Kumari, 2007; Read, 2011; Sabarwal, Evans, & Marshak, 2013; Somerset, 2011). A recent meta-analysis found that the mean effect size for interventions that included instructional materials was just 0.08 standard deviations (SD) (McEwan, 2015). As noted above,

this research suggests that provision of texts alone is not enough, as students must be allowed to actually use the books and teachers must know how to effectively incorporate the texts into their pedagogy. In Kenya, Glewwe, Kremer, and Moulin (2009) found in a randomized controlled trial (RCT) that only the highest-achieving students' scores increased after the provision of additional textbooks.

Another possible explanation for books showing a limited effect on learning, as suggested by Glewwe and colleagues, is that the books are not well designed—they may be leveled too high or in a language that children are just beginning to learn. Studies conducted in India also have pointed to the importance of teaching children at the right level (Banerjee et al., 2016). There is little research that distinguishes between the impact of standard textbooks, often with many flaws, and well-designed textbooks. In the Kenyan literacy and numeracy program that we analyze below—the Primary Math and Reading (PRIMR) Initiative—student textbooks were redesigned to match the curriculum, aligned with current research on literacy and numeracy development, and matched to the actual skill levels of children in grades 1 and 2. Two treatment groups received these student textbooks, but a third treatment group did not. The PRIMR data that we analyze will therefore help to clarify the impact of high-quality textbooks in early grade learning in countries like Kenya.

Our current study also addresses the controversy regarding the effectiveness of an approach using structured teachers' guides provided to teachers. Several rigorous studies have shown that programs using teachers' guides can have a positive effect on learning outcomes and can make it easier for teachers to learn new methods (Reeves, 2010). In the United States, critics have raised concerns about scripted lesson plans potentially limiting both the curriculum and the ability of teachers to adapt content to make it more relevant to their students (Milner, 2014). Researchers have also documented negative impacts of scripted lessons on teacher motivation (Crocco & Costigan, 2007). The potential for structured teachers' guides to assist teachers to change their teaching methods (Reeves, 2010), however, might be particularly helpful in contexts like Kenya's, where teachers generally have less preparation and academic background than in Western contexts. A number of studies have used structured teachers' guides in sub-Saharan Africa (Piper & Korda, 2010; Tilson, Kamlongera, Pucilowski, & Nampota, 2013), and Bridge International Academies is using a tightly controlled scripted lesson plan approach to expand educational access in Liberia, Kenya, Uganda, and Nigeria, with plans to extend across Africa and Asia (see www.bridgeinternationalacademies.com). Despite the spread of this approach, it remains a topic of discussion in sub-Saharan Africa as in the U.S. (Ravitch, 2015). To our knowledge, to date, no studies have examined the impact of structured teachers' guides over and above the impact of other program components, an issue which we address below.

Understanding which academic subjects are most sensitive to interventions is an important part of understanding the essential ingredients to improve learning outcomes. Western-based research on in-service teacher professional development suggests that mathematics is somewhat more sensitive to initial interventions than literacy (Angrist & Lavy, 2001; Bressoux, 1996; Bressoux, Kramarz, & Prost, 2006). However, in previous studies in Kenya, literacy programs had larger effects (Piper, Ralaingita, Akach, & King, 2016) or had small literacy effects and no numeracy effects on learning (Ngware et al., 2015). The comparison is not as simple as looking across subjects, as some evidence suggests that there is significant within-subject variance as well. For example, particular areas of literacy such as letter sounds are more rapidly improved than is reading comprehension (Jukes et al., 2017), and procedural components of mathematics have different impacts than do conceptual mathematics skills (Donovan & Bransford, 2005; National Council of Teachers of Mathematics [NCTM], 2014; Piper, Ralaingita et al., 2016). PRIMR showed a larger effect on literacy, but given the later start of the numeracy portion of the PRIMR program, it is unclear whether the divergence was due to the delay or to the program itself (Piper & Kwayumba, 2014). Therefore, there remains a gap in the literature as to whether instructional interventions in literacy or numeracy will have larger effects.

2.2. The Kenyan context

In 2010, Kenya's educational spending was 6.7% of its gross national product and 23.7% of government expenditures, both figures higher than average for sub-Saharan Africa (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2015). However, the majority of this funding went to teacher salaries, with little left over for in-service teacher professional development and instructional support. Kenya provides a per student capitation grant to cover textbooks and other instructional materials, yet studies have found limited access to learning materials across Kenya (UNESCO, 2016). Even though schools receive approximately US$7 per child per year to provide these materials, the average student–textbook ratio was 3:1 across rural and peri-urban locations (Piper & Mugenda, 2012), a modest improvement from earlier figures (Onsomu, Nzomo, & Obiero, 2005). Additionally, the government's Kiswahili and English textbooks were designed with text that was difficult for children in lower grades to decode. The thematic approach of the syllabuses on which the books were based demanded the books include words with phonetic arrangements difficult for children learning to read in English.

Over 80% of teachers in 72 Kenyan schools reported that they had not received any kind of in-service professional development in the previous year and a half (Ngware, Oketch, & Mutisya, 2014). Government-funded instructional coaches[1] for public school teachers have been in employment for many years, but their administrative roles have gradually expanded, leaving little room for direct teacher coaching and support. Consequently, teaching and learning using professional formative evaluation of teachers has been largely nonexistent in Kenya, although research suggests that instructional improvement depends on collaboration between teachers, and on their engagement in the identification of in-service needs (Wanzare & Ward, 2000).

Despite greater attention to quality in recent years by the government and international donors, outcomes remain poor. In a recent study, less than 5% of first- and second-grade children met government literacy benchmarks (Piper & Mugenda, 2012), and Kenyan primary school learners struggled with both procedural tasks, such as basic addition and subtraction; and more complex conceptual tasks, such as word problems and number sense activities (Reubens, 2009). Classroom observations revealed that pupils spent significant portions of classroom time practicing basic facts, yet results showed very low outcomes on accuracy and fluency of those basic facts (Piper, King, & Mugenda, 2016).

The Kenyan Ministry of Education designed PRIMR as a pilot program to address the gaps in teacher professional development and materials provision, with the goal of producing improved student outcomes in grades 1 and 2 in literacy and numeracy. A randomized controlled trial evaluation approach showed that the

[1] Until 2015, instructional support for teachers in public schools was provided by Teachers' Advisory Centre (TAC) tutors, under Kenya's Teachers' Service Commission (TSC). In early 2016 the TSC changed the title of the position to Curriculum Support Officer (CSO) to better reflect renewed emphasis on teacher support. Although PRIMR (2011–2015) dealt with TAC tutors, this article reflects the updated terminology.

program had an impact on children's literacy and numeracy outcomes, increasing the percentages of children meeting government benchmarks for oral reading fluency in Kiswahili and English and improving students' conceptual and procedural numeracy (Piper, Ralaingita et al., 2016; Piper et al., 2014). As a result, nationwide rollout of the approach began in 2015 and 2016 in two separate programs. The literacy program, called Tusome ("Let's Read" in Kiswahili), will reach more than five million first, second, and third-grade students; 100,000 teachers and head teachers; and more than 1200 public school Curriculum Support Officers (CSOs) and private school instructional coaches by 2018. While the PRIMR program evaluations cited above showed that the combination of teacher professional development, materials provision, and structured teachers' guides had been effective as a whole, further refinement would be needed to push more children past the benchmarks for literacy and numeracy. In order to improve the program, the MoE decided that it was necessary to understand what drove the observed results. This interest in understanding how specific program ingredients affect learning outcomes led to the research questions discussed below.

2.3. Research questions

While the studies reviewed above suggested that teacher professional development, student books, and structured teachers' guides may all be components of successful literacy and interventions in low-resource settings, it is unclear whether all three are required. In severely resource-constrained settings, policy makers need more detailed evidence as to which parts of the program are most strongly driving student outcomes. While the "full" PRIMR treatment (described below) improved students' literacy and numeracy outcomes (Piper et al., 2014), the cost savings that would result from focusing only on the most effective components could lead to faster rollout nationwide as well as improved long-term sustainability. Therefore, in this study, we address the following questions:

RQ1: What is the causal effect of teacher professional development and coaching, revised pupil books at a 1:1 ratio, and structured teachers' guides on student achievement in English, Kiswahili, and mathematics?

RQ2: Do three different combinations of the ingredients have different impacts on student achievement in English, Kiswahili, and mathematics?

3. Methodology

Kenya has been the site of many RCTs in the education sector. For example, several researchers, including those associated with the Massachusetts Institute of Technology's Poverty Action Lab, have undertaken many studies on the impact of various instructional improvement strategies. These included additional classroom books (Glewwe et al., 2009), contract teachers (Duflo, Dupas, & Kremer, 2015), student tracking (Duflo, Dupas, & Kremer, 2011), and after-school tutoring (Duflo & Kiessel, 2012). Other studies have analyzed the effect of older pupils reading to younger pupils outside of classroom time (Jasti, Jukes, Dubeck, Elliott, & Inyega, 2016; Jukes et al., 2016) and the effectiveness of lesson plans and SMS messages (Jukes et al., 2017). Another study showed that implementation of a set of Reading to Learn intervention packages had no effect on written literacy or numeracy in Coastal Kenya, and only a small effect (0.08 SD) on oral literacy (Lucas, McEwan, Ngware, & Oketch, 2014).

The section that follows describes the methodologies of the PRIMR RCT study that served as the foundation for this article.

3.1. The PRIMR intervention

The PRIMR pilot involved 847 government schools in rural zones in Bungoma and Machakos counties. It was designed such that three treatment groups of the PRIMR program could be compared to a control condition. The most basic treatment involved only teacher professional development and instructional support—10 days per year of professional development for teachers, 15 days of professional development for Curriculum Support Officers, and ongoing teacher coaching by the CSOs using observation and feedback tools. The teachers were trained and supported to more effectively use the existing instructional materials in their classrooms. The second version of PRIMR included the same number of days of teacher and CSO professional development, plus 150-page student textbooks for first- and second-grade Kiswahili, English, and mathematics, provided at a 1:1 ratio (the PRIMR books and teachers' guides can be found by using the search term "PRIMR" in the United States Agency for International Development's [USAID's] Development Experience Clearinghouse database, https://dec.usaid.gov/dec/content/search.aspx). Teachers in this second version of PRIMR were encouraged to develop their own lesson plans that incorporated the revised student textbooks. The third version of PRIMR included teacher professional development, instructional support, and revised textbooks, and added teachers' guides, which included 150 days of partially scripted lessons, matched to the revised pupil textbooks.

Each treatment arm therefore increased in intensity over the preceding one, and comparing the impact of the three treatment arms made it possible to test whether the increased costs for professional development and coaching support, student textbooks, and teachers' guides had additive effects on student learning. This study design was intended to test the PRIMR pilot program interventions at a medium scale under real-world conditions, through the government structures and using government employees for support and professional development. This is an important characteristic of the three treatment groups, as they were designed to utilize existing structures so that if any of the treatments showed positive effects, they were more likely to be scaled at a national level—as in fact did occur later under Tusome and the Primary Education Development (PRIEDE) numeracy program.

The PRIMR research showed the effectiveness of the full package of instructional components on numeracy (Piper, Ralaingita et al., 2016), on literacy (Piper et al., 2014), and in mother tongue (Piper, Zuilkowski, & Ong'ele, 2016); using ICT interventions (Piper, Zuilkowski, Kwayumba, & Strigel, 2016); depending on instructional coaching (Piper & Zuilkowski, 2015); and, particularly, for the poor (Piper, Jepkemei, & Kibukho, 2015). For further detail on the design of the PRIMR program, see Piper et al. (2014).

This study, funded by the British Department for International Development (DFID) in Kenya, used the same treatment methodology as other studies on PRIMR that were funded by USAID/Kenya, but went beyond the previous PRIMR research to investigate the specific contributions of particular intervention components. The study was conducted in government schools, in Bungoma and Machakos counties in Kenya, both of which are primarily rural and did not participate in the initial USAID PRIMR pilot. Bungoma is a largely agricultural county, with many families producing sugar cane and maize. Its location on the border with Uganda also results in significant trade activity and cross-border migration. The Bukusu are the largest ethnic group in Bungoma. In Machakos, which is approximately 35 miles southeast of the capital, Nairobi, millet, sorghum, and maize are the main agricultural products. The Kamba ethnic group predominates in Machakos. In Bungoma, 420 government primary schools participated in the overall PRIMR study, and 414 in Machakos.

3.2. Sample and procedures

The impact evaluation of PRIMR used a three-stage stratified cluster design. Twenty-two school zones each within Bungoma and Machakos counties (i.e., 44 zones out of 72 total zones) were randomly selected and assigned to one of the three treatment groups or the control group. Sampling was done at the zone level because the government's Curriculum Support Officers are assigned to support whole zones of typically 12 to 20 schools, and the CSOs constituted the base of the instructional support system. Through a stepped implementation process, all of the zones—treatment and control—were receiving the full PRIMR intervention by 2015. With respect to the complex survey design, the zones were clusters of schools.

The baseline assessment for the DFID PRIMR study sampled between 4 and 10 schools randomly from each of the 44 selected zones for a total of 171 randomly selected schools. The final sampling stage was selection of students within schools; students were stratified by grade and gender and five students per strata were randomly selected for each school. If less than five students were available, they were all selected. A total of 3309 grade 1 and 2 students were assessed in the March 2013 baseline study. Assessments were completed by a trained team of local assessors who had been conducting literacy and numeracy assessments with the same types of instruments (see below for further information) since 2007. Interrater reliability tests during training showed very high average rates of reliability—98.0% for mathematics, 95.5% for English, and 95.0% for Kiswahili. The three treatment groups received the DFID PRIMR program intervention between May 2013 and October 2014. Endline data were collected from 230 randomly selected schools (4566 students) in October 2014. The final analyses presented here include the rural zones in the research design, removing the two peri-urban zones in the two counties.

3.3. Measures

The measures used in these analyses were drawn from the results of two instruments administered one-on-one to pupils—the Early Grade Reading Assessment, or EGRA (Gove & Wetterberg, 2011) and the Early Grade Mathematics Assessment, or EGMA (Platas, Ketterlin-Gellar, Brombacher, & Sitabkhan, 2014)—adapted locally. All assessments were piloted prior to use.

At baseline (March 2013) and endline (October 2014), the EGRA, administered in English and Kiswahili, included measures of letter sound fluency, decoding fluency (nonwords), oral reading fluency, reading comprehension, and the proportion of children at the MoE benchmark; listening comprehension was tested in Kiswahili only (RTI International, 2015). The subcomponent and measure descriptions for English, Kiswahili, and mathematics can be found in Table 1 below. The literacy assessments were equated across the data collection points (Albano & Rodriguez, 2012). In the English assessment, the Cronbach's alphas averaged 0.90. The reliabilities were similar for Kiswahili, with an average of 0.92 and a range from 0.89 (nonword fluency and oral reading fluency) to 0.94 (listening comprehension). In the analyses presented in this paper, we use the average effect size in standard deviations across the instrument subcomponents for each language to present the overall effect of PRIMR on literacy in English and Kiswahili.

The EGMA tool measured both conceptual and procedural numeracy, as described in Table 1. Specific areas included number identification, quantity discrimination, missing number patterns, word problems, addition fluency, and subtraction fluency. The overall Cronbach's alpha for the mathematics assessment was 0.90, with all subtasks having alphas of 0.89 or 0.90. As with literacy, we averaged the effect sizes in standard deviations across the six subcomponents of the numeracy instrument to obtain an overall measure of numeracy.

3.4. Data analysis

In other USAID-funded PRIMR studies in literacy and numeracy (Piper, Ralaingita et al., 2016; Piper, Zuilkowski, Kwayumba, et al., 2016), differences-in-differences (DID) analytic methods were used to account for the fact that, while random selection and assignment were utilized in assigning zones to treatment, baseline equivalence among the various treatment and control groups was not always achieved on the key outcomes. In order to determine which identification strategy was required in this study, we compared student assessment results at the baseline. Table 2 below shows the results by task and for pupils in the control and three treatment groups. The results show some statistically significant differences at baseline between some of the treatment groups. For example, in grade 1, the PD, books, and teachers' guides treatment group performed worse than the control group in oral

Table 1
Measure descriptions for English, Kiswahili, and mathematics.

Instrument	Subcomponent	Description	Timed/ untimed	Measure
English	Letter fluency	Letters read correctly per minute	Timed (1 min)	Correct letters per minute
	Nonword fluency	Nonsense words read correctly per minute	Timed (1 min)	Correct words per minute
	Oral reading fluency	Connected text words read correctly per minute	Timed (1 min)	Correct words per minute
	Reading comprehension	Percentage of comprehension questions correct	Untimed	Percentage correct
	% at benchmark	Percentage of pupils reaching the 30 correct words per minute (cwpm) fluency benchmark	Untimed	Percentage reaching benchmark
Kiswahili	Letter fluency	Letters read correctly per minute	Timed (1 min)	Correct letters per minute
	Nonword fluency	Nonsense words read correctly per minute	Timed (1 min)	Correct words per minute
	Oral reading fluency	Connected-text words read correctly per minute	Timed (1 min)	Correct words per minute
	Reading comprehension	Percentage of comprehension questions correct	Untimed	Percentage correct
	Listening comprehension	Percentage of oral listening comprehension questions correct	Untimed	Percentage correct
	% at benchmark	Percentage of pupils reaching the 17 cwpm fluency benchmark	Untimed	Percentage reaching benchmark
Mathematics	Number identification	Numbers correctly identified	Timed	Correct numbers per minute
	Quantity discrimination	Determining which of two numbers is larger	Untimed	Percentage correct
	Missing number	Determining what is the missing number from a pattern	Untimed	Percentage correct
	Word problems	Correct word problems response	Untimed	Percentage correct
	Addition fluency	Addition problems answered correctly per minute	Timed (1 min)	Correct addition problems per minute
	Subtraction fluency	Subtraction problems answered correctly per minute	Timed (1 min)	Correct subtraction problems per minute

Table 2
Baseline descriptive statistics by treatment group, with statistical significance tests for baseline equivalence for the three treatment groups relative to the control group. Standard errors in parentheses.

Instrument	Subcomponent	Grade 1				Grade 2			
		Control	PD and coaching	PD and books	PD, books, and teachers' guides	Control	PD and coaching	PD and books	PD, books, and teachers' guides
English	Letter fluency	2.0 (0.7)	5.2 (2.1)	1.6 (0.4)	1.7 (0.5)	3.2 (0.4)	10.5* (2.9)	3.1 (0.9)	3.5 (0.9)
	Nonword fluency	0.6 (0.2)	1.9 (1.2)	0.4 (0.1)	0.4 (0.2)	4.4 (1.4)	11.7∼ (3.7)	4.1 (1.1)	4.0 (0.4)
	Oral reading fluency	0.5 (0.1)	1.7 (1.2)	0.4 (0.1)	0.2** (0.0)	5.2 (1.7)	15.3∼ (5.4)	4.9 (1.5)	3.9 (1.0)
	Reading comprehension	0.1 (0.1)	0.6 (0.5)	0.0 (0.0)	0.0 (0.0)	1.9 (0.5)	6.0 (2.5)	1.3 (0.4)	1.2 (0.4)
	% at benchmark	0.0 (0.0)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)	5.1 (2.2)	20.4 (8.7)	7.4 (2.8)	3.2 (1.4)
Kiswahili	Letter fluency	2.6 (0.9)	7.0 (2.9)	3.0 (0.7)	3.5 (0.6)	6.7 (0.9)	14.0* (2.9)	6.7 (2.1)	7.4 (0.7)
	Nonword fluency	0.3 (0.1)	1.5 (1.0)	0.2 (0.1)	0.2 (0.1)	3.6 (1.2)	10.1∼ (3.4)	2.6 (0.8)	3.2 (0.4)
	Oral reading fluency	0.6 (0.2)	2.8∼ (0.2)	0.8 (0.2)	0.6 (0.2)	6.0 (1.2)	15.2∼ (0.3)	5.6 (0.2)	5.0 (2.0)
	Reading comprehension	0.3 (0.2)	2.3 (1.6)	0.2 (0.2)	0.5 (0.3)	3.7 (1.1)	16.2∼ (6.9)	3.8 (1.3)	3.2 (0.5)
	Listening comprehension	7.4 (1.1)	20.3 (8.2)	7.9 (2.5)	10.6 (5.8)	18.6 (2.6)	36.1∼ (8.0)	16.2 (0.7)	17.2 (6.4)
	% at benchmark	0.8 (0.3)	6.6∼ (3.2)	1.1 (0.7)	0.0* (0.0)	17.6 (6.1)	39.2 (11.3)	13.2 (3.7)	13.2 (2.9)
Math	Number identification	4.7 (0.4)	6.9 (1.5)	4.7 (0.3)	4.6 (0.6)	12.3 (0.9)	16.5 (2.4)	11.8 (0.4)	12.2 (0.6)
	Quantity comparison	14.5 (2.0)	19.8 (4.9)	13.3 (1.1)	12.4 (1.8)	37.3 (3.5)	50.1∼ (6.2)	35.1 (1.6)	36.8 (3.4)
	Missing number	10.3 (1.2)	12.9 (2.1)	9.4 (0.3)	10.8 (1.1)	21.1 (1.4)	26.5 (3.3)	19.8 (0.9)	21.2 (1.7)
	Word problems	9.1 (2.1)	12.0 (2.4)	6.5 (0.6)	6.7 (1.2)	19.6 (2.3)	25.1∼ (1.8)	18.9 (2.0)	19.1 (3.4)
	Addition fluency	2.8 (0.4)	3.8 (0.7)	2.3 (0.2)	2.4 (0.3)	6.8 (0.6)	8.4* (0.4)	6.8 (0.1)	6.3 (0.4)
	Subtraction fluency	1.6 (0.3)	2.1 (0.4)	1.1 (0.2)	0.9∼ (0.1)	4.2 (0.6)	5.8* (0.5)	4.0 (0.2)	3.8 (0.4)

∼$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

reading fluency ($p < .05$) and the percentage of children reading at the benchmark ($p < .05$). For grade 2, the PD and coaching group outperformed the control group in English letter fluency ($p < .05$), Kiswahili letter fluency ($p < .05$), addition fluency ($p < .05$), and subtraction fluency ($p < .05$). Given the handful of statistically significant differences at baseline, particularly where the professional development-only treatment group outperformed the control group in grade 2, we decided to fit a differences-in-differences model. In Table 3, we present the key baseline covariates at the baseline and endline for the DFID PRIMR study.

We utilized the differences-in-differences identification strategy to identify causal effects. The DID model differentiated among four sets of children, separating those children into randomly assigned treatment and control schools and by the assessment time, either the baseline in March 2013 or endline in October 2014. The models then separately compared each of the three treatment groups against the control. Given that previous results from DFID PRIMR showed that including control variables made little difference in the overall PRIMR impact estimates (Piper, Ralaingita et al., 2016; Piper, Zuilkowski, & Ong'ele, 2016), we used a parsimonious model to estimate the effect of PRIMR using the svy commands in Stata. Sampling weights were applied to estimates and standard errors were adjusted to account for the three-stage stratified cluster sample design.

PRIMR used a difference-in-difference method where the outcomes are measured for the control and treatment groups across time points. We used notation of treatment group i and time t for a cross-sectional analysis, where $t = 0$ and $t = 1$ denote the baseline and endline time periods, respectively, and $i = 0$ and $i = 1$ represent the control and treatment groups, respectively. Note that $t = 0$ is prior to the intervention.

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Post_t + \beta_3 (Treat \cdot Post)_{it} + \varepsilon_{it},$$

where each unit is denoted as follows:

Y_{it} is the outcome,
Treatment is a dummy when the observation is in the treatment group,
Post is a post-treatment dummy variable, and
ε is random unobserved error.

The coefficient of interest, β_3, is the product of the interaction term $Treat \cdot Post$. It is equivalent to a dummy variable equal to 1 for children in the treatment group during the follow-up data collection. The difference-in-differences estimate is:

$$\bar{\beta}_4 = (\bar{y}_{1,1} - \bar{y}_{1,0}) - (\bar{y}_{0,1} - \bar{y}_{0,0}).$$

The treatment effect for the cross-sectional analysis is the difference-in-difference mean. The effect size measures the magnitude or strength of the treatment effect. The effect size used for the PRIMR study was Cohen's d (Cohen, 1988), which is the difference-in-difference divided by the pooled standard deviation, thus:

$$d = \frac{(\bar{y}_{1,1} - \bar{y}_{1,0}) - (\bar{y}_{0,1} - \bar{y}_{0,0})}{s}$$

Table 3
Student covariates at baseline and differences by covariate at endline, by treatment group (standard errors in parentheses).

Covariate	Control		PD and coaching		PD, coaching, and books		PD, coaching, books, and teachers' guides	
	Baseline	Endline	Baseline	Endline	Baseline	Endline	Baseline	Endline
Student female (%)	48.0 (1.3)	51.9∼ (1.3)	48.7 (0.7)	48.7 (1.3)	50.6 (1.3)	49.6 (0.8)	47.7 (0.5)	48.1 (0.8)
Student age (years)	7.5 (0.2)	7.9* (0.2)	7.4 (0.5)	7.9** (0.4)	7.6 (0.1)	8.0 (0.2)	7.4 (0.1)	7.8 (0.1)
Student wealth index	2.5 (0.2)	2.8* (0.2)	3.0 (0.5)	3.0 (0.4)	2.9 (0.1)	3.1 (0.1)	2.9 (0.1)	3.0 (0.1)
Attended kindergarten or preschool (%)	88.5 (3.5)	79.3 (5.6)	91.5 (1.6)	92.8 (3.7)	92.2 (2.3)	94.9 (1.2)	92.8 (3.0)	89.6 (3.4)
Student absent more than a week	34.7 (3.1)	43.7 (4.2)	43.5 (4.1)	41.5 (2.7)	34.4 (4.9)	41.5 (8.3)	45.5 (7.2)	44.7 (4.2)
Had English book (%)	43.8 (5.3)	52.6 (10.8)	40.6 (5.0)	66.6 (17.8)	40.3 (4.0)	78.7* (5.3)	48.1 (8.1)	75.1 (5.7)
Had Kiswahili book (%)	40.3 (4.6)	46.6 (7.2)	39.2 (5.7)	68.6 (18.7)	37.6 (4.6)	81.9* (5.3)	49.7 (10.1)	73.8 (4.0)

∼$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.
This table presents the baseline covariates for each treatment group and the endline estimates within each treatment group, and statistical significance comparisons between baseline and endline estimates within each treatment.

4. Findings

4.1. RQ1 – What is the causal effect of teacher professional development and coaching, revised pupil books at a 1:1 ratio, and structured teachers' guides on student achievement in English, Kiswahili, and mathematics?

In order to answer our first research question, we fit differences-in-differences models to estimate the effect of each of the three treatment groups on learning outcomes in English, Kiswahili, and mathematics. The results are presented in Table 4, where we show the DID estimate with the standard errors in parentheses. The second column in each pair presents the Cohen's *d* effect size for each estimate.

The results for the professional development and coaching program, the first treatment group, are presented in the left-hand panel for grade 1, and in the right-hand panel for grade 2. The rows present the English results first, then the Kiswahili results, and finally the bottom of the table presents the mathematics results. The PD and coaching program for grade 1 showed no more impacts on learning outcomes than would be expected by chance. There were statistically significant effects for English letter fluency at the 0.10 level, increasing results by 7.1 correct letters per minute (clpm) ($p < .10$); and the percentage of students meeting the English benchmark by 9.2 percentage points ($p < .05$). It had no impacts on Kiswahili or mathematics. In grade 2, the PD and coaching program increased English letter fluency by 13.9 clpm ($p < .05$), English reading comprehension by 6.9 percentage points ($p < .10$), Kiswahili letter fluency by 11.2 clpm ($p < .10$), and word problems by 8.5% at the .10 level. In short, the PD and coaching program improved only 2 of 34 assessment outcomes at the .05 level across the two grades, or no more than would be expected by chance.

For the PD, coaching, and books treatment group in grade 1, the second treatment group, there were statistically significant effects on English, increasing letter fluency by 12.1 correct letters per minute (clpm) at the .10 level; and increasing the percentage of children reading at the benchmark by 9.0 percentage points ($p < .05$). For Kiswahili in grade 1, the PD, coaching, and books treatment group also increased listening comprehension by 8.4 percentage points ($p < .05$). There was no effect for mathematics outcomes. In grade 2, the PD, coaching, and books treatment increased English letter fluency by 20.5 clpm ($p < .05$), nonword fluency by 7.0 cwpm at the 0.10 level, and oral reading fluency by 10.5 correct words per minute (cwpm) at the 0.10 level. The PD, coaching, and books treatment increased Kiswahili outcomes in letter sound fluency by 19.9 clpm ($p < .01$), reading comprehension by 8.5 percentage points ($p < .01$), listening comprehension by 13.2 percentage points ($p < .05$), and the percentage of children at the benchmark by 21.0 percentage points ($p < .05$). Statistically significant effects were identified only for number identification, by 3.5 correct numbers per minute ($p < .05$); and missing number, by 11.3 percentage points ($p < .01$). In total, across both languages and mathematics, PD, coaching, and books increased 7 of the 34 assessment outcomes in grades 1 and 2 at the .05 level.

Finally, the PD, coaching, books, and teachers' guide treatment group increased outcomes at the .05 level on 9 of 17 of the grade 1 measures and 13 of 17 grade 2 measures, for a total of 22 of the 34 measures. Many of the effect sizes for the PD, coaching, books, and teachers' guides treatment were quite large, especially for letter fluency in both English (1.71 SD in grade 1 and 2.15 SD in grade 2) and Kiswahili (1.42 SD in grade 1 and 1.82 SD in grade 2).

A critical literacy measure is the percentage of children reading at least at the MoE benchmark for fluency. The PD and coaching group increased the grade 1 English percentage by 9.2 percentage points ($p = .05$), but had no effect on the grade 2 English percentage, the grade 1 Kiswahili percentage, or the grade 2 Kiswahili percentage. For comparison, the PD, coaching, and books treatment group increased the percentage of children reaching at the benchmark in grade 1 English by 9.0 percentage points ($p < .05$), had no effect on the percentage at benchmark in grade 1 Kiswahili or grade 2 English, and increased the grade 2 Kiswahili percentage by 21.0 percentage points ($p < .05$). Finally, the PD, coaching, books, and teachers' guide treatment group had the largest effect of all of the groups. This treatment group increased the percentage of children at the benchmark by 11.2 percentage points in English grade 1 ($p < .05$), 39.6 percentage points in English grade 2 ($p < .01$), 12.7 percentage points in Kiswahili grade 1 ($p < .10$), and 39.5 percentage points in Kiswahili grade 2 ($p < .01$).

While there is debate regarding the correct benchmarks for effect sizes, Cohen (1988) recommended considering 0.2, 0.5, and 0.8 SD as small, medium, and large effects, respectively. Utilizing Cohen's guidance, the DFID PRIMR treatment groups had significant effects on learning outcomes. The PD, coaching, books, and teachers' guides treatment group had large average effect sizes in grade 1 English (0.93 SD), grade 2 English (1.29 SD) and grade 2 Kiswahili (1.11 SD), and medium effects on grade 1 Kiswahili (0.73 SD) and grade 2 mathematics (0.56 SD). A small average effect was found for grade 1 mathematics (0.38 SD). The PD, coaching, and books treatment group did not have any large effect sizes, but showed medium effects on grade 1 English (0.56 SD), grade 2 Kiswahili (0.58 SD), and grade 2 English (0.71 SD). Small effects were found for grade 1 Kiswahili (0.34 SD) and grade 2 mathematics (0.32 SD), with grade 1 mathematics showing a negligible effect (0.13 SD). There were no subjects that had medium or large effects for the PD and coaching group; PD and coaching had a small effect on grade 1 English (0.42 SD) and grade 2 English (0.30 SD). Negligible effects for the PD and coaching treatment group were found for grades 1 and 2 in both mathematics and Kiswahili. Averaging the effect sizes across grades and subjects, we found that the average effect size for the PD, coaching, books, and teachers' guide group was 0.83 SD; PD, coaching, and books was 0.44 SD; and PD and coaching was 0.20 SD.

The magnitude of the effect sizes from this study are noteworthy. Conn (2017) noted that the average effect size of the highest-quality pedagogical interventions in her meta-analysis was 0.23 SD, and the average effect size of the RCTs (without outliers) was 0.57 SD with a standard error of 0.19. The effect size of the PRIMR PD, coaching, books, and teachers' guide group (0.83 SD) would therefore be on the upper end of Conn's 15 pedagogical interventions assessed, and the PD, coaching and books treatment group would be below the average effect. In comparison to the McEwan (2013) meta-analysis, the PD, coaching, books, and teachers' guide group had some of the largest impacts of any identified in the paper, whether in the instructional materials treatment group (mean = 0.08 SD), the computers and technology group (mean = 0.15 SD), or the treatments with teacher training group (mean = 0.12 SD). Among the treatment arms with comparable cost-effectiveness data, the impact of the PD, coaching, books, and teachers' guide group was larger than in any other study in McEwan (2013). Of particular interest is the much larger effect size in the third treatment group, which differed from the second treatment group only by the inclusion of teachers' guides.

4.2. RQ2 – Do three different combinations of the ingredients have different impacts on student achievement in English, Kiswahili, and mathematics?

In order to answer this research question, we compared the effects of the three treatment groups among each other and have presented the analysis in Table 5. It offers differences-in-differences results comparing the effects of the three treatment

Table 4
Differences in differences effects of three PRIMR treatment groups on student outcomes (standard errors in parentheses) and Cohen's *d* effect sizes.

		Grade 1						Grade 2					
		PD and coaching		PD, coaching, and books		PD, coaching, books, and teachers' guides		PD and coaching		PD, coaching, and books		PD, coaching, books, and teachers' guides	
Instrument	Subcomponent	Effects	Effect size	Effects	Effect size	Effects	Effect size	Effects	Effect size	Effects	Effect size	Effects	Effect size
English	Letter fluency	7.1∼ (3.8)	0.57	12.1∼ (6.1)	1.02	19.2** (5.3)	1.71	13.9* (4.8)	0.87	20.5* (8.5)	1.37	30.2*** (5.1)	2.15
	Nonword fluency	2.4 (1.8)	0.33	2.1 (2.3)	0.34	5.0* (1.9)	0.78	1.1 (3.8)	0.08	7.0∼ (3.4)	0.59	11.4** (3.2)	0.99
	Oral reading fluency	2.3 (2.2)	0.27	3.4 (3.5)	0.41	7.6* (2.6)	0.86	0.5 (5.4)	0.02	10.5∼ (5.6)	0.56	22.5** (6.5)	1.21
	Reading comprehension	1.7 (1.0)	0.27	3.4 (2.3)	0.47	5.1*** (1.2)	0.65	6.9∼ (3.8)	0.41	8.2 (4.7)	0.52	15.2* (5.0)	0.91
	% at benchmark	9.2* (1.9)	0.54	9.0* (14.8)	0.55	11.2* (2.4)	0.65	3.5 (9.4)	0.09	17.1 (10.0)	0.50	39.6** (9.9)	1.21
	Average effect size		0.42		0.56		0.93		0.30		0.71		1.29
Kiswahili	Letter fluency	2.5 (4.8)	0.19	9.1 (5.2)	0.77	15.8* (5.0)	1.42	11.2∼ (6.1)	0.65	19.9** (6.2)	1.27	29.2*** (5.7)	1.82
	Nonword fluency	0.8 (1.7)	0.13	0.5 (1.7)	0.10	2.2 (1.8)	0.42	−3.1 (4.2)	−0.25	3.0 (2.7)	0.28	10.0* (3.4)	0.90
	Oral reading fluency	−0.7 (2.0)	−0.10	0.8 (2.4)	0.13	4.0 (2.4)	0.58	−3.7 (5.0)	−0.24	5.6 (3.2)	0.40	14.8** (3.8)	1.05
	Reading comprehension	−0.3 (2.3)	−0.04	2.5 (1.7)	0.39	6.5* (2.3)	0.77	−2.3 (6.4)	−0.11	8.5** (2.7)	0.50	20.3*** (3.9)	1.14
	Listening comprehension	0.2 (8.2)	0.01	8.4* (3.3)	0.43	12.6* (4.2)	0.61	−1.4 (6.1)	−0.05	13.2* (2.7)	0.54	20.0** (6.0)	0.80
	% at benchmark	−0.9 (5.6)	−0.04	4.3 (7.1)	0.21	12.7∼ (6.6)	0.58	−6.4 (5.6)	−0.14	21.0* (7.1)	0.51	39.5** (6.6)	0.97
	Average effect size		0.03		0.34		0.73		0.16		0.58		1.11
Math	Number identification	−0.5 (1.6)	−0.08	1.3 (1.5)	0.25	4.1*** (1.0)	0.76	1.4 (1.5)	0.16	3.5* (1.3)	0.46	6.5*** (1.5)	0.82
	Quantity discrimination	3.9 (8.0)	0.19	3.1 (4.8)	0.17	8.9∼ (4.2)	0.48	4.3 (8.9)	0.15	8.6 (5.9)	0.32	13.9∼ (7.2)	0.51
	Missing number	3.8 (4.6)	0.33	1.1 (1.6)	0.11	3.3 (2.2)	0.32	4.4 (5.0)	0.26	11.3** (3.7)	0.70	18.1** (5.4)	1.07
	Word problems	7.0 (6.8)	0.34	3.8 (0.0)	0.21	0.8 (3.2)	0.04	8.5∼ (4.6)	0.34	5.5 (5.1)	0.23	3.8 (5.6)	0.16
	Addition fluency	−0.3 (0.8)	−0.07	0.1 (0.6)	0.02	1.1∼ (0.6)	0.31	0.5 (0.8)	0.12	0.1 (0.9)	0.03	1.4 (0.8)	0.34
	Subtraction fluency	0.0 (0.6)	−0.01	0.1 (0.7)	0.03	1.0∼ (0.5)	0.33	−0.1 (0.9)	−0.02	0.7 (1.1)	0.18	1.9 (1.0)	0.49
	Average effect size		0.12		0.13		0.38		0.17		0.32		0.56

∼$p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.
Parameter estimates from differences-in-differences models estimating the impact of PRIMR treatment groups in comparison with the control group. Standard errors correcting for the three-stage clustered design.

Table 5
Differences-in-differences effects of comparisons between PRIMR treatment groups on student outcomes (standard errors in parentheses).

		PD, coaching, and books vs. PD and coaching		PD, coaching, books, and teachers' guides vs. PD and books		PD, coaching, books, and teachers' guides vs. PD and coaching	
Instrument	Subcomponent	Grade 1	Grade 2	Grade 1	Grade 2	Grade 1	Grade 2
English	Letter fluency	5.1 (6.6)	6.5 (9.4)	7.1 (7.6)	9.8 (9.6)	12.1∼ (5.9)	16.3* (6.5)
	Nonword fluency	−0.3 (2.5)	5.9 (4.2)	2.9 (2.6)	4.4 (3.7)	2.6 (2.2)	10.3* (4.1)
	Oral reading fluency	1.1 (3.7)	10.0 (5.7)	4.1 (3.9)	12.1 (6.7)	5.3∼ (2.8)	22.0** (6.5)
	Reading comprehension	1.7(2.1)	1.3 (4.7)	1.7 (2.2)	7.0 (5.7)	3.3** (0.9)	8.3 (5.0)
	% at benchmark	−0.2 (6.9)	13.6 (10.3)	2.1(7.5)	22.5∼ (10.7)	1.9 (5.7)	36.1** (10.1)
Kiswahili	Letter fluency	6.7 (6.6)	8.7 (7.9)	6.7 (6.8)	9.3 (7.6)	13.4∼ (6.4)	18.0* (7.5)
	Nonword fluency	−0.3 (2.0)	6.1 (3.9)	1.7 (2.1)	7.0* (2.9)	1.5 (2.1)	13.1* (4.4)
	Oral reading fluency	1.5 (2.8)	9.3 (5.0)	3.2 (3.1)	9.3* (3.8)	4.6 (2.8)	18.5** (5.4)
	Reading comprehension	2.7 (2.8)	10.8 (6.1)	4.0 (2.8)	11.8** (3.4)	6.8∼ (3.1)	22.6** (6.7)
	Listening comprehension	8.2 (8.2)	14.6∼ (6.9)	4.2 (4.2)	6.8 (6.9)	12.4 (8.6)	21.4* (7.6)
	% at benchmark	5.2 (7.8)	27.4* (11.7)	8.4 (8.5)	18.5 (10.4)	13.6 (7.3)	45.9** (13.0)
Math	Number identification	1.8 (1.9)	2.1 (1.5)	2.7 (1.5)	3.0∼ (1.5)	4.6* (1.5)	5.1* (1.7)
	Quantity discrimination	−0.7 (8.3)	4.3 (8.1)	5.7 (4.8)	5.4 (6.2)	5.0 (8.0)	9.7 (9.1)
	Missing number	−2.7 (4.5)	6.9 (3.9)	2.2 (1.8)	6.8 (4.5)	−0.5 (4.7)	13.7* (5.6)
	Word problems	−3.1 (7.6)	−3.0 (3.6)	−3.0 (4.5)	−1.7 (4.8)	−6.2 (6.8)	−4.7 (4.4)
	Addition fluency	0.4 (0.8)	−0.4 (0.7)	1.0 (0.6)	1.3 (0.8)	1.4 (0.8)	1.0 (0.6)
	Subtraction fluency	0.1 (0.7)	0.8 (0.9)	0.9 (0.7)	1.2 (1.0)	1.0 (0.6)	2.0* (0.8)

∼$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$.
The statistical significance tests compare the magnitude of the differences-in-differences effects of the three treatment groups. Statistical significance tests in the first two columns identify differences in the magnitude of the impacts of the PD and coaching and the PD, coaching, and books treatment groups. Significance tests in the next two columns identify differences in the magnitude of the impacts of the PD, coaching, and books treatment group and the PD, coaching, books, and teachers' guides group. Finally, significance tests in the final two columns identify differences in the magnitude of the impacts of the PD and coaching and the PD, coaching, books, and teachers' guides treatment groups.

groups. Given that the discussion above examined whether the effects of each treatment group were statistically significantly different from those of the control, this analysis fit regression models comparing the treatment groups against each other. The first set of results compares the PD and coaching group with the PD, coaching, and books treatment group. Causal estimates from this comparison would reveal whether providing revised textbooks in addition to PD and coaching made a difference on learning. Results showed one statistically significant effect at the .05 level and one other at the .10 level, both of which favored the PD, coaching, and books treatment group over the PD and coaching treatment group.

The next set of values compares the PD, coaching, and books treatment group with the PD, coaching, books, and teachers' guides treatment group. This comparison examines the causal effect of adding structured teachers' guides to the textbooks and PD with coaching. Two of the regressions showed statistically significant effects at the .05 level, and two other comparisons had statistically significant effects at the .10 level. All of the significant effects favored the PD, coaching, books, and teachers' guides treatment group over the PD, coaching, and books treatment group.

The final columns compare the PD and coaching treatment group with the treatment group with PD, coaching, books, and structured teachers' guides. Nine of these regressions showed statistically significant effects at the .05 level and four others demonstrated significant effects at the .10 level. All of the significant effects favored the PD, coaching, books, and teachers' guide treatment group over the PD and coaching treatment group.

To examine the question of comparisons between the effects of the various treatment groups in greater depth, we present Fig. 1. This figure compares the average effect sizes of the three treatment groups. The PD, coaching, books, and teachers' guide treatment group had the highest average effect size for all six groupings of subject and grade. For each grouping, the PD, coaching and books treatment group had the second highest effect size. Finally, for

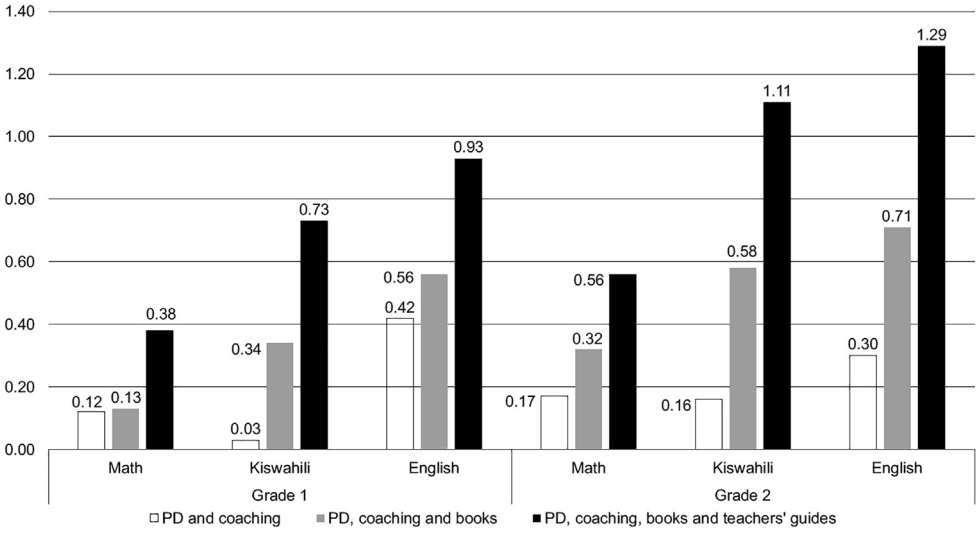

Fig. 1. Effect size comparisons by treatment group, subject, and grade.

each grouping, the PD and coaching treatment group had the smallest average effect size. Note that, as Table 5 shows, the individual effects are not always statistically significant when compared between the treatment groups.

5. Discussion

The differences-in-differences regression results showed that two of the three treatment groups had statistically significant positive learning impacts, while the PD and coaching treatment group had impacts no different from what would be experienced by chance. This suggests that providing teachers with only professional development plus instructional support results in very modest improvements. Letter sound fluency appeared to be the most sensitive to such an approach, but the other subjects and measures did not respond to this treatment. We conclude that instructional support and teacher training alone were insufficient to improve learning outcomes in the selected counties. Average effects for the PD and coaching treatment group were negligible for Kiswahili and mathematics for grades 1 and 2 (between 0.03 and 0.17 SD), and small for English grade 1 (0.42 SD) and grade 2 (0.30 SD).

The PD, coaching and books treatment group differed from the PD and coaching treatment group by the provision of new PRIMR-developed books for every child in the classroom. In this treatment group, teachers were given support and professional development in how to develop lesson plans themselves that matched those pupil books. The size of these improvements was small for mathematics grades 1 and 2 (0.13 SD and 0.32 SD, respectively), medium for Kiswahili for grades 1 and 2 (0.34 and 0.58 SD, respectively), and medium for English grades 1 and 2 (0.56 and 0.71 SD, respectively). It appears that having the combination of books for pupils, teacher PD, and instructional support was sufficient to modestly improve learning outcomes in all three subjects. These findings align with those from a recent study of 26 programs in developing countries which found that providing textbooks alongside teacher training resulted in increases in student achievement scores of 0.36 standard deviations, on average (Popova, Evans, & Arancibia, 2016).

The PD, coaching, books, and teachers' guides effects differed from those of the PD, coaching, and books treatment group because teachers were given structured teachers' guides with daily lesson plans that helped them teach using the provided pupil books. The trainings focused on how the teachers could effectively implement the lessons utilizing the partially scripted approach in the PRIMR teachers' guides. Overall, the PD, coaching, books, and teachers' guides treatment group had the greatest impacts on learning outcomes in the three subjects, and several effects were large. Average effect sizes for this treatment group ranged from 0.38 to 0.56 for mathematics, which are moderate effects; and from to 0.73 to 1.29 SD for English and Kiswahili, which are moderate to large average effect sizes. These results show that adding teachers' guides to the package of PD, instructional support, and pupil books had a dramatic impact on learning outcomes across all three subjects and in nearly all of the instrument subcomponents in the study. Teachers' guides made a significant difference to improving learning outcomes across grades and subjects.

We interpret these findings to mean, in general, that providing teachers with teacher training and instructional support alone has very modest impacts on learning outcomes. Adding a redesigned pupil book is essential, as it seems to have a much larger effect on learning outcomes and instruction than PD and coaching support alone. In many sub-Saharan African countries, it is the book that drives instruction (Commeyras & Inyega, 2007; Dubeck et al., 2012; Lee & Zuilkowski, 2015). Critically, we found the largest gains when teachers' guides with lesson plans were added to the package. In other words, while teachers' guides, and so-called scripted lesson plans, are contentious and resisted by many scholars and some teachers, the results in Kenya suggest that they caused the largest differences in learning outcomes. They appeared to work in tandem with revised books to help students substantially improve their literacy and numeracy outcomes.

Recall that our first research question simply looked at whether the three treatment groups had statistically significant effects on learning outcomes as compared to control, whereas our second research question examined whether the effects that were identified were different from each other, statistically. This analysis was somewhat underpowered, given that all three treatment groups had at least moderate effects on learning outcomes, although some of the effects were statistically insignificant. However, there were some interesting statistically significant differences in learning outcomes between the treatment groups.

Learning outcomes in many developing countries, particularly in early grade literacy and numeracy, are not sufficient to support later learning. Could the problem be solved by simply training teachers on current research-based pedagogical methods and providing them with instructional support? Our study suggests that such an approach will have limited impact. Could the problem be solved by improving the quality of books and providing them to pupils? Our study found that textbooks are an important ingredient in improved instruction, and can have a moderate impact on learning outcomes over and above PD and support. Could the problem be solved by adding teachers' guides with lesson plans? We found that adding these teachers' guides to textbooks can have a meaningful impact on learning, much more than adding textbooks alone.

Policy makers need information on costs, and we analyzed the marginal costs of each ingredient. The cost of teacher training and coaching support was constant throughout the three treatment groups, at US$5.63 per pupil for the three subjects. The PD, coaching, and books training group added pupil books at a 1:1 ratio in the three subjects delivered to the zonal training sites, at a further cost of US$2.38 per pupil. Finally, the teachers' guides added only US$0.16 per pupil to the overall cost.

The cost-effectiveness results (see Fig. 2.) suggest that for every additional US$100 spent on the PD and coaching training group, an additional 1.9 students were able to read at or beyond the government benchmark in English and 3.4 *fewer* students in Kiswahili. For every additional US$100 spent on the PD, coaching, and books treatment group, an additional 6.4 students were able to read at benchmark in English and 7.9 students in Kiswahili. Finally, for every US$100 spent on the PD, coaching, books, and teachers' guide treatment group, an additional 14.7 students were able to read at the English benchmark and 14.7 students at the Kiswahili benchmark. These results show that adding pupil books to teacher professional development and coaching created a major increase in cost-effectiveness. Moreover, the minimal cost of the teachers' guides as an ingredient resulted in a large increase in cost effectiveness, more than double in English and nearly double in Kiswahili. The teachers' guides were a remarkably cost-effective investment.

6. Conclusion

Globally, many countries are increasingly concerned about the quality of education provided to early primary grade learners. Faced with the increased awareness that learning outcomes are poor, as demonstrated by assessments such as Uwezo's annual measures (see Uwezo., 2015, for an example) and by various administrations of the EGRA (refer to examples in Gove & Cvelich, 2011), many countries have increased emphasis on learning. Given the scale of the problem, policy makers seem to be

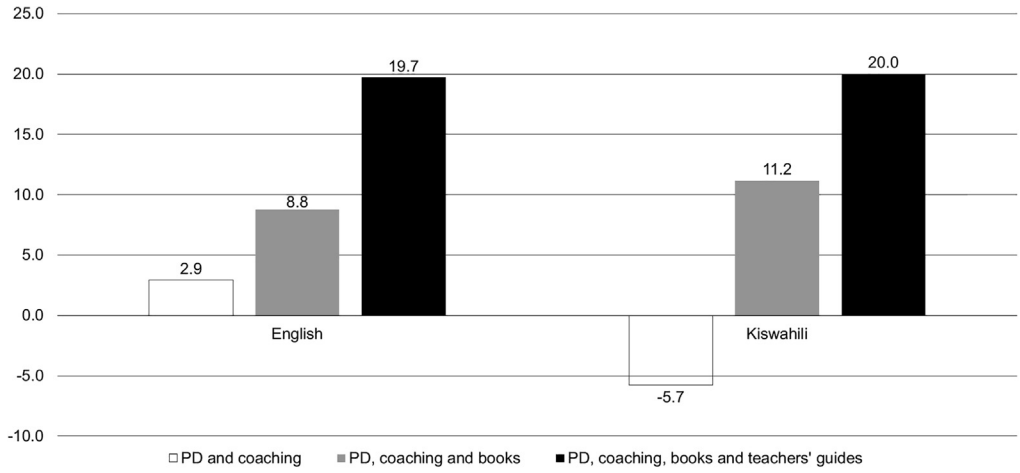

Fig. 2. Additional learners reading at MoE benchmark per US$100 spent, by treatment group.

implementing various approaches to improving literacy, such as providing teacher professional development (Kenya, Ethiopia, Malawi, Rwanda, Uganda), increasing access to tablets (Kenya), supplying textbooks at a ratio of one book per pupil (Kenya, Rwanda, Uganda), providing teachers' guides (Kenya, Ethiopia, Malawi, Uganda), and supporting teachers with classroom supervision (Kenya, Malawi, and Uganda). Many of these interventions are quite expensive, particularly supplying ongoing PD and providing books at a 1:1 ratio. Others are less expensive monetarily but have an opportunity cost, namely teacher coaching by government officers. Providing teachers' guides with lesson plans is a less costly option per pupil, but this practice is somewhat contested in the region, given some resistance to scripted lesson plans.

Without evidence as to the relative effectiveness of all of these ingredients to improved learning outcomes, countries are likely to invest in ingredients that have little impact on learning, or leave aside ingredients essential to improve outcomes in their context because of the lack of supporting evidence. The DFID PRIMR study, then, born of the desire of Kenyan policy makers to determine the most effective and cost-effective ingredients for improved learning, could serve as a starting point for education policy makers in other contexts. We encourage other countries with learning outcome results similar to those of Kenya to consider the implications of this research on the ingredients in the policies and interventions that are designed to improve learning, and invest in student books at a 1:1 ratio, but also to invest in the highly cost-effective structured teachers' guides to accompany those books.

Given our findings, we suggest that Kenya consider utilizing student books at a 1:1 ratio and teachers' guides for all three subjects, as well as instructional support and PD, to increase learning in early primary education. Encouragingly, Tusome, the national scale-up of PRIMR, is utilizing the evidence from the USAID and DFID PRIMR studies in its national implementation, as is the PRIEDE program, the scale-up of the mathematics component. Tusome and PRIEDE both use structured teachers' guides and textbooks at a 1:1 ratio, alongside of PD and coaching. The external impact evaluation of the Tusome program showed dramatic improvements in learning outcomes at the national level (Freudenberger & Davis, 2017), doubling or tripling the percentage of learners meeting Kenyan literacy benchmarks. It is worth considering how much of that very large impact was due to the policy decision by Kenyan education leaders to include structured teachers' guides in the program, even though there was some resistance to teachers' guides by some educationists in the sector.

Our study begins a strand of research focused on ingredients analysis in the education sector in the developing world. While government policy makers and educational researchers have preferences for what ingredients are most likely to be effective, they typically have limited evidence as to what works best. The DFID PRIMR results showed significant effects for 1:1 books, and additional effects for teachers' guides. The analysis suggests that a combination of coaching, professional development, pupil books, and teachers' guides is remarkably cost effective. Further research should investigate whether similar findings are found in other contexts and in either pre-primary education or later in the primary education cycle.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgments

The authors acknowledge the leadership of the Kenyan Ministry of Education, including the Tusome National Technical Team housed at the MoE and its members from across the Ministry, Teachers' Service Commission, and other organizations. The then DFID/Kenya education advisor, Sandra Barton, designed a study which asked important research questions. We appreciate the excellent editorial work of Erin Newton, and the technical support and leadership of Roxanne Zerbonia, Jessica Mejia, Joseph DeStefano, and Melinda Taylor.

Funding

The Primary Math and Reading (PRIMR) Rural Expansion program was supported by the United Kingdom Department for International Development (DFID)/Kenya from 2012 to 2015, component code 202657–108, purchase order 40064649. DFID/Kenya granted the authors permission to use the data set for additional analysis and dissemination. The initial PRIMR pilot was supported by the United States Agency for International Development (USAID/Kenya), under the Education Data for Decision Making (EdData II) Blanket Purchase Agreement, Task Order No. AID-623-M-11–00001 (RTI International Task 13). The lead author, Dr. Benjamin Piper, served as the Chief of Party on the PRIMR Initiative. The contents of the article and decisions about publication are the sole responsibility of the authors. The authors' views expressed in this publication do not necessarily reflect the views of DFID, the United Kingdom government, USAID, the United States government, or RTI International.

References

Ackers, J., & Hardman, F. (2001). Classroom interaction in Kenyan primary schools. *Compare: A Journal of Comparative and International Education, 31*(2), 245–261. https://doi.org/10.1080/03057920123230.

Albano, A. D., & Rodriguez, M. C. (2012). Statistical equating with measures of oral reading fluency. *Journal of School Psychology, 50*(1), 43–59. https://doi.org/10.1016/j.jsp.2011.07.002.

Angrist, J., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics, 19*(2), 343–369. https://doi.org/10.1086/319564.

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

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal, 111*(1), 7–34. https://doi.org/10.1086/653468.

Bressoux, P. (1996). The effects of teachers' training on pupils' achievement: The case of elementary schools in France. *School Effectiveness and School Improvement, 7*(3), 252–279. https://doi.org/10.1080/0924345960070303.

Bressoux, P., Kramarz, F., & Prost, C. (2006). *Teachers training, class size and students' outcomes: Evidence from third grade classes in France*. Unpublished paper.

Carlisle, J. F., Cortina, K. A., & Katz, L. A. (2011). First-grade teachers' response to three models of professional development in reading. *Reading & Writing Quarterly: Overcoming Difficulties, 27*(3), 212–238. https://doi.org/10.1080/10573569.2011.560482. Special issue: Research on Teacher Quality: Improving, Reading and Writing.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Commeyras, M., & Inyega, H. N. (2007). An integrative review of teaching reading in Kenyan primary schools. *Reading Research Quarterly, 42*(2), 258–281. https://doi.org/10.1598/RRQ.42.2.3.

Conn, K. (2014). *Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of rigorous impact evaluations*. New York: Columbia University.

Conn, K. M. (2017). Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of impact evaluations. *Review of Educational Research, 87*(5), 863–898. https://doi.org/10.3102/0034654317712025.

Crocco, M. S., & Costigan, A. T. (2007). The narrowing of curriculum and pedagogy in the age of accountability: Urban educators speak out. *Urban Education, 42*(6), 512–535. https://doi.org/10.1177/0042085907304964.

Donovan, M. S., & Bransford, J. D. (Eds.). (2005). *How students learn: Science in the classroom*. Washington, DC: National Academies Press.

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. https://doi.org/10.1016/j.ijedudev.2014.11.022.

Dubeck, M. M., Jukes, M. C. H., & Okello, G. (2012). Early primary literacy instruction in Kenya. *Comparative Education Review, 56*(1), 48–68. https://doi.org/10.1086/660693.

Dubeck, M. M., Okello, G., & Jukes, M. C. H. (2012). Early primary literacy instruction in Kenya. *Comparative Education Review, 56*(1), 48–68. https://doi.org/10.1086/660693.

Duflo, A. & Kiessel, J. (2012). *Teacher community assistant initiative (TCAI)*. Policy brief 4004. London: International Growth Centre, London School of Economic and Political Science. Retrieved on 08 July 2016 from http://www.theigc.org/wp-content/uploads/2015/07/Duflo-Kiessel-2012-Policy-Brief.pdf.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *The American Economic Review, 101*(5), 1739–1774. https://doi.org/10.1257/aer.101.5.1739.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics, 123*, 92–110. https://doi.org/10.1016/j.jpubeco.2014.11.008.

Evans, D. K., & Popova, A. (2015). *What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews*. Washington, DC: World Bank. https://doi.org/10.1596/1813-9450-7203.

Fehrler, S., Michaelowa, K., & Wechtler, A. (2009). The effectiveness of inputs in primary education: Insights from recent student surveys for sub-Saharan Africa. *The Journal of Development Studies, 45*(9), 1545–1578. https://doi.org/10.1080/00220380802663625.

Freudenberger, E., & Davis, J. (2017). *Tusome external evaluation—Midline report*. Prepared for the Ministry of Education of Kenya, USAID/Kenya, and the UK Department for International Development under Contract No. AID-615-TO-16-00012. Washington, DC: Management Sciences International (MSI), a Tetra Tech company. http://pdf.usaid.gov/pdf_docs/PA00MS6J.pdf.

Fuller, B. (1987). What school factors raise achievement in the Third World? *Review of Educational Research, 57*(3), 255–292. https://doi.org/10.3102/00346543057003255.

Ganimian, A. J., & Murnane, R. J. (2016). Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research*. https://doi.org/10.3102/0034654315627499.

Glewwe, P., Hanushek, E. A., Humpage, S., & Ravina, R. (2013). School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. In P. Glewwe (Ed.), *Education policy in developing countries* (pp. 13–64). Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226078854.003.0002.

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics, 1*(1), 112–135. https://doi.org/10.1257/app.1.1.112.

Gove, A., & Cvelich, P. (2011). *Early reading: Igniting education for all. A report by the Early Grade Learning Community of Practice* (Rev. Ed.). Research Triangle Park, NC: RTI International. Retrieved from https://www.rti.org/sites/default/files/resources/early-reading-report-revised.pdf.

Gove, A., & Wetterberg, A. (2011). *The Early Grade Reading Assessment: Applications and interventions to improve basic literacy*. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2011.bk.0007.1109.

Harbison, R., & Hanushek, E. A. (1992). *Educational performance of the poor: Lessons from rural Northeast Brazil*. New York: Oxford University Press.

Heyneman, S. P., & Jamison, D. T. (1980). Student learning in Uganda: Textbook availability and other factors. *Comparative Education Review*, 24(2, Part 1), 206–220. https://doi.org/10.1086/446116.

Heyneman, S. P., Jamison, D. T., & Montenegro, X. (1984). Textbooks in the Philippines: Evaluation of the pedagogical impact of a nationwide investment. *Educational Evaluation and Policy Analysis, 6*(2), 139–150 http://epa.sagepub.com/cgi/doi/10.3102/0162373700600213.

Jasti, C., Jukes, M. C., Dubeck, M. M., Elliott, & Inyega, H. N. (2016). *A child-to-child approach to improve literacy: Design of the Buddy Reading program in Kenya*. Manuscript submitted for publication.

Jukes, M. C. H., Dubeck, M. M., Adelman, E., Sheppard, M., Jasti, C., & Turner, E. (2016). *The impact of child-to-child reading on reading skills and motivation in Kenya*. Manuscript submitted for publication.

Jukes, M. C. H., Turner, E. L., Dubeck, M. M., Halliday, K. E., Inyega, H. N., Wolf, S., ... Brooker, S. J. (2017). Improving literacy instruction in Kenya through teacher professional development and text messages support: A cluster randomized trial. *Journal of Research on Educational Effectiveness, 10*(3), 449–481. https://doi.org/10.1080/19345747.2016.1221487.

Kim, Y.-S. G., Boyle, H., Zuilkowski, S. S., & Nakamura, P. (2017). *Landscape report on early grade literacy*. Washington, DC: USAID. Retrieved from https://globalreadingnetwork.net/sites/default/files/research_files/LandscapeReport.pdf.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science, 340*(6130), 297–300. https://doi.org/10.1126/science.1235350.

Kremer, M., & Holla, A. (2009). Improving education in the developing world: What have we learned from randomized evaluations? *Annual Review of Economics, 1*(1), 513–542.

Krishnaratne, S., White, H., & Carpenter, E. (2013). Quality education for all children? *What works in education in developing countries: International Initiative for Impact Evaluation*. Working Paper 20. New Delhi: International Initiative for Impact Evaluation (3ie). Retrieved on 06 July 2016 from http://www.3ieimpact.org/media/filer/2013/09/10/wp_20.pdf.

Lee, J., & Zuilkowski, S. S. (2015). 'Making do': Teachers' coping strategies for dealing with textbook shortages in urban Zambia. *Teaching and Teacher Education, 48*, 117–128. https://doi.org/10.1016/j.tate.2015.02.008.

Lockheed, M. E., Vail, S. C., & Fuller, B. (1986). How textbooks affect achievement in developing countries: Evidence from Thailand. *Educational Evaluation and Policy Analysis, 8*(4), 379–392. https://doi.org/10.3102/01623737008004379.

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda. *Journal of Policy Analysis and Management, 33*(4), 950–976. https://doi.org/10.1002/pam.21782.

Matsumura, L. C., Garnier, H. E., Correnti, R., Junker, B., & Bickel, D. D. (2010). Investigating the effectiveness of a comprehensive literacy coaching program in schools with high teacher mobility. *The Elementary School Journal, 111*(1), 35–62. https://doi.org/10.1086/653469.

McEwan, P. J. (2013). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research, 85*(3), 353–394. https://doi.org/10.3102/0034654314553127.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research, 85*(3), 353–394. https://doi.org/10.3102/0034654314553127.

Milner, H. R. (2014). Scripted and narrowed curriculum reform in urban schools. *Urban Education, 49*(7), 743–749. https://doi.org/10.1177/0042085914549685.

Ministry of Education (MoE), Republic of Kenya. (2014). *National Education Sector Plan, volume one: Basic Education Programme, rationale and approach, 2013/2014 – 2017/2018*. Nairobi: MoE. Retrieved on 06 July 2016 from http://www.globalpartnership.org/content/education-sector-plan-2013-2018-kenya.

Mohammed, R., & Kumari, R. (2007). Effective use of textbooks: A neglected aspect of education in Pakistan. *Journal of Education for International Development, 3*(1), 1–11.

National Council of Teachers of Mathematics (NCTM) [United States]. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: NCTM. Retrieved on 06 July 2016 from http://www.nctm.org/PrinciplestoActions.

Ngware, M. W., Abuya, B., Oketch, M., Admassu, K., Mutisya, M., & Musyoka, P. (2015). Randomized impact evaluation of education interventions: Experiences and lessons from a reading to learn intervention in East Africa. *International Journal of Research & Method in Education, 38*(4), 430–451. https://doi.org/10.1080/1743727X.2014.965252.

Ngware, M. W., 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. https://doi.org/10.1016/j.ijedudev.2014.01.004.

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*. Kenya Working Report, SACMEQ Educational Policy Research Series. Harare and Nairobi: Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) and Kenya Ministry of Education. Retrieved on 06 July 2016 from http://datatopics.worldbank.org/hnp/files/edstats/kenstu05.pdf.

Piper, B., & Korda, M. (2010). *EGRA Plus: Liberia. Program evaluation report*. Prepared for USAID/Liberia under the Education Data for Decision Making (EdData II) project, Early Grade Reading Assessment (EGRA): Plus Project, Task Order No. EHC-E-06-04-00004-00 (RTI Task 6). Research Triangle Park, NC: RTI International. Retrieved from http://pdf.usaid.gov/pdf_docs/pdacr618.pdf.

Piper, B., & Kwayumba, D. (2014). *The Primary Math and Reading Initiative (PRIMR). Kisumu information and communication technology (ICT) intervention endline report*. Prepared under the USAID EdData II project, Task Order No. AID-623-M-11-00001 (RTI Task 13). Research Triangle Park, NC: RTI International. Retrieved from http://pdf.usaid.gov/pdf_docs/pa00k285.pdf.

Piper, B., & Mugenda, A. (2012). *The Primary Math and Reading (PRIMR) Initiative: Baseline report*. Prepared under the USAID EdData II project, Task Order No. AID-623-M-11-00001 (RTI Task 13). Research Triangle Park, NC: RTI International. Retrieved from http://pdf.usaid.gov/pdf_docs/pa00hx75.pdf.

Piper, B., & Spratt, J. (2017, September). *Cambodia teacher professional development policy options brief*. Prepared for the United States Agency for International Development (USAID)/Cambodia under All Children Reading – Asia (ACR–Asia), All Children Reading – Cambodia, Contract No. AID-OAA-TO-16-00017, REQ-ASIA-16-00017. Research Triangle Park, NC: RTI International. Retrieved from http://pdf.usaid.gov/pdf_docs/PA00N2RT.pdf.

Piper, B., King, S., & Mugenda, A. (2016, September; revised from 2014). *USAID/Kenya Primary Math and Reading (PRIMR) Initiative: Endline impact evaluation – revised edition*. Prepared under the USAID EdData II project, Task Order No. AID-623-M-11-00001 (RTI Task 13). Research Triangle Park, NC: RTI International. Retrieved from https://globalreadingnetwork.net/eddata/usaidkenya-primary-math-and-reading-primr-initiative-endline-impact-evaluation-revised.

Piper, B., Jepkemei, E., & Kibukho, K. (2015). Pro-Poor PRIMR: Improving early literacy skills for children from low-income families in Kenya. *Africa Education Review, 12*(1), 67–87. https://doi.org/10.1080/18146627.2015.1036566.

Piper, B., Ralaingita, W., Akach, L., & King, S. (2016). Improving procedural and conceptual mathematics outcomes: Evidence from a randomised controlled trial in Kenya. *Journal of Development Effectiveness, 2016*. https://doi.org/10.1080/19439342.2016.1149502.

Piper, B., & Zuilkowski, S. S. (2015). Teacher coaching in Kenya: Examining instructional support in public and nonformal schools. *Teaching and Teacher Education, 47*, 173–183. https://doi.org/10.1016/j.tate.2015.01.001.

Piper, B., Zuilkowski, S. S., Kwayumba, D., & Strigel, C. (2016). Does technology improve reading outcomes? Comparing the effectiveness and cost-effectiveness of ICT interventions for early grade reading in Kenya. *International Journal of Educational Development, 49*, 204–214. https://doi.org/10.1016/j.ijedudev.2016.03.006.

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. https://doi.org/10.1016/j.ijedudev.2014.02.006.

Piper, B., Zuilkowski, S. S., & Ong'ele, S. (2016). Implementing mother tongue instruction in the real world: Results from a medium-scale randomized controlled trial in Kenya. *Comparative Education Review, 60*(4), 776–807. https://doi.org/10.1086/688493.

Platas, L. M., Ketterlin-Gellar, L., Brombacher, A., & Sitabkhan, Y. (2014). *Early Grade Mathematics Assessment (EGMA) toolkit*. Research Triangle Park, NC: RTI International.

Popova, A., Evans, D. K., & Arancibia, V. (2016). *Inside in-service teacher training: What works and how do we measure It?* Working paper. Retrieved on 10 July 2016 from http://www.riseprogramme.org/sites/www.riseprogramme.org/files/Evans%20Inside%20In-Service%20Teacher%20Training%20-%20CLEAN%20-%20v2016-06-22.pdf.

Ravitch, D. (2015, August 3). *Are scripted lessons in for-profit schools "an audacious answer" for poor kids in Africa and Asia?* [Web log post]. Retrieved on 06 July 2016 from https://dianeravitch.net/2015/08/03/are-scripted-lessons-in-for-profit-school-an-audacious-answer-for-poor-kids-in-africa-and-asia/.

Read, T. (2011). *Learning and teaching materials: Policy and practice for provision*. London: Department for International Development.

Reeves, J. (2010). Teacher learning by script. *Language Teaching Research, 14*(3), 241–258. https://doi.org/10.1177/1362168810365252.

Reubens, A. (2009). *Pilot of the Early Grade Mathematics Assessment: Final report*. Prepared for USAID under the Education Data for Decision Making (EdData II) project, Task Order No. EHC-E-02-04-00004-00 (RTI Task 2). Research Triangle Park, NC: RTI International. Retrieved on 06 July 2016 from http://pdf.usaid.gov/pdf_docs/PNADS440.pdf.

Riddell, A. R., & Nyagura, L. M. (1991). *What causes differences in achievement in Zimbabwe's secondary schools?* Washington, DC: World Bank. Retrieved on 10 July 2016 from http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/1991/06/01/000009265_3961001140949/Rendered/PDF/multi_page.pdf.

RTI International (2015). *The Primary Math and Reading (PRIMR) Initiative DFID/Kenya Rural Expansion Programme, Bungoma and Machakos endline study*. Prepared for DFID/Kenya under Contract 202657–108. Research Triangle Park, NC: Research Triangle Institute.

Sabarwal, S., Evans, D., & Marshak, A. (2013). *The permanent textbook hypothesis: School inputs and student outcomes in Sierra Leone*. Retrieved on 06 July 2016 from http://cega.berkeley.edu/assets/cega_events/61/4C_Inputs_to_Education.pdf.

Somerset, A. (2011). Strengthening educational quality in developing countries: The role of national examinations and international assessment systems. *Compare: A Journal of Comparative and International Education, 41*(1), 141–144. https://doi.org/10.1080/03057925.2011.534851.

Teemant, A. (2014). A mixed-methods investigation of instructional coaching for teachers of diverse learners. *Urban Education, 49*(5), 574–604. https://doi.org/10.1177/0042085913481362.

Tilson, T., Kamlongera, A., Pucilowski, M., & Nampota, D. (2013). *Evaluation of the Malawi Professional Development Support (MTPDS) Program: Final evaluation report*. Prepared for USAID under Requisition No. REQ-612-12-000029. Washington, DC: Social Impact, Inc., Retrieved on 06 July 2016 from http://pdf.usaid.gov/pdf_docs/PDACX458.pdf.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2015). *EFA global monitoring report*. Paris: UNESCO. Retrieved on 06 July 2016 from http://en.unesco.org/gem-report/report/2015/education-all-2000-2015-achievements-and-challenges#sthash.e2C5th7M.dpbs.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2016). *Every child should have a textbook*. Global Education Monitoring Report, Policy Paper 23. Paris: UNESCO. Retrieved on 06 July 2016 from http://unesdoc.unesco.org/images/0024/002433/243321E.pdf.

United Nations (2015). *Transforming our world: The 2030 agenda for sustainable development*. New York: United Nations.

Uwezo. (2015). *Are our children learning? The state of education in Kenya in 2015 and beyond*. Fifth Uwezo annual learning assessment report. Narobi: Twaweza East Africa. Retrieved on 06 July 2016 from http://www.twaweza.org/uploads/files/kenya.pdf.

Wanzare, Z., & Ward, K. L. (2000). Rethinking staff development in Kenya: Agenda for the twenty-first century. *International Journal of Educational Management, 14*(6), 265–275. https://doi.org/10.1108/09513540010348052.

Yara, P. O., & Otieno, K. O. (2010). Teaching/learning resources and academic performance in mathematics in secondary schools in Bondo District of Kenya. *Asian Social Science, 6*(12), 126–132. https://doi.org/10.5539/ass.v6n12p126.

Zuilkowski, S. S., & Piper, B. (2017). Instructional coaching in Kenya: Supporting teachers to improve literacy outcomes. In M. Akiba & G. LeTendre (Eds.), *The Routledge international handbook on teacher quality and policy* (1st ed., pp. 505–516). New York: Routledge.