Journal of Educational Change
https://doi.org/10.1007/s10833-018-9325-4

Scaling up successfully: Lessons from Kenya's Tusome national literacy program

Benjamin Piper[1] · Joseph Destefano[2] · Esther M. Kinyanjui[3] · Salome Ong'ele[1]

© The Author(s) 2018

Abstract

Many successful piloted programs fail when scaled up to a national level. In Kenya, which has a long history of particularly ineffective implementation after successful pilot programs, the Tusome national literacy program—which receives funding from the United States Agency for International Development—is a national-level scale-up of previous literacy and numeracy programs. We applied a scaling framework (Crouch and DeStefano in Doing reform differently: combining rigor and practicality in implementation and evaluation of system reforms. International development group working paper no. 2017-01, RTI International, Research Triangle Park, NC, 2017. https://www.rti.org/publication/doing-reform-differently-combining-rigor-and-practicality-implementation-and-evaluation) to examine whether Tusome's implementation was rolled out in ways that would enable government structures and officers to respond effectively to the new program. We found that Tusome was able to clarify expectations for implementation and outcomes nationally using benchmarks for Kiswahili and English learning outcomes, and that these expectations were communicated all the way down to the school level. We noted that the essential program inputs were provided fairly consistently, across the nation. In addition, our analyses showed that Kenya developed functional, if simple, accountability and feedback mechanisms to track performance against benchmark expectations. We also established that the Tusome feedback data were utilized to encourage greater levels of instructional support within Kenya's county level structures for education quality support. The results indicated that several of the key elements for successful scale-up were therefore put in place. However, we also discovered that Tusome failed to fully exploit the available classroom observational data to better target instructional support. In the context of this scaling framework, the Tusome literacy

The authors were involved in the implementation of the USAID Tusome Early Grade Reading Activity and the Primary Math and Reading Initiative that preceded it. One author (Esther Kinyanjui) served as the National Coordinator of Tusome while working for the Ministry of Education, another author (Dr. Benjamin Piper) was the Chief of Party of Tusome and PRIMR, and another author (Salome Ong'ele) currently serves as the Chief of Party of Tusome after previously being the Deputy Chief of Party of Tusome and PRIMR. The authors appreciate the generous funding of USAID Kenya as well as the focused leadership of the Ministry of Education.

Extended author information available on the last page of the article

Published online: 11 July 2018

program's external evaluation results showed program impacts of 0.6–1.0 standard deviations on English and Kiswahili learning outcomes. The program implemented a functional classroom observational feedback system through existing government systems, although usage of those systems varied widely across Kenya. Classroom visits, even if still falling short of the desired rate, were far more frequent, were focused on instructional quality, and included basic feedback and advice to teachers. These findings are promising with respect to the ability of countries facing quality problems to implement a coherent instructional reform through government systems at scale.

Keywords Literacy · Systems · National · Evaluation · Reading · Implementation · Reform

Introduction

Several countries have recently begun large-scale educational interventions to respond to low learning outcomes. The evidence used to justify national or large-scale programs typically has emanated from pilot initiatives that increasingly have provided rigorous evidence of impact on learning outcomes. For example, several recent meta-analyses showed a dramatic increase in results from pilot programs focused on improving learning outcomes while also using rigorous research designs (Conn 2017; McEwan 2015; Murnane and Willett 2011).

Although an increased dependence on causal evidence to justify large-scale implementation strengthens the research base for large programs, the decisions to take these programs to scale have largely been undertaken without a body of literature that examines the barriers to successful large-scale educational implementation or whether the external validity assumptions for these comparisons hold (Bates and Glennerster 2017). This failure to develop robust scale-up literature and practice has allowed the field of educational development to focus heavily on proof of concept, with randomized controlled trial (RCT) studies estimating the program impact of small- or medium-scale interventions in several contexts. We found a handful of recent studies and research that made explicit a model by which evidence from pilot programs should interact with scale-up designs. However, this evidence is too recent to influence the design decisions made by governments. Instead, based on the guidance of educational researchers, governments often depend on pilot evidence as sufficient for scale-up decisions, without undertaking a thorough investigation of the aspects of system capacity that will be most critical to producing similar impact at scale.

The scale-up literature is pessimistic about both the initial take-up of educational actors and the long-term impact on learning outcomes, to say the least. Scale-up efforts in a variety of areas have proven ineffectual, including school-based teacher professional development initiatives (Chege 2011), cluster-based teacher support initiatives (Piper 2009), in-service teacher professional development programs (Villegas-Reimers 1998), the development of head teachers' instructional leadership

skills (Cambridge Education Consultants 1998), the expansion of mathematics and science instructional capacity in East Africa (Kamau et al. 2014), tablet-based interventions for children (Christia et al. 2014), tablet-based school support (Doykos et al. 2015), and television-based classroom learning programs (Assefa 2017). We wish to emphasize how infrequently large-scale educational interventions have had any statistically significant impact on teacher practice (Jukes et al. 2017) and, as a result, how often such interventions have had a negligible impact on student learning.

Some recent exceptions to this pessimistic view of the scalability of educational reforms exist. The Gauteng Primary Language and Mathematics Strategy program in Gauteng Province in South Africa sustained increased learning outcomes across several grades over a few years (Fleisch et al. 2016). Escuela Nueva has shown ongoing improvement in teaching and learning in Colombia and a dozen other countries, mostly in Latin America, and at increasingly large scale (Rincón-Gallardo and Fleisch 2016). Working outside of government systems, Pratham's work has consistently shown significant effects on literacy in thousands of schools and communities in India (Banerji and Chavan 2016). These examples of successful scale-up were found in a recent issue of the *Journal of Educational Change*,[1] book-ended by reflections on improving outcomes at scale by Fullan (2016) and Elmore (2016). However, even when reflecting on the large-scale impacts identified in these and other programs in the special issue, Elmore (2016) questioned the entire enterprise of focusing on systematic instructional reform given the complexity of having individual teachers determine whether to implement the program behind closed doors.

The large-scale interventions reviewed in the special issue of the *Journal of Educational Change* used various implementation designs. The studies showed that each program faced project-specific hurdles to achieving program impact. Some of the programs failed because they did not sufficiently account for the demand side of the intervention, whereas others failed because they were overly technological solutions to instructional programs. Still others did not succeed because they expected teachers to invest more time and effort than they were willing to provide on a consistent basis, and some failed because they were implemented outside of government systems and, therefore, did not change the day-to-day practice of education officials.

However, the experiences of the programs reviewed in the *Journal of Educational Change* did point to a set of factors or conditions under which core pedagogical practices can be changed at scale (from Elmore's 1996 study, also cited in Rincón-Gallardo 2016). Attention to the instructional core is one of them, including setting up the means to articulate and communicate what good pedagogy looks like. Another factor involves reinforcing that core of instructional practice by developing new normative structures and putting in place organizational relationships that intensify teachers' motivation to adopt those practices. Systems also need structures to promote, provide ample practice with, and thus reproduce the new pedagogical approaches; as well as incentive systems that are tied to those structures.

[1] *Journal of Educational Change* special issue: Volume 17, issue 4, November 2016, https://link.springer.com/journal/10833/17/4/page/1.

Fullan and Quinn (2016) argued that the difference between failure and success in large-scale education system improvement is determined by whether interventions focus on the right or wrong drivers of reform. They classified as "wrong drivers" interventions focused on punitive accountability, individualism, technology, and fragmented policies. Their "right drivers" included capacity building, pedagogy, and systemic policies (Fullan and Quinn 2016). Attention to pedagogy as the driver of improved learning outcomes (whether at the micro or system level) was reiterated by the findings of the International Initiative for Impact Evaluation's (3ie's) summary review of the impact of education programs, which showed that the largest and most consistent positive effects on learning resulted from programs that used structured approaches to improving pedagogy (Snilstveit et al. 2016). The experience in Gauteng Province in South Africa mentioned above also demonstrated the importance of structured approaches to pedagogy, with additional emphasis placed on the institutional infrastructure needed to support those approaches (Fleisch et al. 2016). That infrastructure includes a variety of high-quality learning materials, implemented in combination with support to teachers (in the form of ongoing coaching) that enables them to utilize those learning materials and that establishes an environment of professional instructional accountability (Fleisch 2017). This notion of institutional infrastructure is consistent with Pritchett's (2015) idea of a system that is "coherent for learning," in which the key principal-agent relationships that govern the operation of the education system are aligned to and supportive of specific improvements in teaching and learning (Pritchett 2015, p. 4).

While Fullan (2016) also argued that there are limitations to an overly prescriptive approach to pedagogy, the evidence from 3ie (Snilstveit et al. 2016), McKinsey and Company's theoretical model (Mourshed et al. 2010), and the United States Agency for International Development's (USAID's) *Landscape Report on Early Grade Literacy* (Kim et al. 2016) indicates that developing countries trying to improve very low-functioning education systems should focus on introducing highly specific approaches to instructional change and putting in place the supports needed for that change to be realized across large numbers of classrooms. In Kenya, Piper et al. (2018) examined the relative impact on learning and cost-effectiveness of training, coaching, 1:1 literacy materials for children, and structured teachers' guides. They found that the entire package resulted in much greater impacts than books and training or training alone. This growing body of evidence makes a strong case for focusing on curriculum, pedagogy, and materials aligned around a structured sequence of lessons; supporting teachers through relatively frequent coaching in specific instructional techniques, including how to use new materials; and regularly monitoring teacher adoption of those practices and students' learning outcomes.

Informed by this recent literature's attention to how education systems do or do not assemble the ingredients for large-scale improvements in learning outcomes, very recent efforts have begun to expand our understanding of large-scale program impact. Five reports in particular have driven a revision of the educational development field's thinking regarding scale-up. First, the *Millions Learning* research by the Brookings Institution (Robinson et al. 2016) examined how small pilot programs expanded into more and larger settings. However, that research paid little attention to how and whether the programs were embedded in government systems, which is,

in many contexts, a prerequisite for successful large-scale implementation. Second, a framework developed by Management Sciences International (MSI) (Cooley and Kohl 2006) on scale-up provided a typology for how scaling happens. This typology has driven some donor-funded programs to at least address the scalability of interventions. However, few program designs have systematically applied MSI's complex and linear framework; therefore, its relevance for national or large-scale implementation in education is difficult to assess. Third, a manuscript examining large-scale program implementation in the literacy realm was intended to help USAID missions design and manage such programs for sustainability (DeStefano and Healey 2016). This theoretical framework drew on the *Millions Learning* research and the MSI frameworks mentioned above and focused specifically on the challenges associated with improving at scale the teaching of reading in the early grades. However, this framework has only nominally been applied to existing programs, and it has yet to be used to guide the initial design or implementation of large-scale education programs. Fourth, the special issue of the *Journal of Educational Change* reported several recent attempts to implement instructional change at scale, with an opening essay describing the contours of a systematic review of the scale-up literature (Rincón-Gallardo and Fleisch 2016). Fifth and finally, applying the principles of "doing development differently" (The DDD Manifesto Community 2017) to the challenge of achieving large-scale improvements in learning outcomes, Crouch and DeStefano (2017) identified a core set of capacities that education systems should focus on to shift away from bureaucratic administration toward the development of management relationships that are coherent for learning (as argued for by Pritchett 2015) and able to support large-scale instructional change. Crouch and DeStefano (2017) defined the three core functions that education systems need to conduct rapid cycles of planning, action, reflection, and revision: (1) setting and communicating expectations, (2) monitoring and guaranteeing accountability for meeting those expectations, and (3) intervening to ensure the support needed to assist students and schools that are struggling.

Much of this literature on scale-up has focused on how innovations are introduced and then taken to scale. The emphasis is on identifying the conditions needed for innovation to spread and be adopted or adapted in a growing number of sites. However, often, the challenge is not simply how to support or promote the spread of innovation. Successfully scaled programs require building the institutional capacity of an education system to itself first enable a new approach to be implemented at scale and then to support and sustain schools as they work to incorporate that approach into their day-to-day operations. Therefore, we propose an analytical approach that examines how one educational reform program currently being implemented at national scale, the Kenya Tusome Early Grade Reading Activity, has performed in its attention (or lack thereof) to the core functions defined by Crouch and DeStefano (2017). We will demonstrate whether and how Tusome has reinforced education system capacity to better fulfill these core functions. The comparison of this framework to the available evidence in Kenya is supported by an analysis of the Tusome literacy program's impact on student learning in Kenya.

Tusome, the national literacy program of the Kenyan Ministry of Education, was designed and implemented in a context of growing interest in large-scale program

implementation. Based on rigorous evidence from previous research on improving literacy outcomes in Kenya (Gove et al. 2017), and funded by USAID, Tusome began in January 2015 with a formal launch by Kenyan President Uhuru Kenyatta. Tusome supported all of Kenya's Class 1 (grade 1) students with materials in English and Kiswahili and training for teachers in 2015, expanded to all of Kenya's Class 2 students in 2016, and expanded again to Class 3 in 2017. What makes Tusome a plausible research subject for the scaling literature is that it drew on the evaluations of several pilot studies in Kenya and enjoyed the benefits of careful design focused on large-scale implementation possibilities. Before Tusome, the Primary Math and Reading (PRIMR) Initiative was a 547-school research study supported by USAID/Kenya and undertaken by the Ministry of Education with technical support from RTI International. Its intent was to assess whether a low-cost intervention implemented primarily through the existing infrastructure could support sufficient initial take-up by teachers to result in increased learning outcomes (Piper and Zuilkowski 2015). Additional pilot conditions were tested in the PRIMR Rural Expansion Programme, funded by the UK Department for International Development (DFID) Kenya, in 847 schools, with additional treatment groups assigned to test various conditions essential to improving outcomes (Piper et al. 2016a, e, 2018). Note that the PRIMR intervention and the subsequent Tusome program focused on a particular set of literacy skills that can be assessed orally and relatively simply, rather than emphasizing a broader definition of literacy's ability to change an individual's situation even as it changes the individual (Bartlett 2008) or the potential of literacy to support political freedom (Friere 1970). Some of the large-scale interventions described in the special issue of the *Journal of Educational Change* included a slightly broader definition of what literacy improvement means, including the Escuela Nueva program in Colombia (Colbert and Arboleda 2016) and the Learning Community Project in Mexico (Rincón-Gallardo 2016).

Tusome was one of the first experiences of taking a piloted literacy program to national scale through government systems. The findings from the pilots were used to design key elements of the national-scale implementation (Piper et al. 2015b, 2016b), and several other large-scale literacy programs were then designed based on Tusome's experience, even before clear evidence of the program's scalability or impact was documented. External evaluation data on Tusome's impact and both school- and classroom-level take-up by individual education actors (including teachers) and learning outcomes data from Tusome's first year of implementation are now in hand (baseline results, MSI 2015; midline results, Freudenberger and Davis 2017). Thus, Tusome's transition from a medium-scale pilot to national-scale implementation in 23,000 public primary schools and 1500 low-cost private schools [called Alternative Provision of Basic Education and Training (APBET) institutions in Kenya] offers an interesting case study for the application of the Crouch and DeStefano (2017) core functions framework.

Kenya as the site

Kenya is a lower-middle-income country with a per capita annual gross domestic product of US$1455 (World Bank 2017a) and a population estimated at 48 million (World Bank 2017b). The 2010 Constitution transitioned Kenya from a centralized system of service delivery to a system decentralized at the county level to increase accountability (Republic of Kenya 2010). The education sector was the largest social sector that did not decentralize financial control, making investigating scaling education a more complex exercise than in other sectors in Kenya.

Low learning outcomes is an area of increasing concern and interest in Kenya. An Uwezo household survey assessment (Uwezo 2016) showed no national-level changes in literacy outcomes from 2010 to 2014, despite the implementation of more than 25 interventions across Kenya in 2014 (Kibukho, personal communication, October 15, 2014). In addition, data from Early Grade Mathematics and Reading Assessments (EGMA and EGRA respectively)[2] showed that 40% of pupils could not read or comprehend what they read by the end of Class 2 (Piper and Mugenda 2012) and that numeracy outcomes in both conceptual and procedural tasks (Piper et al. 2016c) were lower than Ministry of Education-established performance benchmarks (Reubens 2009). In 2012, on average, only 5% of Class 2 (grade 2) children reached the benchmark fluency rates for comprehension in Kiswahili, and 7% did so in English (Piper and Mugenda 2012).

Kenya has long been a location for rigorous studies examining the impact of educational programs that attempt to work at scale (Bold et al. 2013; Kremer et al. 2009). Kenya is the country in sub-Saharan Africa with the highest number of studies cited in the Abdul Latif Jameel Poverty Action Lab's (J-PAL's) database of educational interventions (https://www.povertyactionlab.org/evaluations). In fact, Kenya trails only India and the United States in the number of evaluations in J-PAL's database.

Kenya's status as a preferred research site for pilot programs, however, has not protected it from also being the setting for failed large-scale initiatives, including basic education programs of various designs. The implementation of the 2002 curriculum struggled because of its adherence to relatively outdated instructional routines (Kenya Institute of Curriculum Development, KICD 2017). School-based teacher professional development programs struggled because of a perception that the key resource teachers were not chosen by merit (Chege 2011). The government's system of instructional support through Teachers' Advisory Centre (TAC) tutors suffered from the burden of administrative tasks, and as a result, by the beginning of Tusome's implementation in 2015, very few classrooms were being supported by this professional cadre, outside of donor-funded programs (Zuilkowski and Piper 2017). The Strengthening of Primary Education program for head teachers to provide classroom support failed to develop incentives for this instructional leadership function (Cambridge Education Consultants 1998). A large-scale literacy and

[2] For more about the construction, content, and administration of the EGMA and EGRA instruments, see RTI International (2014, 2016).

numeracy intervention in Western Kenya struggled to have teachers consistently implement the program, leading to no impact or negative results (Lucas et al. 2014). An instructional improvement program called Literacy Boost had an impact when managed by a nongovernmental organization but none when managed by the Ministry of Education (Bold et al. 2013). Clearly, Kenya is a context open to research and experimentation, but success at scale has remained elusive.

The Tusome intervention

Background

In 2012, Kenya began implementing the PRIMR Initiative, which was the precursor to the Tusome intervention (2014–2019). PRIMR encompassed two separate research programs (one funded by USAID and the other funded by DFID Kenya), organized into a set of RCTs with various treatment groups to determine the most cost-effective means to improve early literacy and numeracy. The PRIMR model required that the actual training and classroom support be done by existing government officers and that research be undertaken to understand whether and how these officers would be able to accommodate the PRIMR activities in their daily work. This is an important consideration that many pilot programs do not take into account (Gove et al. 2017). Local government officers' attention and time are sparse, as they face many competing demands. Many small programs that have had statistically significant effects have been unable to be scaled given inattention to this inherent weakness. The PRIMR design showed 0.2–1.0 standard deviation (SD) effects on literacy and 0.1–0.6 SD effects on numeracy (Piper et al. 2016c, 2018). The literacy results showed somewhat larger effect sizes for PRIMR's impacts on reading skills than on comprehension outcomes, but PRIMR did achieve consistent improvements in reading comprehension in both languages. Furthermore, PRIMR's findings indicated that coaches did improve the literacy program and that 15:1 was a more cost-effective school-to-coach ratio than 10:1 (Piper and Zuilkowski 2015); that learning impacts were possible after only 1 year (Piper et al. 2014); that the impact of PRIMR was sufficient to reduce the poverty gap (Piper et al. 2015a); that performance on reading assessments administered in mother tongues could also be improved, even without mother tongue instruction (Piper et al. 2016f); that the most cost-effective information and communication technology intervention was tablets for coaches (rather than for students or teachers; Piper et al. 2016a, e); and that a package of teachers' guides and learner books was more cost-effective than programs that offered only training without these materials (Piper et al. 2018). To the credit of the Kenyan Ministry of Education, Tusome was designed according to the research evidence collected from PRIMR.

The Tusome program design was structured to implement the most cost-effective elements that PRIMR showed would be able to work through government systems.

The learning materials (student textbooks and teachers' guides) in Tusome were adapted and refined from PRIMR's, the coaching model was the same, and the teacher training model was similar (although only 6 days per year were affordable at the national scale, compared with 10 days of teacher professional development under PRIMR). However, while PRIMR was able to provide consistent ongoing support to Curriculum Support Officers (CSOs)[3] implementing the program, the sheer size of Tusome precluded this arrangement. Therefore, whether Tusome would realize higher levels of take-up and implementation than found in other large-scale education projects was unclear. We examined Tusome's scale-up implementation outcomes to determine whether the core functions of the education system enabled pedagogical change to be implemented consistently enough to produce widespread improvements in learning (Crouch and DeStefano 2017). The analysis presented below compares the impact of Tusome to that of the pilot program that preceded it and describes the conditions within Tusome that supported or hindered achieving results at scale.

It should be noted that Tusome is focused on a particular set of program elements that were determined at the national level and are being adhered to at the local level. Some might argue that such an approach is inadequate to describe the complex processes that occur in policy implementation (Datnow and Park 2009). For Tusome, given that the research used to design the program provided significant evidence that a successfully implemented version of the previously piloted program would be effective, this criticism might be less relevant. Better understanding how the Tusome initiative was received by local actors and actualized by individual educators far from the national-level policy maker is essential to fully understanding the impact of Tusome at scale.

The core functions framework

Pilot programs often achieve success on a small scale by intervening directly to change instruction in a given set of teachers or schools. System-level change requires identifying the mechanisms that will ensure that the majority of schools and teachers adopt the new strategy. Crouch and DeStefano (2017) argued that an education system's set of core functions determine whether classroom practice can be improved at scale. These functions are essential to increase the likelihood of all schools succeeding in any new education reform effort. Their key contributions were the notions that the institutional capacity to implement varies in scaled programs and that the contours and content of that institutional capacity are "what makes an education system a 'system' and not just a collection of schools" (Crouch and DeStefano 2017, p. 8). These core functions are as follows:

- *Setting expectations for the outcomes of education* Decision-making at all levels should be driven by clearly articulated (and communicated) expectations for

[3] In 2016, the TAC tutor responsibilities and credentials were revised, and the position was renamed "Curriculum Support Officer." More discussion about this change appears later in the text.

what students should learn and be able to do, for what teachers should do on a day-to-day basis, and for what others in the system should do to support teachers and students to achieve those expectations.

- *Monitoring and holding schools accountable for meeting those expectations* At the local or decentralized level, the education system should monitor whether expectations are being met. For example, are the basic inputs of learning materials and administrative requirements being provided to schools in a timely manner, and are schools and teachers using those inputs in ways that promote increased learning? Finally, are students demonstrating the expected levels of skill development?
- *Intervening to support the students and schools that are struggling to meet expectations and holding the system accountable for delivering that support* If schools and teachers are going to be evaluated and held accountable according to their ability to meet the expectations, the system must ensure that schools, teachers, and students have the basic inputs they need. Furthermore, the system needs to be able to intervene in response to any weaknesses that monitoring the schools reveals. Schools that are succeeding should be recognized, and those that are struggling should be given additional help in a timely and supportive way. Rather than administering the status quo of inequitable learning performance, this core function requires that systems recognize the need for and be able to deliver differentiated responses based on school and student performance.

The focus on education system capacity to assure these three core functions constitutes a deliberately limited framework. The setting of goals, provision of inputs, and monitoring of progress rely on a technical-rational perspective of policy development and implementation. Using this rationalistic framework allows us to examine ways in which the more complex features of education change are addressed in the case of Tusome in Kenya. Attempting to shift the normative culture of the education bureaucracy, and thereby changing the nature of the relationships among teachers, school directors, and their supervisors, presents an inherently complex web of challenges. Using a framework focused on a specific set of technical implementation challenges may provide some insight into what enables the more complex dimensions of system change to be addressed. However, it must be acknowledged that this perspective does not address all the socio-political and other contextual factors that should be taken into account to improve national education systems.

Research questions

Given the limited evidence regarding whether and how large-scale interventions are implemented through existing government education systems, we applied the Crouch and DeStefano (2017) core functions framework to answer the following research questions.

- RQ1: Has Tusome reinforced system capacity to set expectations for the outcomes of education?
- RQ2: Has Tusome reinforced system capacity to monitor and hold schools accountable for meeting these expectations?
- RQ3: Has Tusome reinforced system capacity to provide the inputs necessary to implement the reading program at scale, to intervene to support students and schools that are struggling to meet expectations, and to enforce accountability for delivering these types of support?
- RQ4: Given the ability of the system to fulfill its core functions, what has been the impact of Tusome on learning outcomes?

Methodology

Sources of data

To answer the research questions posed above, we systematically examined several sources of data, from within the Tusome project and from a recently completed external evaluation. Tusome program documents provided an account of how benchmarks were established for student performance in reading (and math), and training documents and official reports recounted how expectations for teachers and students were defined and communicated (RQ1). Implementation of the national program included introducing the use of tablet-based computers to systematically record classroom observations and snapshots of student progress in reading skill development. Data gleaned from a "dashboard" (data visualization tool) that centralizes, compiles, and performs first-order analysis of those findings were reviewed to show how the system has been monitoring and holding itself accountable for meeting expectations (RQ2). Internal project monitoring information, derived in part from ongoing observations by project staff of the performance of ministry officials, enabled us to evaluate how well the education system is fulfilling its responsibilities vis-à-vis teachers and schools (RQ3). Lastly, the external evaluation provided data on learning outcomes showing whether student performance in reading was improving compared with baseline performance before the at-scale implementation of Tusome (RQ4).

Applying the core functions framework to Tusome

As noted earlier, to evaluate the Tusome literacy program's contribution to system capacity at scale, we fit the core functions framework to the Tusome program. The Tusome program has included building the capacity of the Kenyan education system in each of the core functions mentioned in the model; therefore, we analyzed whether the core functions are operating effectively. The Tusome program has also worked alongside the relevant Kenyan government institutions to develop their capacity to set and communicate expectations, to have CSOs (or in the case of APBET schools, project-hired instructional coaches) monitor instructional change

and learning outcomes, and to ensure basic inputs and ongoing, increasingly differentiated support for schools. Below, we examine the evidence from Tusome related to Kenya's ability to fulfill each of these functions.

Findings

RQ1: Setting expectations

Research question 1 asks whether Tusome made expectations for learning outcomes clear. To answer this question, we investigated the program's design documents and the training manuals used in coach and teacher professional development. We found clear evidence that Tusome communicated the Ministry of Education's benchmarks for literacy outcomes within the design documents and training materials, but more importantly, we found that those benchmarks were widely disseminated, understood, and reinforced. With the support of PRIMR and, later, Tusome, the Ministry of Education has defined expectations for reading achievement by setting benchmarks for proficient reading in both Kiswahili and English in the first years of primary school (Piper et al. 2016b), as shown in Table 1 below. Analyzing Tusome training documents from two national-scale trainings in 2015, three national-scale training programs in 2016, and two more national-scale trainings in 2017 showed that teachers were trained on the existence of the literacy benchmarks and that actual data from Kenya were consistently shared during termly trainings to demonstrate and reinforce how Kenya's children were performing relative to these expectations. Evidence from training reports generated during these sessions suggests that the benchmarks were typically understood by the participants and that they could recall the benchmarks and what the literacy results meant by grade and language in terms of the progress being made (or failing to be made).

Second, instructional expectations are embodied in the teachers' guides distributed to teachers across the program. The guides (and the training in using the guides) explicitly communicate to teachers the sequence of lessons they are to implement to build the reading skills expected by the national primary curriculum. The external evaluation results showed that these teachers' guides were widely available, with 95% of teachers at the national level using the guides during classroom observations (Freudenberger and Davis 2017).

Table 1 Expectations (benchmarks) for student oral reading fluency performance

Language	Benchmark levels, in correct words per minute	
	Emergent	Fluent
Kiswahili	17	45
English	30	65

These benchmarks were established and approved by the Kenya National Examinations Council (KNEC), with advice from the PRIMR and Tusome project teams

Third, the Tusome training program encouraged teachers to use their guides consistently to deliver high-quality lessons. The emphasis in Tusome was on program fidelity, and during training, the fact that teachers should expect to see progress toward literacy benchmarks based on their consistent use of the lessons as laid out in the teachers' guides was specifically highlighted. To reinforce this message, classroom support was similarly focused on the daily implementation of the lessons in the teachers' guides.

Fourth, CSOs were expected to regularly visit schools and were given tools and training to enable them to monitor and support teachers in adopting the instructional methods and using the provided Tusome learning materials. During these visits, CSOs uploaded the results of a benchmark assessment designed by the Ministry of Education to be specific to the material covered each term and administered to three children in each classroom.[4]

To determine whether Tusome was able to support Kenyan government officers to focus on fidelity, we reviewed whether the instructional support function had been internalized by the government system. Tusome's work has been done primarily by government officers, and a key focus has been working with the Teachers' Service Commission to reinvigorate the classroom instructional support role of the CSOs. CSOs represent the front line of the public education system in Kenya—the place where the government's administrative apparatus interacts most directly with schools.

Under the PRIMR pilot program, the project supported the Teachers' Service Commission to evaluate and reform the role of the TAC Tutors (now CSOs). In 2016, using a new job description for CSOs, the TSC re-interviewed and rehired CSOs under a job description that focused more specifically on instructional support. This job description emphasized visiting schools regularly, observing teachers' lessons, and providing constructive feedback on their instructional practice—a significant departure from the previous de facto role of CSOs that stressed supervising school administration.

Under Tusome, CSOs were trained three times a year, given tablets with instructional support tools to help teachers, and reimbursed for travel to enable them to visit schools, record their observations, advise teachers and head teachers, and do spot assessments of student progress. In the 2016 academic year, CSOs across Kenya recorded and uploaded 113,604 classroom lesson observations at schools during the seven key instructional months in the Kenyan calendar, corresponding to an average of more than 90 visits per CSO; this rate of school supervision and support far surpassed what the education system was doing before (Kisirkoi 2012). Encouragingly, the 2017 observation data indicated even more classroom visits than in the prior year. The more frequent visits, and the observation of teachers during visits, communicated and reinforced the new expectations for teachers. Having CSOs show up regularly to observe their lessons let teachers know that the education system expected them to deliver a specific sequence of lessons and employ the instructional techniques on which they had been trained.

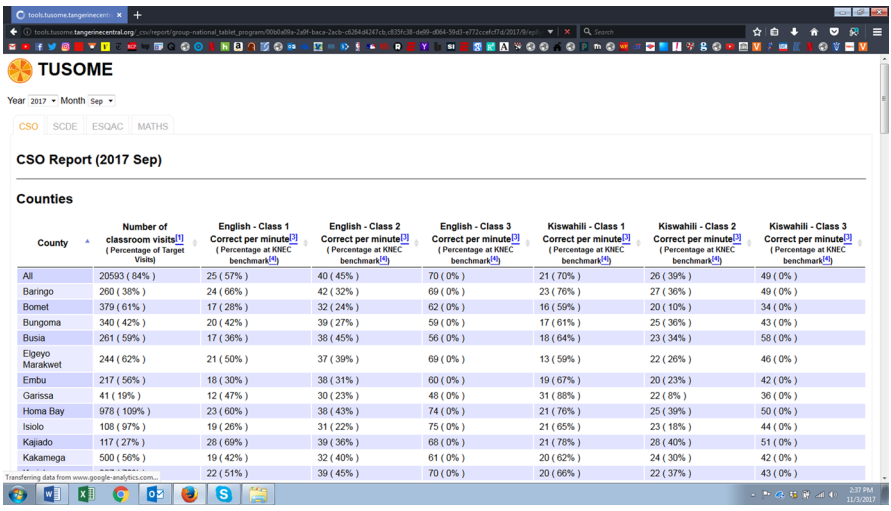
[4] Kenya's academic calendar is divided into three terms: Term 1, January–April; Term 2, May–August; and Term 3, September–November.

RQ2: Monitoring progress

To determine whether Tusome supported the government's core function of monitoring progress toward the new system, teacher, and student expectations, we reviewed program documents and the large-scale database of classroom outcomes shared in the Tusome dashboard. The work of the CSOs described above communicated and reinforced expectations and monitored the extent to which those expectations were met. We found that Tusome had built system capacity to monitor expectations in four important ways.

First, as mentioned above, at the end of each classroom observation, the CSOs randomly selected and evaluated the oral reading fluency of three students. Although a small number of students were sampled at each visit, cumulatively, over the course of several visits within a sub-county or county, the number of students added up to a reasonably representative sample for that geographic area. These data were centralized and viewable on the dashboard, which presented both national and county-level data, as shown in the sample in Fig. 1. The dashboard is shared each term with all CSOs, in front of the Ministry of Education's County Directors of Education to whom they are responsible; as we have argued elsewhere (Piper et al. 2017a), this is a meaningful expansion of the accountability function of the Ministry of Education.

The dashboard in Fig. 1 shows the average oral reading fluencies and percentages of students meeting performance benchmarks in English and Kiswahili in Classes 1–3 in each county. For example, as of September 2017, across all the counties in Kenya, 57% of Class 1 students tested were reading at or above the benchmark in English, and 70% were meeting the benchmark in Kiswahili. However, when we look at each county, this dashboard shows considerable variation, especially in performance on the reading fluency measure, with some counties performing


TUSOME

Year 2017 Month Sep

CSO SCDE ESQAC MATHS

CSO Report (2017 Sep)

Counties

County	Number of classroom visits[1] (Percentage of Target Visits)	English - Class 1 Correct per minute[2] (Percentage at KNEC benchmark[4])	English - Class 2 Correct per minute[2] (Percentage at KNEC benchmark[4])	English - Class 3 Correct per minute[2] (Percentage at KNEC benchmark[4])	Kiswahili - Class 1 Correct per minute[2] (Percentage at KNEC benchmark[4])	Kiswahili - Class 2 Correct per minute[2] (Percentage at KNEC benchmark[4])	Kiswahili - Class 3 Correct per minute[2] (Percentage at KNEC benchmark[4])
All	20593 (84%)	25 (57%)	40 (45%)	70 (0%)	21 (70%)	26 (39%)	49 (0%)
Baringo	260 (38%)	24 (66%)	42 (32%)	69 (0%)	23 (76%)	27 (36%)	49 (0%)
Bomet	379 (61%)	17 (28%)	32 (24%)	62 (0%)	16 (59%)	20 (10%)	34 (0%)
Bungoma	340 (42%)	20 (42%)	39 (27%)	59 (0%)	17 (61%)	25 (36%)	43 (0%)
Busia	261 (59%)	17 (36%)	38 (45%)	56 (0%)	18 (64%)	23 (34%)	58 (0%)
Elgeyo Marakwet	244 (62%)	21 (50%)	37 (39%)	69 (0%)	13 (59%)	22 (26%)	46 (0%)
Embu	217 (56%)	18 (30%)	38 (31%)	60 (0%)	19 (67%)	20 (23%)	42 (0%)
Garissa	41 (19%)	12 (47%)	30 (23%)	48 (0%)	31 (88%)	22 (8%)	36 (0%)
Homa Bay	978 (109%)	23 (60%)	38 (43%)	74 (0%)	21 (76%)	25 (39%)	50 (0%)
Isiolo	108 (97%)	19 (26%)	31 (22%)	75 (0%)	21 (65%)	23 (18%)	44 (0%)
Kajiado	117 (27%)	28 (69%)	39 (36%)	68 (0%)	21 (78%)	28 (40%)	51 (0%)
Kakamega	500 (56%)	19 (42%)	32 (40%)	61 (0%)	20 (62%)	24 (30%)	42 (0%)
		22 (51%)	39 (45%)	70 (0%)	20 (66%)	22 (37%)	43 (0%)

Transferring data from www.google-analytics.com...

Fig. 1 Sample Tusome dashboard display, September 2017. *Note* as of 2018, KNEC was in the process of establishing Class 3 fluency benchmarks for the first time. Thus, the dashboard shows only 0% placeholder values for Class 3 percentages at the KNEC benchmark, for both languages

particularly poorly. How data such as these are or are not being used to direct and target support resources is discussed below.

Second, in addition to monitoring student outcomes, Tusome equipped and trained CSOs to monitor teacher adoption and application of the desired instructional practices. At each visit, the CSO recorded which lesson the teacher was delivering that day, noted whether each component of the lesson was well delivered, and indicated whether the teacher employed the techniques on which he or she had been trained. This information formed the basis for the feedback offered by the CSO and was aggregated to generate a picture of whether teachers in a district were meeting expectations for improved instruction. In addition, the Tusome technical team often followed the CSOs to the classroom and to their subsequent one-on-one sessions with teachers and then gave feedback on the quality of the CSOs' instructional support. Based on those visits, the Tusome field staff shared daily updates with the appropriate government officers on teachers' progress in delivering the expected sequence of daily lessons. Education officers then gave instructions to the CSOs—and head teachers—to support those teachers who were falling behind in Tusome lesson delivery.

Figure 2 presents the progress through the sequence of lessons during classroom observations in October 2016, the last month of Tusome's second year of implementation. The figure shows small variations by grade and by language, but on average, Class 1 and 2 teachers had taught more than 124 Tusome lessons by that point in the year, meaning that they were within 10 lessons (i.e., two school weeks) of where they should have been. The lowest quartile of teacher performance on this measure corresponded to those teachers who were 4 weeks behind what would be appropriate for October, which is the middle of Term 3.

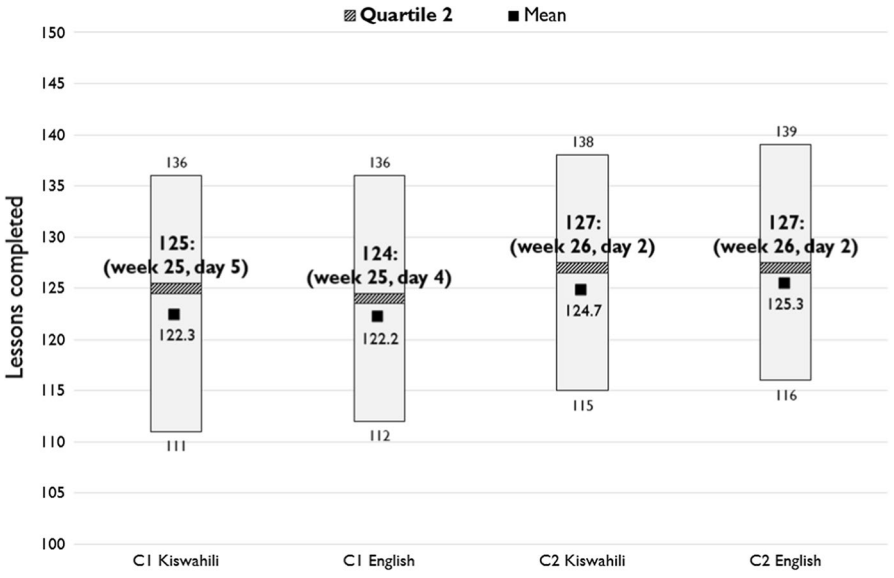


Fig. 2 Number of lessons completed as of the end of the last lesson observed in October 2016. *Note* C1, C2 = Class 1 or 2

While CSOs monitored teachers and assessed students as described above, the CSOs were also monitored. Each time a CSO visited a school, the tablet they used to record their visit and lesson observations took a global positioning system (GPS) reading to verify the location. The classroom observation data and GPS location were then uploaded and aggregated into monthly reports showing whether the CSO made the target number of visits to that location every month.

The data on the number of classroom visits were used to determine the travel reimbursement each CSO received—depending on whether they met the target number of visits *for each school* (i.e., not simply visiting the nearest schools multiple times while ignoring harder-to-reach locations). This type of monitoring, coupled with the explicit incentive created by linking travel allowances to the fulfillment of specific school-visit responsibilities, greatly increased the accountability of CSOs for meeting the expectations placed on them for school and teacher support. The accountability structure of Tusome also expanded beyond the CSOs, with the dashboard showing the percentages of target visits at the county and national levels. These data were consistently used by Ministry of Education leadership to draw attention to and increase accountability for instructional support. The Tusome program has supported the use of these data by both national-level education leaders and the county-level officers who are much closer to the actual CSOs whose behavior the program is trying to influence.

Fourth, our analysis of the Tusome data suggested that the accountability structures available in Tusome have driven demand for similar accountability structures in other programs. Kenya's Primary Education Development project (2015–2019), an early grade mathematics program operating with funds from the Global Partnership for Education, has created mathematics benchmarks for Classes 1 and 2 and a similar tablet-based, GPS-tracked tool for CSO observations of mathematics classrooms. Furthermore, the quality assurance and standards system, which serves as an inspectorate in Kenya, also created a set of tablet-based tools and a dashboard for their system. Finally, the sub-county directors of education, similar to district education officers, were provided with a Ministry of Education-designed tool that is available on the Tusome dashboard platform. The demand for benchmarking and accountability systems in these other areas is evidence that Tusome is helping Kenya institutionalize mechanisms to monitor and reinforce expectations for decentralized school support, teacher instructional behavior, and student outcomes.

The demand for the accountability structures in Tusome is encouraging, but implementation still requires improvement. For example, the number of visits by CSOs to mathematics classrooms typically corresponded to just over 10% of the number of visits to Tusome-supported literacy classrooms. Moreover, quality assurance officers and sub-county directors of education undertook only a fraction of the expected numbers of class and school visits, and most of these officers did not upload observation data consistently.

RQ3: Providing differentiated support

If the education system in Kenya is indeed creating new expectations for teachers and students, then the system must also provide teachers and students the resources

they need. The external evaluation of Tusome (Freudenberger and Davis 2017) showed that Tusome had assisted the Ministry of Education in ensuring provision of the basic inputs necessary to implement an early grade reading program at scale. In more than 97% of classrooms across the country, teachers had the teachers' guides containing the lesson plans for their grade, 99% of classrooms had a 1:1 ratio of books to students in both subjects, and teachers reported having participated in training on early grade reading instruction frequently, with the median teacher reporting having been involved in four to five training sessions. The provision of coaching, as mentioned above, was also being assured, albeit with somewhat less fidelity than the provision of training and materials. More than 80% of Class 1 and 2 teachers reported being observed by a CSO at least once per term (compared to the official target of three times per term).

Data gathered from ongoing support visits and aggregated at the county and then zonal level indicate the instructional challenges teachers are facing as they work to adopt the Tusome approach and could be used to better structure future trainings and immediate support. Our findings were mixed in the area of providing differentiated support. We found that, in fact, Tusome used the large-scale data collected from classrooms to inform the design of trainings that were responsive to the needs of teachers. Specifically, we found that the classroom observation briefs collected by Ministry of Education officers and Tusome staff inspired a robust email-based discussion of these findings on a daily basis. These data included specific feedback to individual CSOs on how to respond to the challenges that teachers encountered during their observations and summaries of the challenges the teachers described. The trainings organized during the term breaks took those challenges into account and could be focused in response to teachers' needs within each county and zone.

While the data above showed how Tusome responded to monitoring data via its aggregate training programs, the review of the Tusome program showed no evidence that data were used to inform county- or zone-specific interventions when the data showed weak levels of implementation. The monitoring data did reveal to each CSO the schools in their jurisdiction where teachers were not meeting instructional expectations and/or where students were underperforming in reading. The dashboard mentioned above could be used to indicate to individual CSOs which schools need more support and to Tusome and Ministry of Education officers which CSOs need support. However, the evidence suggests that this support was not provided frequently, if at all, and that no plan based on classroom visit data was implemented to help CSOs know which teachers to visit more frequently.

RQ4: Impact on learning outcomes

To answer RQ4, we reviewed the external evaluation report prepared by MSI (Freudenberger and Davis 2017) to investigate the impact of Tusome on learning outcomes. Given that Tusome was implemented nationally and that there is no counterfactual, the external evaluation report compared the learning outcomes in a nationally representative sample of schools between 2015 and 2016. Table 2 shows the impact of Tusome on Kiswahili reading outcomes (for an explanation of

Table 2 Kiswahili average scores on reading subtasks at baseline and midline. *Source*: Adapted from Freudenberger and Davis (2017, Table 2, p. 4)

Subtask	Class 1				Class 2			
	Baseline	Midline	Difference	Effect size	Baseline	Midline	Difference	Effect size
Letter sound knowledge	16.6	29.7	13.1*	0.75	16.2	39.7	23.4*	1.32
Syllable fluency	11.0	21.5	10.4*	0.66	20.9	37.5	16.6*	0.80
Invented/nonword decoding	4.7	8.3	3.6*	0.45	10.2	16.1	5.8*	0.50
Oral reading fluency	4.9	12.2	7.3*	0.75	13.5	24.5	11.0*	0.71
Reading comprehension	0.4	0.9	0.5*	0.62	1.1	2.0	1.0*	0.69
Listening comprehension	1.2	2.0	0.8*	0.52	1.9	2.9	0.9*	0.52
Average effect size				0.63				0.76

*Statistically significant difference at the $p < 0.01$ level

the reading subtasks assessed at midline, see Annex III of Freudenberger and Davis 2017, p. 65). The external evaluation assessed children in both Class 1 and Class 2 on six Kiswahili subtasks. The data showed statistically significant increases in national-level learning outcomes for each of the six subtasks and for both Classes 1 and 2. The table also shows the effect sizes of the differences between baseline and midline. The effect sizes were 0.50 SD or higher for all estimates, except for invented word decoding, which was 0.45 SD in Class 1. The average effect size for Class 1 was 0.63 SD, and that for Class 2 was 0.76 SD. These effects are considered large in the education literature that looks primarily at small-scale pilot programs (J-PAL 2018) and are significantly larger than the Kiswahili effects in the PRIMR pilot studies (Piper et al. 2016b).

Table 3 presents the impact of Tusome on English in Classes 1 and 2 and was drawn from findings presented in Freudenberger and Davis (2017). Similar to Kiswahili, the results showed statistically significant gains in each of the eight tasks and in both grades. The magnitude of the impacts was typically large, with effect sizes above 0.50 SD except for vocabulary (Classes 1 and 2) and a short reading comprehension task (Classes 1 and 2). The average effect size for English Class 1 was 0.67 SD, and for Class 2, it was 1.04 SD. These impacts were significantly larger than that of PRIMR on English outcomes determined in previous studies (Piper and Zuilkowski 2015; Piper et al. 2014, 2018).

Encouragingly, the percentage of Class 2 children who met the emergent benchmarks increased from 34 to 65% for English and from 37 to 66% for Kiswahili (Freudenberger and Davis 2017), nearly doubling both proportions. These are massive gains for research on educational interventions in general but even more so considering that Tusome is implemented at the national level.

Discussion

Our intention in this analysis was to apply a framework to determine whether key elements of education system institutional capacity were contributing to the successful large-scale implementation of a particular set of instructional changes. Kenya's Tusome program provided a relevant example because it represents a fully national-scale implementation of a successful pilot program aimed at improving the teaching and learning of reading in English and Kiswahili. The Crouch and DeStefano (2017) framework, when applied to Kenya, enabled us to recount how changes in three core education system functions were contributing to the ability of the education system in Kenya to ensure instructional improvement. The first three research questions we posed at the beginning of this paper led us to investigate whether Tusome had reinforced system capacity to

- set and communicate expectations;
- monitor and hold schools accountable for meeting expectations; and
- provide basic inputs, intervene to support students and schools that were struggling, and hold the system accountable for delivering those inputs and supports.

Table 3 Average scores on English reading subtasks at baseline and midline. *Source*: Adapted from Freudenberger and Davis (2017, Table 1, p. 4)

Subtask	Class 1				Class 2			
	Baseline	Midline	Difference	Effect size	Baseline	Midline	Difference	Effect size
Phoneme segmentation	1.1	3.8	2.6*	1.07	0.6	5.0	4.5*	2.57
Letter sound knowledge	15.1	26.3	11.3*	0.71	10.2	32.6	22.4*	1.63
Invented/nonword decoding	5.7	10.4	4.7*	0.52	10.4	18.6	8.3*	0.68
Vocabulary	5.9	7.8	1.9*	0.48	8.2	10.2	1.9*	0.41
Oral reading fluency (short)	10.6	22.3	11.7*	0.67	23.8	43.6	19.9*	0.72
Reading comprehension (short)	0.2	0.5	0.3*	0.40	0.5	1.0	0.5*	0.49
Oral reading fluency (long)	9.7	22.0	12.4*	0.73	21.8	44.2	22.5*	0.86
Reading comprehension (long)	0.2	0.8	0.6*	0.75	0.6	1.7	1.2*	0.94
Average effect size				0.67				1.04

*Statistically significant difference at the $p < 0.01$ level

The analyses presented in this paper allowed us to answer RQ1, and we found that expectations had successfully been created for the acquisition of specific literacy competencies in the first years of primary school. Expected outcomes for students had been codified in the form of benchmarks for oral reading fluency and comprehension in English and Kiswahili that students should achieve by the end of Classes 1 and 2 (Piper et al. 2016b). In addition, the learning materials used in Tusome were structured around a sequence of lessons designed to develop those competencies (Mourshed et al. 2010). The evidence suggests that expectations were created for how teachers would follow that sequence of lessons by using the teachers' guides for their classes (Piper et al. 2017b). Teacher training and the materials provided to teachers and students communicated those instructional and curricular expectations. However, we found little evidence of whether and how these expectations were communicated to teachers outside of Tusome-specific activities, although anecdotal evidence gathered by researchers and Tusome staff suggested that it did happen intermittently.

With respect to the second research question on whether and how the Tusome reform was able to monitor and hold schools accountable, we found positive results. The regular visits by CSOs reinforced teachers' adherence to the sequence of lessons in the teachers' guides and generated data on student learning progress. It is becoming increasingly clear for all involved that progress in delivering lessons and progress in student performance are important initial indicators for the system. Somewhat surprisingly, we found less evidence that the *quality* of the CSO feedback made a difference in the implementation of Tusome in classrooms (Zuilkowski and Piper 2017). We did find examples where low aggregated fluency scores from CSO visits were used to create specific interventions for particular teachers or schools. Enabling CSO visits to include structured feedback to teachers relative to expected instructional practice and having those visits generate student fluency data are two examples of how the Ministry of Education's institutional capacity was being developed to monitor and hold schools accountable—a positive indication of progress toward the second of the core functions undertaken in Tusome.

Interestingly, school- and student-level accountability under Tusome is not typically expressed through sanctions imposed for not meeting expectations, for either individual teachers or CSOs. We hypothesize that the accountability is being expressed in terms of a change in organizational culture rather than a more stringent punitive accountability model. Elmore (1996), reflecting on why incentives alone do not lead to dramatic uptake of instructional change in education systems, reasoned that unless the prevailing social norms under which teachers operate are also changed, the incentives will not induce the desired behaviors in ways that economists might predict. Having CSOs (and their head teachers at their schools) visit and conduct structured observations of teachers' classroom instruction on a dramatically more frequent and regular basis represents a profound change in the prevailing norms under which teachers and education officials typically work in sub-Saharan Africa. For years, if not decades, schools and teachers have operated in virtual isolation from the education system—not receiving classroom support (Zuilkowski and Piper 2017) or system- or county-level oversight and, therefore, accepting the prevailing norm that they are not actually expected to perform their job up to any

reasonable standard (such as, most fundamentally, showing up regularly to teach). Early signs in Kenya suggest that the systems for regular monitoring introduced by Tusome are leading to this type of normative shift, even without hard negative consequences (or, conversely, remunerative rewards) being tied to the outcomes of the CSO monitoring data.

Our analysis of RQ3 showed that providing support to schools and teachers that is more responsive to their expressed needs is a higher-order capacity, which the education system in Kenya has only begun to manifest. On the positive side, we found that training content was taking into account what coaching visits to schools revealed about teacher take-up. However, support resources—whether in the form of more attention from CSOs or other parts of the system or in the form of compensatory allocations that would enable schools to add staff, materials, or supplementary instructional opportunities—were not being directed based on what data were revealing about school, teacher, and student performance. Such an approach to the provision of support would require the education system to accept that the pursuit of equitable outcomes demands inequitable allocation of support resources. This approach would also oblige the system to have the data management capacities to consistently review the findings of the data (which appeared to be the case in many counties in Kenya) and to also apply resources to respond to those findings. The second aspect of this balancing act remains elusive in Kenya, but the potential is at least there.

The changes in system management that Tusome has been able to support could be perceived as limited in that they focus on adherence to what Fullan (2016) may characterize as an overly prescriptive pedagogical approach. Indeed, the basic inputs and instructional changes (i.e., use of structured teachers' guides) supported at scale under Tusome fall short of the "deep change" exemplified by the dispersed and collective engagement in the construction of innovation at the local level in programs such as Escuela Nueva or the Learning Communities Project discussed in the special issue of the *Journal of Educational Change* (Fullan 2016). However, one can consider that in a large education system, such as in Kenya, shifting the culture to expect the provision of basic inputs, the use of available class time to assure the delivery of basic instruction, and the provision of regular follow-up and monitoring could be an important first step in achieving deeper, more collective and, thus, more sustainable improvement. The authors' experiences suggest that focusing on specific instructional strategies is what enables the cultural shift to take place. Continuing to monitor and learn how a large system such as Kenya's moves from being able to deliver a structured and somewhat "prescriptive" model to delivering one that is "deeper" would contribute to further understanding how bureaucracies can come to emulate more dispersed, collective, and networked systems [such as argued for by Pritchett (2013) and described as a "starfish"].

Our fourth research question investigated the combined ability of the Kenyan system to provide these three core functions and its impact on learning outcomes. The results showed that Tusome's impact was large and meaningful. The gains in Tusome after 1 year exceeded both the gains made in other neighboring countries over several years and those of the preceding pilot programs, particularly the PRIMR RCT research used to design Tusome (Freudenberger and Davis 2017). The effect sizes, which were

greater than 0.7 SD for both English and Kiswahili at the national level, were remarkable in terms of both their magnitude and their equitable distribution. Tusome impacts on the ultimate outcome of reading comprehension were somewhat smaller than in the other tasks, and somewhat lower in English than Kiswahili. This suggests that careful consideration should be given to the expansion of oral language and vocabulary skills in a language that is a second language—or even more frequently, a third or fourth language—for most children in Kenya (Piper et al. 2016d).

The external evaluation results showed that the effects were relatively stable across the four lowest wealth quartiles and that the causal gains were similar in the low-cost private schools (i.e., APBET institutions) and public schools that were part of the sample (Freudenberger and Davis 2017). Note, however, that the PRIMR RCT research had treatment and control groups, whereas the Tusome external evaluation results presented in Freudenberger and Davis (2017) simply compare year-on-year gains without a counterfactual. Even with this caveat, the gains were remarkably large.

Compared to the impacts identified in a study of 12 separate RCTs in Uganda looking at the impact of literacy improvements over a period of either 1, 2, or 3 years, the impacts of Tusome in Kenya were larger than those for any of the languages in Uganda (Brunette et al. 2018). They also were larger than those achieved in the national program in Rwanda (Education Development Center, EDC 2017) and under similarly designed pilot programs in South Africa (Fleisch et al. 2016; Taylor and von Fintel 2016). The fact that the gains in Tusome were achieved at the national level is striking.

The external evaluation did not involve the qualitative research necessary to produce a detailed set of explanatory characteristics for these meaningful impacts, but it did show that Tusome's provision of student books, teachers' guides, teacher professional development, and classroom support was similarly high across the country. It also showed that the teachers perceived the program as successful, and the data from classroom observations revealed a meaningful percentage of teachers progressing through the sequence of lessons at a pace consistent with almost daily adherence to the program in both languages. The combination of the institutionalized capacity mentioned above with materials and training (which benefited from several rounds of revisions because of the PRIMR experimental pilots implemented since 2011) appears to have positioned Kenya to see large effects at scale. Although the external evaluation results did not include the data needed to test this proposition, national-level ownership, from the Presidency to the Ministry to the county-level leadership, appeared to be important. Additional research should examine these conditions in more depth to reveal how this ownership and leadership were experienced in local settings in Kenya and, thus, to understand the relationship between that local level decision-making and the assurance of the core functions. In addition, more implementation analyses are needed in this sector. Indeed, the Tusome example shows that understanding and focusing on teacher behavior change, the dissemination of changes in the normative environment for teachers and administrators, and increased accountability at the local level are essential elements of successful large-scale implementation (Datnow and Park 2009).

Conclusion

When considering why the impact of Tusome on learning outcomes over 1 year was somewhat larger than the gains experienced in the much smaller (but still substantial) pilot programs that preceded it (Piper et al. 2014, 2018), we note that the gains are both counterintuitive and outliers considering the literature on the typical impacts achieved when a reform moves from pilot to scale (e.g., Moore et al. 2017). We posit that the relatively successful implementation of the first two areas of the core functions model allowed the program to focus on particular aspects of institutional capacity relating to these areas. This focus appears to have ensured a high level of implementation fidelity in materials provision, teacher professional development, and even the more difficult task of instructional support. The fidelity of implementation, in turn, had an impact on the normative culture in the education sector and galvanized support around a limited, but important, set of behaviors for which key actors saw positive effects. These initial impacts, demonstrated by tangible evidence available at the local level, created a virtuous cycle that overcame the initial resistance by some in the sector and encouraged those who experienced changes in their classrooms, their schools, and their zones to focus more heavily on implementation fidelity in their locality. If this change process is what happened in Kenya in 2015 and 2016, potential remains for additional growth as Kenya attempts to more systematically address the third core function. Similarly, the focus on the utilization of the teachers' guides in the initial stages of Tusome may eventually give way to a more open approach to instructional support, in which expert teachers can expand above and beyond what the basic model will provide, allowing, perhaps, for even greater impact on learning (Piper et al. 2018).

The findings of this analysis suggest that there is a place in educational research for both the medium-scale RCT initiatives that underpinned the initial design of Tusome and rapid and large-scale implementation and fidelity analyses, such as the one presented here. The special issue of the *Journal of Educational Change* cited here showed that in the field of international education, research interest is growing in understanding not only whether programs are working but also whether the theoretical perspectives that underpin scale-up and expansion efforts are borne out in reality. Recall from the discussion above that within the special issue, Elmore (2016) was pessimistic about the exercise of attempting national-scale interventions at all. Our findings suggest that Elmore's pessimism may have been unfounded. When countries like Kenya make radical implementation decisions, invest in the resources needed to follow these decisions and policy steps with robust teacher support structures that focus on reforming the core of their instructional practice, there is hope for large-scale instructional reform after all. Research on implementation and fidelity is important not only for analyzing whether programs were ultimately successful but also for providing programs with rapid analysis of whether the conditions essential to successful large-scale implementation are present so that course corrections can be made.

It is unfortunate to tout the provision of foundational inputs, such as materials and basic teacher professional development, as a significant accomplishment of an education system, but that is the reality faced by teachers and students in many

developing countries, including Kenya. Too often, and too consistently, books in the hands of children and training for their teachers are not adequately assured, and very seldom are teachers provided any meaningful or consistent instructional support. Teacher professional development—when it is delivered at all—is of low quality, too theoretical, or mistimed. Quantities of materials too often are insufficient to allow every child to hold and use their own book. Failure to provide these basic supports to schools substantially erodes the potential impact of innovations as they are taken to scale. The midline evaluation of Tusome showed that these basic supports were being provided to schools, teachers, and students at scale (Freudenberger and Davis 2017). The bigger challenge for the education system going forward is developing the capacity to ensure ongoing support of the type that enables teachers and students to make good day-to-day use of those basic inputs. On this count, the dramatic improvement in the frequency of school visits by CSOs is evidence that the Kenyan education system is also developing this capacity. Visits, even if they fell short of the desired rate, were more frequent, were focused on what was going on in classrooms, and included some feedback and advice to teachers. Whether those CSO visits add as much value to instruction as they could remains an open question. CSOs themselves are not expert instructional coaches and have only had some training intended to reorient them away from a default tendency to assume a supervisory/inspectorate stance in their relationships with teachers. The evolution of the system's capacity to provide deep, rich support to classrooms remains a future challenge. Our current analysis of the development of system capacity to provide support to schools indicated that instructional change is not being driven by any substantial ongoing coaching technical input but is, rather, a first-order result of the changed norms around learning expectations and the provision of the basic inputs alluded to above. What the learning gains in a country such as Kenya could be if the more sophisticated elements of the third core function were in place remain to be seen.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Assefa, T. (2017). Educational technology implementation in Ethiopian high schools: Benefits and challenges of the instructional plasma TV. In A. Marcus-Quinn & T. Hourigan (Eds.), *Handbook on digital learning for K-12 schools* (pp. 413–427). New York, NY: Springer. https://doi.org/10.1007/978-3-319-33808-8_25.

Banerji, R., & Chavan, M. (2016). Improving literacy and math instruction at scale in India's primary schools: The case of Pratham's Read India program. *Journal of Educational Change, 17*(4), 453–475. https://doi.org/10.1007/s10833-016-9285-5.

Bartlett, L. (2008). Literacy's verb: Exploring what literacy is and what literacy does. *International Journal of Educational Development, 28*(6), 737–753. https://doi.org/10.1016/j.ijedudev.2007.09.002.

Bates, M. A., & Glennerster, R. (2017). The generalizability puzzle. *Stanford Social Innovation Review, Summer,* 50–54.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). *Scaling up what works: Experimental evidence on external validity in Kenyan education.* Working paper no. 321, Center for Global

Development, Washington, DC. Retrieved June 18, 2018, from https://www.cgdev.org/publication/scaling-what-works-experimental-evidence-external-validity-kenyan-education-working.

Brunette, T., Piper, B., Jordan, R., & Nabacwa, R. (2018). *Differential impacts of mother tongue instruction on literacy outcomes: Understanding what works at scale in Uganda.* Manuscript under review.

Cambridge Education Consultants. (1998). *Strengthening primary education in Kenya: An evaluation of the strengthening primary education project 1992–1996. (EV629)*. London: DFID.

Chege, M. P. (2011). *Effects of School Based Teacher Development (SBTD) programme on teaching-learning process in public primary school in Makuyu Division, Kenya*. Doctoral dissertation, University of Nairobi, Nairobi, Kenya.

Christia, J., Czerwonko, A., & Garofalo, P. (2014). Does technology in schools affect repetition, dropout and enrollment? Evidence from Peru. *Journal of Applied Econometrics, 17*(1), 89–111. https://doi.org/10.1016/s1514-0326(14)60004-0.

Colbert, V., & Arboleda, J. (2016). Bringing a student-centered participatory pedagogy to scale in Colombia. *Journal of Educational Change, 17*(4), 385–410. https://doi.org/10.1007/s10833-016-9283-7.

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.

Cooley, L., & Kohl, R. (2006). *Scaling up—from vision to large-scale change. A management framework for practitioners*. Washington, DC: MSI.

Crouch, L., & DeStefano, J. (2017). *Doing reform differently: Combining rigor and practicality in implementation and evaluation of system reforms.* International development group working paper no. 2017-01, RTI International, Research Triangle Park, NC. Retrieved June 18, 2018, from https://www.rti.org/publication/doing-reform-differently-combining-rigor-and-practicality-implementation-and-evaluation.

Datnow, A., & Park, V. (2009). Conceptualizing policy implementation: Large-scale reform in an era of complexity. In D. Plank, B. Schneider, & G. Sykes (Eds.), *AERA handbook on education policy research* (pp. 348–361). New York: Routledge.

DeStefano, J. & Healey, F. H. (2016). *Scale-up of early grade reading programs.* Prepared for USAID under the Education Data for Decision Making (EdData II) project, Data for Education Programming in Asia and the Middle East (DEP/AME), task order no. AID-OAA-BC-11–00001 (RTI Task 15), RTI International, Research Triangle Park, NC. Retrieved June 18, 2018, from http://pdf.usaid.gov/pdf_docs/PBAAF430.pdf.

Doykos, B., Silvernail, D. L., & Johnson, A. F. (2015). *Preliminary examinations of the relationships between the use levels of Maine learning technology initiative devices and school-level poverty*. Portland, ME: Maine Education Policy Research Institute.

EDC. (2017). *Literacy, language, and learning initiative (L3). National fluency and mathematics assessment of Rwandan schools: Endline report.* Prepared for USAID under cooperative agreement AID-696-A-11-0006. EDC, Washington, DC. Retrieved June 18, 2018, from http://l3.edc.org/documents/EDC-L3-Endline-Evaluation.pdf.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), 1–27. https://doi.org/10.17763/haer.66.1.g73266758j348t33.

Elmore, R. F. (2016). "Getting to scale…" It seemed like a good idea at the time. *Journal of Educational Change, 17*(4), 529–537. https://doi.org/10.1007/s10833-016-9290-8.

Fleisch, B. (2017). Teachers, the politics of the governed and educational development: Insights from South Africa. In C. Day (Ed.), *The Routledge international handbook of teacher and school development* (pp. 185–193). New York, NY: Routledge.

Fleisch, B., Schöer, V., Roberts, G., & Thornton, A. (2016). System-wide improvement of early-grade mathematics: New evidence from the Gauteng primary language and mathematics strategy. *International Journal of Educational Development, 49,* 157–174. https://doi.org/10.1016/j.ijedudev.2016.02.006.

Freudenberger, E., & Davis, J. (2017). *Tusome external evaluation—Midline report.* Prepared for the Ministry of Education of Kenya, USAID/Kenya, and the UK DFID under contract no. AID-615-TO-16–00012. MSI, A Tetra Tech Company, Washington, DC. Retrieved June 18, 2018 from http://pdf.usaid.gov/pdf_docs/PA00MS6J.pdf.

Friere, P. (1970). Cultural action and conscientization. *Harvard Educational Review, 40*(3), 452–477.

Fullan, M. (2016). The elusive nature of whole system improvement in education. *Journal of Educational Change, 17*(4), 539–544. https://doi.org/10.1007/s10833-016-9289-1.

Fullan, M., & Quinn, J. (2016). *Coherence: The right drivers in action for schools, districts, and systems*. Thousand Oaks, CA: Corwin.

Gove, A., Korda Poole, M., & Piper, B. (2017). Designing for scale: Reflections on rolling out reading improvement in Kenya and Liberia. In P. McCardle, A. Mora, & A. Gove (Eds.), *Progress toward a literate world: Early reading interventions in low-income countries.* Special issue of *New Directions for Child and Adolescent Development, 2017*, issue 155 (pp. 77–95).

J-PAL. (2018). *Evaluation* [Interactive database]. Retrieved June 26, 2018 from https://www.povertyactionlab.org/evaluations.

Jukes, M. C. H., Turner, E. L., Dubeck, M. M., Halliday, K. E., Inyega, H. N., Wolf, S., et al. (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.

Kamau, K. J., Wilson, K. L., & Thinguri, R. (2014). An evaluation of effectiveness of SMASSE program in performance in science and mathematics in primary schools in Kenya. *International Journal of Education and Research, 2*(6), 1–10.

KICD, Ministry of Education, Republic of Kenya. (2017). *Basic education curriculum framework*. Nairobi: KICD.

Kim, Y.-S. G., Boyle, H., Zuilkowski, S. S., & Nakamura, P. (2016). *Landscape report on early grade literacy*. Washington, DC: USAID.

Kisirkoi, F. K. (2012). Effectiveness of Teacher Advisory Centres (TACs) in teacher professional development in Nairobi County. *International Journal of Current Research, 4*(4), 297–302.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics, 91*(3), 437–456. https://doi.org/10.1162/rest.91.3.437.

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.

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.

Moore, A. M., Gove, A., & Tietjen, K. (2017). Great expectations: A framework for assessing and understanding key factors affecting student learning of foundational reading skills. In P. McCardle, A. Mora, & A. Gove (Eds.), *Progress toward a literate world: Early reading interventions in low-income countries.* Special issue of *New Directions for Child and Adolescent Development, 2017*, issue 155 (pp. 13–30). https://doi.org/10.1002/cad.20192.

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. London: McKinsey & Company.

MSI. (2015, revised January 2016). *Tusome revised baseline study.* Prepared for USAID under the Kenya Support Program (KSP), contract no. AID-623-I-12-00001 (task order no. AID-623-TO-14-00004). MSI, Washington, DC. Retrieved June 18, 2018 from http://pdf.usaid.gov/pdf_docs/PA00MR51.pdf.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York: Oxford University Press.

Piper, B. (2009). *Impact study of SMRS using Early Grade Reading Assessment in three provinces in South Africa*. Prepared for USAID/Southern Africa under the Integrated Education Program, contract no. 674-C-00–04–00032–00. RTI International, Research Triangle Park, NC. Online education survey course page. Retrieved June 18, 2018 from https://globalreadingnetwork.net/eddata/education-survey-course. Handout, Reading 3. Retrieved June 18, 2018 from https://globalreadingnetwork.net/sites/default/files/Module-2-Handout_3_FINAL_RSA_SMRS_EGRA_Impact_Study.pdf.

Piper, B., Jepkemei, E., & Kibukho, K. (2015a). 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., Jepkemei, E., Kwayumba, D., & Kibukho, K. (2016a). Kenya's ICT policy in practice: The effectiveness of tablets and e-readers in improving student outcomes. *Forum for International Research in Education, 2*(1), 3–18.

Piper, B., King, S., & Mugenda, A. (2016b, September; revised from 2014). *USAID/Kenya Primary Math and Reading (PRIMR) Initiative: Endline impact evaluation—Revised edition.* Prepared under the USAID Education Data for Decision Making (EdData II) project, task order no. AID-623-M-11–00001 (RTI Task 13), RTI International, Research Triangle Park, NC. Retrieved June

18, 2018 from https://globalreadingnetwork.net/eddata/usaidkenya-primary-math-and-reading-primr-initiative-endline-impact-evaluation-revised.

Piper, B., Kwayumba, D., Oyanga, A., & Jepkemei, E. (2015b). *The Primary Math and Reading (PRIMR) Initiative endline impact evaluation on the DFID Kenya Rural Expansion Programme*. Prepared for DFID Kenya under contract 202657-108. RTI International, Research Triangle Park, NC.

Piper, B., & Mugenda, A. (2012). *The Primary Math and Reading (PRIMR) Initiative: Baseline report*. Prepared under the USAID Education Data for Decision Making (EdData II) project, task order no. AID-623-M-11–00001 (RTI Task 13). RTI International, Research Triangle Park, NC. Retrieved June 18, 2018 from http://pdf.usaid.gov/pdf_docs/pa00hx75.pdf.

Piper, B., Oyanga, A., Mejia, J., & Pouezevara, S. (2017a). Implementing large-scale instructional technology in Kenya: Changing instructional practice and developing accountability in a national education system. *International Journal of Education and Development Using Information and Communication Technology, 13*(3), 57–59.

Piper, B., Ralaingita, W., Akach, L., & King, S. (2016c). Improving procedural and conceptual mathematics outcomes: Evidence from a randomised controlled trial in Kenya. *Journal of Development Effectiveness, 8*(3), 404–422. https://doi.org/10.1080/19439342.2016.1149502.

Piper, B., Schroeder, L., & Trudell, B. (2016d). Oral reading fluency and comprehension in Kenya: Reading acquisition in a multilingual environment. *Journal of Research in Reading, 39*(2), 133–152. https://doi.org/10.1111/1467-9817.12052.

Piper, B., Sitabkhan, Y., Mejia, J., & Betts, K. (2017b). *How scripted is too scripted? Mixed-methods analysis of RTI's teachers' guides in developing countries.* Report on internal research and development study, FY 2017, RTI International, Research Triangle Park, NC.

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., Dubeck, M., Jepkemei, E., & King, S. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Coaching, teacher professional development, improved books, and teachers' guides. *World Development, 106,* 324–336. https://doi.org/10.1016/j.worlddev.2018.01.018.

Piper, B., Zuilkowski, S. S., Kwayumba, D., & Strigel, C. (2016e). 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. (2016f). 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.

Pritchett, L. (2013). *The rebirth of education: Schooling ain't learning*. Washington, DC: Center for Global Development.

Pritchett, L. (2015). *Creating education systems coherent for learning outcomes: Making the transition from schooling to learning* RISE [Research on Improving Systems of Education] working paper 15/005, Center for Global Development, Washington, DC. Retrieved June 18, 2018 from https://www.riseprogramme.org/sites/www.riseprogramme.org/files/inline-files/RISE_WP-005_Pritchett_1.pdf.

Republic of Kenya. (2010). *The constitution of Kenya, 2010*. Nairobi: National Council for Law Reporting.

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), RTI International, Research Triangle Park, NC. Retrieved June 18, 2018 from http://pdf.usaid.gov/pdf_docs/PNADS440.pdf.

Rincón-Gallardo, S. (2016). Large scale pedagogical transformation as widespread cultural change in Mexican public schools. *Journal of Educational Change, 17*(4), 411–436. https://doi.org/10.1007/s10833-016-9286-4.

Rincón-Gallardo, S., & Fleisch, B. (2016). Bringing effective instructional practice to scale: An introduction. *Journal of Educational Change, 17*(4), 379–383. https://doi.org/10.1007/s10833-016-9288-2.

Robinson, J. P., Winthrop, R., & McGivney, E. (2016). *Millions learning: Scaling up quality education in developing countries*. Washington, DC: The Brookings Institution.

RTI International. (2014). *Early grade mathematics (EGMA) toolkit*. Research Triangle Park, NC: RTI. Retrieved June 18, 2018, from https://shared.rti.org/content/early-grade-mathematics-assessment-egma-toolkit.

RTI International. (2016). *The Early Grade Reading Assessment toolkit, second edition.* Prepared for USAID under the Education Data for Decision Making (EdData II) project, task order no. AID-OAA-12-BC-00003 (RTI Task 20), USAID, Washington, DC. Retrieved June 18, 2018 from http://pdf.usaid.gov/pdf_docs/PA00M4TN.pdf.

Snilstveit, B., Stevenson, J., Menon, R., Philips, D., Gallagher, E., Geleen, M., et al. (2016). *The impact of education programmes on learning and school participation in low- and middle-income countries.* Systematic review summary 7, 3ie, London. Retrieved June 18, 2018 from http://www.3ieimpact.org/media/filer_public/2016/09/20/srs7-education-report.pdf.

Taylor, S., & von Fintel, M. (2016). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach. *Economics of Education Review, 50,* 75–89. https://doi.org/10.1016/j.econedurev.2016.01.003.

The DDD Manifesto Community. (2017). *Doing development differently: The manifesto.* Retrieved October 13, 2017, from http://doingdevelopmentdifferently.com/the-ddd-manifesto.

Uwezo. (2016). *Are our children learning (2016)? Uwezo Kenya sixth learning assessment report, December 2016.* Twaweza East Africa, Nairobi. Retrieved June 18, 2018 from http://www.twaweza.org/uploads/files/UwezoKenya2015ALAReport-FINAL-EN-web.pdf.

Villegas-Reimers, E. (1998). *The preparation of teachers in Latin America: Challenges and trends*. Washington, DC: Human Development Department, World Bank, Latin America and the Caribbean Regional Office.

World Bank. (2017a). *World Bank indicators: GDP per capita (current US$)* [Interactive database]. Retrieved October 13, 2017, from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD.

World Bank. (2017b). *World Bank indicators: Population, total* [Interactive database]. Retrieved October 13, 2017, from https://data.worldbank.org/indicator/SP.POP.TOTL.

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

Affiliations

Benjamin Piper[1] · Joseph Destefano[2] · Esther M. Kinyanjui[3] · Salome Ong'ele[1]

✉ Benjamin Piper
bpiper@rti.org

Joseph Destefano
jdestefano@rti.org

Esther M. Kinyanjui
maigaek@gmail.com

Salome Ong'ele
bpiper@rti.org

[1] RTI International, Nairobi Regional Office, The Westwood, 5th Floor, Vale Close, Westlands, Village Market, P.O. Box 1181, Nairobi 00621, Kenya

[2] RTI International, 3040 Cornwallis Rd, P.O. Box 12194, Research Triangle Park, NC 27709-2194, USA

[3] Ministry of Public Service Youth and Gender Affairs, State Department of Public Service and Youth, P.O. Box 12881, Nakuru, Kenya