Journal of Development Economics 148 (2021) 102539

Contents lists available at ScienceDirect

Journal of Development Economics

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

Regular Article

How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia☆

Alex Eble [a,*], Chris Frost [b], Alpha Camara [c], Baboucarr Bouy [c], Momodou Bah [c], Maitri Sivaraman [c], Pei-Tseng Jenny Hsieh [d], Chitra Jayanty [1], Tony Brady [e], Piotr Gawron [e], Stijn Vansteelandt [b,f], Peter Boone [g], Diana Elbourne [b]

[a] Teachers College, Columbia University, USA
[b] London School of Hygiene and Tropical Medicine, UK
[c] Effective Intervention, The Gambia
[d] Centre for Comparative and International Education, Department of Education, University of Oxford, UK
[e] Sealed Envelope, Ltd, UK
[f] Department of Applied Mathematics, Computer Science and Statistics at Ghent University, Belgium
[g] Effective Intervention, UK

A B S T R A C T

Despite large schooling and learning gains in many developing countries, children in highly deprived areas are often unlikely to achieve even basic literacy and numeracy. We study how much of this problem can be resolved using a multi-pronged intervention combining three interventions known to be separately effective. We conducted a cluster-randomized trial in The Gambia evaluating a literacy and numeracy intervention designed for primary-aged children in remote parts of poor countries. The intervention combines para teachers delivering after-school supplementary classes, scripted lesson plans, and frequent monitoring focusing on improving teacher practice (coaching). A similar intervention previously demonstrated large learning gains in rural India. After three academic years, Gambian children allocated to the intervention scored 46 percentage points (3.2 SD) better on a combined literacy and numeracy test than control children. Our results demonstrate that, in this type of area, aggressive interventions can yield far greater learning gains than previously shown.

1. Introduction

While children in developing countries are much more likely than before to be in school, in dozens of countries, they are still highly unlikely to acquire the skills expected of them at each grade level (Bold et al., 2017; Pritchett, 2013). Learning levels are often even lower in rural areas of these countries (Chaudhury et al., 2006; Glewwe, 2002). Reviews show that hundreds of studies have evaluated a wide range of different interventions attempting to raise learning levels in these contexts (Evans and Popova, 2016; Ganimian and Murnane, 2016; McEwan, 2015). Among the many studies yielding positive results, the majority find modest test score or ability changes, usually in the range of 0.1–0.5 test score standard deviations, or SDs (Kremer et al., 2013). This suggests that, to date, we know very little about how to generate the type of large gains necessary to close the learning gap between developing and developed countries (Glewwe and Muralidharan, 2016).

In this paper, we report results from a randomized controlled trial asking the following research question: if we deliver a multi-pronged, well-resourced intervention to children in a highly deprived setting, how much of this learning gap can we close? The intervention we study

☆ We would like to acknowledge the contributions of our team in The Gambia, particularly Lenin Balan, Lamin Janneh, Alieu Jibba, Jenieri Sagnia, and Abhishek Thakur. This study was approved by the Institutional Review Board of the London School of Hygiene and Tropical Medicine, protocol number 8767. Funding for this study was provided by Effective Intervention, a UK-registered charity.

* Corresponding author.
E-mail addresses: eble@tc.columbia.edu (A. Eble), chris.frost@lshtm.ac.uk (C. Frost), a.camara@effint.net (A. Camara), bbouy@effint.net (B. Bouy), newsbah@yahoo.com (M. Bah), maitri.effectiveintervention@gmail.com (M. Sivaraman), ptjhsieh@gmail.com (P.-T.J. Hsieh), chitra.jayanty@gmail.com (C. Jayanty), support@sealedenvelope.com (T. Brady), support@sealedenvelope.com (P. Gawron), stijn.vansteelandt@lshtm.ac.uk (S. Vansteelandt), pb@effint.org (P. Boone), Diana.Elbourne@lshtm.ac.uk (D. Elbourne).
[1] Independent Consultant

https://doi.org/10.1016/j.jdeveco.2020.102539
Received 7 November 2019; Received in revised form 24 June 2020; Accepted 2 August 2020
Available online 19 September 2020

combines three well-known levers for improving learning: i) the use of para teachers, instead of civil servants or volunteers, to deliver after school lessons (Banerjee et al., 2007; Duflo et al., 2015; Lakshminarayana et al., 2013; Muralidharan and Sundararaman, 2013); ii) an improved, scripted curriculum targeted at students' current learning levels (Banerjee et al., 2007, 2017; Lakshminarayana et al., 2013; Piper et al., 2014); and iii) extensive monitoring of these teachers, the aim of which was to provide regular feedback on teaching methods and practice (Kraft et al., 2018; Muralidharan et al., 2017). This intervention was originally designed by The Naandi Foundation, an Indian non-governmental organization, and raised learning levels among primary-aged children by 0.75 SD in rural Telangana, India after two years of implementation (Lakshminarayana et al., 2013). In this study, we partly attempt to learn whether the gains in learning achieved by this intervention in India can, with appropriate adaptation, be realized in a new, more challenging context: rural areas of The Gambia.

We find that this intervention generates extremely large learning gains in rural Gambia. Children receiving the intervention scored 46 percentage points (3.2 SD) better on a composite literacy and numeracy test than children in the control villages. This comprises a large change across the distribution of scores: for example, a child at the 10th percentile of the distribution of scores for intervention children would be at approximately the 80th percentile of the control group score distribution. Furthermore, these learning gains are meaningful in absolute, as well as relative, terms. For comparison, rural Gambian children who received the intervention performed, on average, as well or better in all comparable English reading subtasks than children in a nationally representative assessment of these skills among third grade students in the Philippines, a country with a per-capita GDP several times greater than that of The Gambia. Their scores also compare favorably to scores of similarly-aged children from other developing countries, such as Uganda, Egypt, Morocco, and Iraq.[2]

We worked in 169 small villages in the two central regions of The Gambia. We began with a census of 6–8 year-olds in these villages whose caregivers planned to enroll them in primary school the following year, and we followed these children for the next three years. Villages were randomized in clusters, with half receiving the intervention and half not. The intervention ran from early January 2016 to early May 2018. Delivery of the intervention constituted recruiting, training, and deploying para teachers to deliver scripted, supplementary lessons in mathematics and reading for 12 h per week in addition to their normal schooling. The intervention was adapted from that in Lakshminarayana et al. (2013) to the Gambian setting, advancing with children through the first three years of the official primary school curriculum of The Gambia. At baseline, schools in these regions of The Gambia operated with traditional pedagogy, a typical civil servant structure for recruiting and retaining teachers, and normal resource constraints. The intervention addresses each dimension – changing pedagogy to scripted lessons, changing the incentives facing teachers, and substantially increasing the resources spent on staff, materials, and staff development.

The pre-specified primary outcome of the trial was the arithmetic mean of the child's score on Early Grade Reading and Mathematics Assessment-style tests (also known as EGRA and EGMA tests: Dubeck and Gove, 2015; Platas et al., 2014) administered at the end of the trial. Attrition from enrollment to the endline test was less than 14%. The differences between control and intervention child scores are similar for the mathematics and reading tests. Each test comprises subtasks of varying difficulty. Intervention children had substantially higher scores than control children on subtasks that are relatively simple (e.g., letter recognition and number recognition), moderately difficult (e.g., familiar word reading and single-digit addition), and more difficult (e.g., reading comprehension and two- or three-digit subtraction with borrowing).

We argue that the dramatic increase in learning we observe in the intervention group is likely due to two main factors. The first is that the intervention combines several tools – para teachers delivering after school lessons; an improved, scripted, and targeted curriculum; and extensive monitoring of these para teachers with an emphasis on pedagogical improvement – known to be effective in isolation (e.g., Banerjee et al., 2017; Kraft et al., 2018; Muralidharan et al., 2017; Muralidharan and Sundararaman, 2013; Piper et al., 2018a, 2018b).[3] The second factor is the very low learning levels of children in rural parts of The Gambia, even among those enrolled in school (Blimpo et al., 2011). For example, in our control villages, more than two thirds of children could not read a single short (two-to seven-letter) word taken from the second-grade curriculum, and half could not successfully complete even one single-digit addition problem at endline.[4] This makes exceptionally large relative improvements possible.[5]

Our results suggest that the upper bound for the magnitude of intervention-driven learning gains in similar settings is much higher than previously observed (McEwan, 2015). This study, in conjunction with Lakshminarayana et al. (2013), shows that there exists a demonstrated way to reach the learning gains that many have called for in particularly disadvantaged areas (Pritchett, 2013; Glewwe and Muralidharan, 2016). This echoes other work on the efficacy of "bundled" interventions combining multiple separate interventions. While substantially more expensive than typical interventions, these often show large impacts on otherwise difficult to move outcomes (Banerjee et al., 2017; Brudevold-Newman et al., 2017; Bandiera et al., 2020).

The main policy question this raises is about implementation: should governments and donors who wish to reap such gains attempt to operationalize this within the government system, or contract it out? There are two core challenges. One is cost: in its current form, this after-school intervention costs more than the average government spend per pupil. The other is the set of challenges involved in scaling up effective NGO interventions, as described in Banerjee et al. (2017) and Bold et al. (2018). As a result, we see our results primarily as indication of a path forward for addressing this type of problem, rather than an immediate policy recommendation.[6]

We argue that these findings also provide evidence on the relative importance of supply of and demand for education in explaining low levels of learning in such settings. The low outcomes we observe could reflect very poor schools, or a lack of demand for education, or both. We obtain large learning gains with an experimentally induced change in supply, showing that these poor outcomes can be improved without inducing a change in demand. This suggests that low learning levels in these areas are at least primarily supply-driven.

The rest of the paper proceeds as follows: in Section 2 we describe the

[2] EGRA data for these countries come from https://earlygradereadingbarometer.org/overview, accessed October 16, 2019. GDP data come from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD, accessed on the same date.

[3] This combination of multiple known best practices, offered through a contracted provider, parallels solutions from studies of the provision of healthcare in similarly remote and disadvantaged areas (Salehi et al., 2018) and of "graduation" programs helping individuals out of extreme poverty (Banerjee et al., 2015).

[4] These learning levels are substantially worse than those observed in similar assessments of children's learning in other developing countries, such as India and Tanzania (Pratham, 2010; Rajani, 2010) and similar to what some of us have observed in recently completed work in rural Guinea Bissau (Fazzio et al., 2020). This pattern also appears in regular national assessments of child learning using EGRA- and EGMA-style tests administered by the Gambian government and by third parties (e.g., Sprenger-Charolles, 2008).

[5] Burde and Linden (2013) find that a supply-side intervention yields learning gains of 0.4–0.65 SD in rural Afghanistan, a setting with comparably low baseline literacy and numeracy.

[6] We also contribute to the growing set of studies looking at the replication, scalability, and generalizability of results from RCTs, particularly those that try to raise learning levels in developing countries (Banerjee et al., 2017; Bold et al., 2018; Lucas et al., 2014).

intervention we study and our setting. In Section 3 we provide an overview of our research design. In Section 4 we present our main results. In Section 5, we discuss the reliability of and potential mechanisms behind our results. Section 6 concludes. In the Appendices, we include a further description of the intervention and its motivation, the full statistical analysis tables according to our pre-specified statistical analysis plan, a sample size calculation, the final test papers we used to assess child reading and mathematics ability, and a series of other robustness checks and exploratory analyses.

2. Intervention and setting

In this section we briefly describe the setting we work in, the logic of the intervention we study, and its implementation.

2.1. Setting

The Gambia is a small country in West Africa, roughly 475 km long and 25–50 km wide. Its main exports are tourism and agricultural products. It was formerly a British colony and home to a large part of Britain's slave trade. Much of The Gambia is rural and hard to reach. Its per-capita gross domestic product was estimated to be US $483 in 2017 by the World Bank (market rate, not purchasing power parity), placing it among the 10 poorest countries in the world according to that ranking (The World Bank, 2018a).

The education system in The Gambia comprises basic education – six years of primary schooling (lower basic) and three years of middle school (upper basic) – followed by three years of high school or vocational training, and then four years of university. Its gross primary school enrollment rate was 100 percent in 2017, slightly higher than that for sub-Saharan Africa overall (97 percent).[7] Population-level data on the average years of education completed are sparse, particularly for older generations, but in a census conducted in 2013, literacy rates for individuals aged 15 or older were estimated to be approximately 42% overall, with higher levels for males (51%) than females (34%) (The World Bank, 2018b). Since 2007, The Gambia has run regular assessments of the literacy of its children using EGRA-style tests, with the addition of EGMA-style tests beginning in 2013. The first assessment yielded startling results – country-wide, 46% of third grade students could not read a single word in a sentence, and fewer than 20% were at the level expected of them by the national curriculum (Sprenger-Charolles, 2008). In the regions in which our study takes place, these levels are much lower.[8]

The Gambia has six administrative regions – Banjul (the capital), West Coast, Lower River, North Bank, Central River, and Upper River. The Gambia's Ministry of Basic and Secondary Education (MoBSE) recommended we work in the Lower River and North Bank regions, located in the center of the country. The reasoning behind this recommendation was that these regions were needier than the Western regions (Banjul and West Coast) and had fewer ongoing NGO interventions than the Eastern Regions (Central River and Upper River).

2.2. Intervention design

Our intervention originated with the intervention evaluated in Lakshminarayana et al. (2013). That study ran a cluster-randomized trial evaluating a similar para teacher intervention. That intervention was designed by The Naandi Foundation, which had been implementing it in multiple Indian states for several years prior to the start of the trial in 2008. The trial reported in Lakshminarayana et al. (2013) took place in 214 villages in rural Telangana (then Andhra Pradesh), India, over a period of two academic years. The study found that the intervention yielded a 0.75 SD increase in reading and mathematics test scores among children in intervention villages, relative to those in control villages.

In The Gambia, our main goal was to see if this model could be similarly or more effective in a context where one of its core components – using previously untrained people to deliver basic literacy and numeracy skills – might be an even better fit. This potential improvement in fit comes from the premise that teaching at lower learning levels is easier and requires lower baseline learning than teaching at higher levels, such as those taught to the fourth and fifth grade students in Lakshminarayana et al. (2013). Furthermore, English is an official national language of The Gambia and its main language for instruction, making it easier to maintain quality control in adapting the materials from Lakshminarayana et al. (2013) to the local context.

In a separate project, some of us took this model to Guinea Bissau to try to raise learning levels in remote rural areas there (Fazzio et al., 2020). In that context, there is essentially no reliable state implementation of education. As a result, the core research question of that study diverged, and instead aimed to see whether the main thrust of the model could be delivered via a dramatically different method. The intervention studied in Fazzio et al. (2020) replaced the early grades of government primary schools with privately operated schools which use only two of the core insights – scripted lessons and frequent monitoring. Because of the extremely low levels of human capital in Guinea Bissau, implementing the "para teacher" component – hiring literate and numerate individuals locally and training them to serve as teachers – was impossible, and trained teachers had to be hired instead.

2.3. Implementation

In trial villages randomly selected to receive the intervention, implementation proceeded according to the following steps: first, we held a meeting in each community, announcing the intervention and asking all community members for their support. Second, with the community, we attempted to identify an adult from the village with at least a 12th grade education, to serve as a para teacher. In the absence of such an individual, we relaxed either the locality or education requirement (or both) and found the most qualified individual who met a set of minimum qualifications, passed a proficiency test, and who was willing to reside in the village, with a preference for those from nearby communities.[9] We paid these individuals a post-tax salary of 3550 Gambian Dalasis (GMD) per month (US $81.12), roughly 14 percent more than the

[7] Data taken from https://data.worldbank.org/indicator/se.prm.enrr on May 22, 2019.

[8] The Gambian government has recognized the need for such an intervention since 2008 (Sprenger-Charolles, 2008). Furthermore, after initial discussions, the government stated that, were the intervention found to be effective, it would be interested in the potential for subsequently implementing the intervention itself.

[9] We sought to hire individuals primarily from the village in which they served, though in approximately half of our villages, we were unable to find someone with sufficient education to do so, and so we had to recruit from nearby communities. The goal here was to exploit the "informational and motivational advantages" that come with hiring local individuals - also known as local delivery agents – to administer services (Bandiera et al., 2018). We required all teachers to live in the village in which they served, improving the ability of children and parents to address teacher attendance issues directly, i.e., by going to the teacher's home to find them should they be absent.

government teachers received (3085 GMD per month, or US $70.55).,[10][11] Due to other benefits from the government and greater monetary and in-kind transfers from parents to government teachers, overall compensation was roughly equivalent for the two groups. Para teacher salaries were higher in our project than in other para teacher work (Muralidharan and Sundararaman, 2013) for three reasons. First, the market clearing wage for individuals qualified to do our work was higher in our project area because their outside option was to migrate to the city, where they could earn nearly twice the monthly wage we paid them. Second, our demands on their time were greater than those on government teachers. Finally, in most villages, we found few individuals who were qualified to do the work.

To emphasize their role in the community, these para teachers were hired under the official title of "community educator", or "CE." At the start of the intervention, we provided an initial two-month pre-service training for the recruited CEs in pedagogical content knowledge related to our curriculum. This curriculum was based on the official Gambian national curriculum for these grades, but incorporated scripted daily lesson plans and a series of activities for each lesson, using the core tools from Lakshminarayana et al. (2013). It was also designed to be easily implementable by our team of CEs who, when hired, had no teaching experience.[12] Recently, interventions which include scripted lessons have been shown to work in numerous contexts (Piper et al., 2018a, 2018b; Romero et al., 2020). These scripted lessons aim to provide scaffolding to enhance the quality of instruction in early grade learning, particularly where teacher training is suboptimal and there is limited scope for supervision.[13]

After their training, CEs returned to their villages to commence the intervention.[14] Each CE administered 12 hours of after-school lessons per week using our daily scripted lesson plans. These lessons took place either in the local school or, in the absence of a nearby school, a structure which we supplied with mats for sitting and a chalkboard. In these cases, the community either furnished a suitable place to hold the lessons – such as a local madrassa or meeting hall – or constructed a one-room structure in which to hold the lessons. We monitored the CEs throughout the course of the intervention and provided regular training as the curriculum progressed. In Appendix A, we provide more information on the setup, implementation, and motivating principles of the intervention.

The intervention targeted all children in the village who were eligible for participation in the study. To be eligible, the child had to be aged between six and eight years old at enumeration and intended to enroll in the first grade, for the first time, in the following school year.[15] We describe eligibility criteria more fully in the next section. We describe levels of enrollment in school over the course of the trial, by randomization assignment (intervention vs. control), in Section 4. In Appendix B, we provide a series of tables, as specified in our Statistical Analysis Plan,[16] which give greater detail about the socio-economic and demographic characteristics of participants in our study.

Our main interaction with the government education system in the intervention arm was to request the use of government school facilities to hold lessons after-hours and store our learning materials. In almost all cases, this support was generously provided. We chose not to engage in formal coordination with government teachers about our students or our intervention. We made this decision with the hope of minimizing disruption of normal school activities and to avoid adding to the workload of government teachers. In the classrooms of government schools where intervention children attended school during the day, most classrooms were a mix of participants and non-participants. This occurred for two reasons. First, our eligibility criteria excluded older students who were entering the first grade at age 9 or later. Second, some families delayed their children's enrollment in school, leading to a staggered start of government school for intervention children. For example, at the end of the first year, less than 50 percent of intervention children were actually enrolled in the first grade, and at the end of the three years, only 40 percent of intervention children had progressed to the third grade. In Section 5.2, we describe how the government schools may have conditioned the effect of the intervention.

3. Research design

This section describes our research design. We published a study protocol prior to executing the study, which describes the study design in further detail and specifies our primary and secondary outcomes and analysis methods (Boone et al., 2015).

3.1. Eligibility and enrollment of villages and children

We began with a list of the 323 villages in the Lower River and North Bank regions with between 15 and 300 households according to the 2013 Gambian national census. In each village on this list, we conducted a census of all dwellings in order to enumerate the number of children born between January 1, 2006 and December 31, 2010 currently resident in the village, counting only those children whom we could meet face-to-face.[17] For a village to be eligible for inclusion in the study, we required that it have at least 10 eligible children. For a child to be eligible, they had to be born between September 1, 2007 and August 31, 2009, not yet enrolled in the first grade, and their caregiver must have expressed

[10] Education spending in The Gambia is similar to that of its neighbors. The Gambia spends roughly 8.5% of GDP per capita on primary education annually. Two of its nearest geographic neighbors, Guinea and Senegal, spend 6.8% and 10.9%, respectively, according to World Bank Statistics (https://data.worldbank.org/indicator/SE.XPD.PRIM.PC.ZS?view=map, accessed April 5, 2020). Guinea is closer to The Gambia in terms of GDP per capita than Senegal, and pays its teachers roughly the same.

[11] All GMD to USD conversions are made using the average exchange rate over the period 2015–2018 taken from https://www.exchangerates.org.uk/on May 15, 2019.

[12] The main pre-service training happened at the beginning of the study. This comprised 30 days of training, for roughly 6 hours a day, or 180 hours of training. As the program progressed, we occasionally brought on new CEs on an as-needed basis to fill vacancies left by departing or underperforming CEs. These CEs received a shorter, more intense 10-day training (80 hours total) before being sent to work, and received additional supervision on an ad hoc basis as deemed necessary by supervisors to complete their onboarding.

[13] This intuition is also borne out in Muralidharan et al. (2019), who show that an after-school program providing adaptive learning software via tablet computers with teacher input can be highly effective at raising student learning levels. They also find that the largest gains are among academically weaker students, the students in that study who are perhaps most similar to the students in our context.

[14] Prior to being hired, the CEs were most frequently farmers, workers in shops or shopkeepers, or doing other small scale business in their villages. They were encouraged to keep their other employment as long as they dedicated 4 hours to teaching from Monday to Saturday (comprising 2.5 hours in the class teaching children, including startup and wind-down, and 1.5 hours for preparing their lessons, correcting exercise books and giving extra support to children as needed). They were also required to attend periodic training and review meetings. To the best of our knowledge, all CEs maintained some version of their main source of income after joining our project.

[15] We were only able to capture parent intention to enroll their child. As a result, our sample includes many children who did not ultimately enroll in school the following school year and some children (a subset of this group) for whom, this intervention is not "supplementary," but rather the only formal education that they received. We present relevant summary statistics on enrolment over the course of the study in Table 8.

[16] In addition to our published study protocol (Boone et al., 2015), we also wrote out a more detailed statistical analysis plan prior to analysis of the data. The full results according to this analysis plan are given in the Appendix. The accompanying text of the analysis plan is available from the authors on request.

[17] Birthdate data were confirmed with birth and health records in all cases where possible.

intention to enroll them in the first grade in the coming academic year.[18]

Among villages that were eligible, we drew a circle with a radius of 2.5 km around each village to serve as a buffer area. We then generated clusters of villages, the unit of randomization, from contiguous groups of village buffer zones. If there were three or more villages in a given cluster, we removed one or more of these villages from the trial to generate the maximum number of clusters such that there were at least 5 km between the GPS coordinates of any village in the cluster to those of all villages in all other clusters. This left us with 169 villages grouped into 111 clusters. We enrolled all eligible children in these 169 villages into the trial, obtaining consent from village chiefs and each child's primary caregiver.

We conducted our randomization by these clusters of villages to avoid the risk of spillover, e.g., children in control villages being able to walk into intervention villages and avail themselves of intervention classes there. We used a random number generator, stratifying on two criteria: whether a cluster was in the Lower River or North Bank region, and whether the cluster was above or below the median distance to the main road in its region. In Appendix C, we give the sample size calculation we used to ensure that we would have adequate statistical power.

3.2. Data collection

We collected information from families, villages, and schools at baseline as part of our census of potentially eligible villages. After finalizing which villages would be included in the trial, we conducted a second survey of all eligible children in these villages to enroll them in the trial, as we could not enroll children prior to determination of their village's eligibility. At this point, we also collected additional data on the child and their caregiver. We collected end-of-school-year surveys at the end of the first and second academic years (i.e., May-June 2016 and 2017) to measure child migration, enrollment in school, and attendance. Between February and May 2018, we conducted an endline survey with caregivers of children to measure attitudes, time use, and other child- and family-level variables.

Finally, between May and June 2018, we conducted EGRA and EGMA-style tests among all of the randomized children who we could locate in the villages and who were willing to take the tests (regardless of whether or not they were enrolled in school). EGRA (Early Grade Reading Assessments) and EGMA (Early Grade Mathematics Assessments) are reading and mathematics tests, respectively. They were initially designed by RTI International and are intended to test a set of basic skills related to each subject (Dubeck and Gove, 2015; Platas et al., 2014).[19] These tests are oral assessments conducted one-on-one with the child, avoiding potential floor effects associated with requiring the child to complete a paper-based assessment.

In Table 1, we briefly describe the skills tested in each of the two tests. The full test papers are given in Appendix D. One of the authors (Hsieh) was contracted to design our tests to be consistent with prior EGRA and EGMA-style assessments in The Gambia and to ensure consistency of training and implementation by assessors. This design, training, and implementation followed RTI's guidelines for creating reliable and accurate EGRA and EGMA-style assessments. In Appendix E, we discuss how we addressed the threat of floor and ceiling effects influencing our results and their interpretation.

Table 1
Description of EGRA and EGMA test content.

EGRA	EGMA
1: Read a letter's sound (e.g., "eh" for e)*	1: Read a number (e.g., 1, 5, 22)
2: Differentiate sounds (e.g., which word starts with a different sound: book, dog, or boy?)*	2: Choose the larger number (e.g., 7 or 5?)
3: Read a made-up word (e.g., ked)	3: Complete a sequence (e.g., 2 4 6 _)
4: Read a familiar word (e.g., but)	4a: Simple addition (e.g., 3 + 2) 4b: Two- and three-digit addition (e.g., 38 + 26)
5a: Read a short passage	5a: Simple subtraction (e.g., 5 – 3)
5b: Answer questions on the passage's content	5b: Two- and three-digit subtraction (e.g., 59 – 37)
6: Listen to a different short passage, answer questions on the passage's content	6: Solve a simple word problem read aloud

Note: this table provides descriptions of the different types of questions asked on the reading (EGRA) and mathematics (EGMA) tests, respectively. Later in the text, these are referred to as "tasks" or "subtasks", by the number given in this table. * The Gambian Ministry of Basic and Secondary Education began using a phonics instruction system called "Jolly Phonics" in 2007. This system trains students to recognize letter sounds in precisely the manner tested in subtasks 1 and 2 on the EGRA test, meaning that these subtasks, while potentially not aligned with national curricula in some contexts, are well-aligned with the Gambian national curriculum.

3.3. Pre-specified primary and secondary endpoints and analysis method

Our published study protocol (Boone et al., 2015) contains key information on our analysis plan, including primary and secondary outcomes and method of analysis, specified prior to conducting the trial. The primary investigators agreed upon and signed off on a detailed statistical analysis plan prior to the start of statistical analysis.[20] Our primary outcome is a composite test score, calculated as the arithmetic mean of a child's scores (scaled 0–100) on consecutively administered reading (EGRA) and mathematics (EGMA) tests.[21]

As described in Section 2, these tests are aligned with the national curriculum and assessment standards. MoBSE has used EGRA-style tests since 2007 to assess child learning and government teacher performance, with the addition of EGMA-style tests in 2013. Over the study period there was also a nationwide initiative providing continuous training to government teachers on how to teach reading. Our choice of primary endpoint is therefore aligned with the government system's standards and goals for its students and teachers.[22] Participants were aware of their randomization status; nonetheless, the objective nature of the primary outcome data makes this knowledge unlikely to bias our results (Wood et al., 2008).

We use a linear regression model to compare child-specific composite endline test scores between intervention and control groups. In our model, we control only for the two stratification factors included in the randomization (region: lower river or north bank; distance to main road in region: above or below median distance) by including these as dummy

[18] We started with the wider range – 2006 to 2010 – to avoid parental misreporting of a child's birth date in order to satisfy the September 1, 2007 to August 31, 2009 eligibility rule. In the list of children we collected, we see no bunching around either the earlier or later eligibility cutoff.

[19] They are intended to be adapted to a local context for each administration. Test questions differ from test to test to conform to local needs and standards while following standardized EGRA/EGMA protocol design guidelines.

[20] As mentioned earlier, the full tables appear in Appendix B and the accompanying text is available by request.

[21] We chose this composite score because we needed to select one outcome to serve as the primary outcome of our study; we acknowledge that this is a departure from the intended and conventional use of EGRA and EGMA tests. To align with more conventional use, we present subtask scores in Figs. 3.1 and 3.2. Other commonly presented scores – e.g., fluency scores and zero scores – are available on request.

[22] In cases such as India, where the curriculum advances rapidly, there is reason to believe that a study such as ours with one specific endpoint might distort teacher effort (Muralidharan and Sundararaman, 2011).

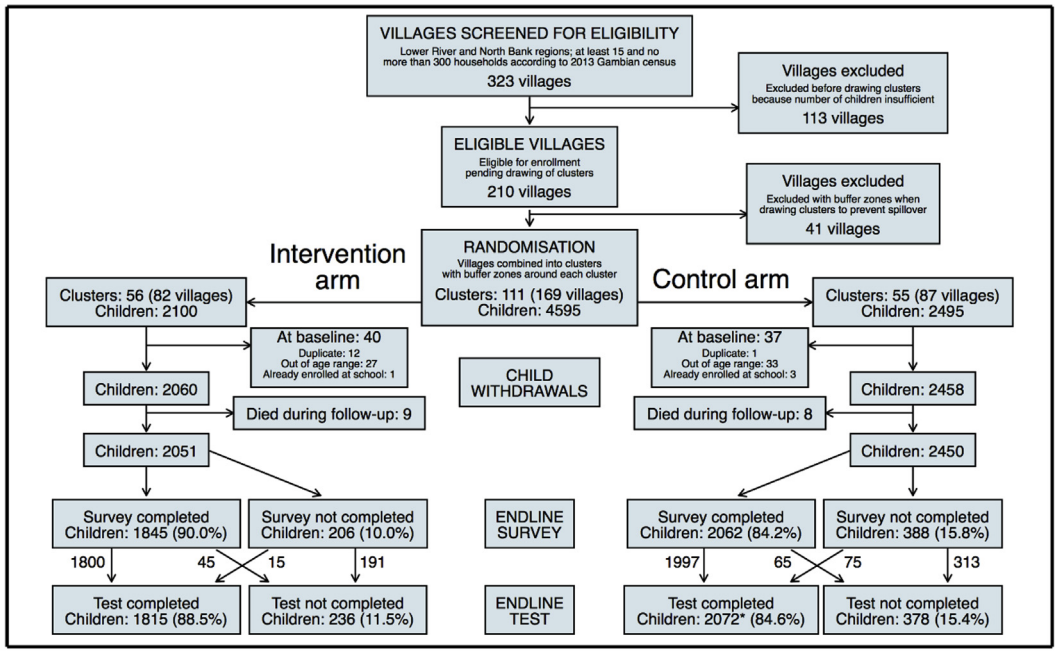


Fig. 1. CONSORT-style flowchart of participants through trial.

variables. Here, and in other analyses, we perform hypothesis tests and calculate 95% confidence intervals using robust standard errors, allowing for correlated responses by cluster of villages. For the primary outcome, we divide the adjusted difference in means by the standard deviation (SD) of the test score in the control group to give a standardized difference in SD terms.[23] We present this standardized difference together with a nonparametric bootstrap confidence interval.[24] We also compute Lee bounds (Lee 2009) to explore impact of differential attrition between the two groups.[25] We also perform a number of exploratory mediation analyses to seek to understand the extent to which school enrolment, school grade in the year of testing, school attendance and adherence to the intervention might have influenced the effect of the intervention.

In our prespecified secondary analyses of the composite test scores, we extend the linear regression model described above to investigate interactions by ethnic group, gender, wealth, caregiver education, and two measures of geographic location. We also investigate heterogeneity in effect size based on whether there was a school in the village; this is an exploratory (not-prespecified) analysis. The test performance data by subtask, which comprise the primary outcome, are presented in bar charts. Our pre-specified secondary outcomes include school attendance, enrollment, performance on nationally administered exams,[26] and school-related time use of parents and children.[27] For dichotomous secondary outcomes – such as whether the child was enrolled in school – we present odds ratios with 95% confidence intervals obtained from a logistic regression model fitted within a generalized estimating equation (GEE) framework with a binary outcome, a logit link, and a 'working' assumption of independence, with robust standard errors to take account of clustering. Our logistic regression model includes the stratification factors as well as the randomization group as a binary dummy variable. This gives estimates of the effect of the intervention expressed as odds ratios that are conditional on the stratification factors.[28] By re-fitting the models and switching to an identity link, rather than a logit link, we also express results as differences in proportions.

We conduct all primary analysis on an intent-to-treat basis, including children in the group that their village was randomized to regardless of attendance at classes or school. In Appendix G, we describe and report pre-specified, secondary, "per-protocol" analyses of the primary outcome in villages and children whose class schedule (villages) and child attendance (villages and children) met pre-specified attendance thresholds.

3.4. Randomization and balance

In Fig. 1, we present a CONSORT-style diagram (Campbell et al., 2012) showing the numbers of villages, clusters, and children at various stages of the trial. This diagram shows how villages and children flowed through the trial from consideration for eligibility to the endline test. In Table 2, we show the baseline characteristics of the clusters in our trial, the level at which we randomized. Fifty-six clusters were randomized to the intervention group and fifty-five clusters randomized to the control group; as a result, clusters receiving the intervention have a somewhat smaller mean number of eligible children per cluster than those in the

[23] Here we use the total SD for the control group estimated by fitting a linear mixed model that allowed for between and within cluster variability.

[24] We use a bootstrapped confidence interval here because the distribution of a standardized difference needs to take account of sampling variability in both the estimated difference in means and the estimated standard deviation. This can be achieved by bootstrapping. Here and elsewhere, bootstrap confidence intervals are bias corrected and accelerated, computed from 2000 bootstrap resamples of the pre-defined clusters of villages stratified by randomized group. Elsewhere we use bootstrap confidence intervals when variables are highly skewed and likely to be far from normally distributed even when aggregated at cluster level.

[25] We use the bootstrap to obtain confidence intervals on these bounds.

[26] After we finalized our analysis plan, a government policy reduced the frequency of nationally administered exams. As a result, we are unable to conduct this analysis because children in our study were never administered these exams.

[27] Other pre-specified analysis of parents' spending on education, and spillover learning to siblings and family members are presented in appendix F.

[28] As an illustration of how to interpret odds ratios, suppose that the proportion of children in a particular stratum who have the outcome is 60% in the control arm, but 75% in the intervention arm. The odds of having the outcome in the control arm is 1.5 (60/40) and in the intervention arm it is 3 (75/25). The odds ratio is therefore 2 (3/1.5).

Table 2
Baseline characteristics of clusters.

Variable	Intervention	Control
Number of clusters	56	55
Number of clusters by stratum		
Region: North Bank vs. Lower River	36:20	35:20
Distance to main road:above vs. below median	28:28	27:28
Mean cluster distance to road in km (SD)	2.00 (2.92)	1.65 (2.80)
Number of randomized eligible children per cluster: mean (SD)	36.8 (20.8)	44.7 (35.7)
Number of villages per cluster:		
One village	35	32
Two villages	18	16
Three villages	1	5
Four villages	2	2
Mean cluster population (SD)	1188 (556)	1415 (1007)

Note: this table presents baseline characteristics of control and intervention clusters. Results correspond to Table 1 of analysis plan, see Appendix B.

control group. After excluding ineligible children (see Fig. 1) there were 2060 and 2458 children in the trial at baseline in the intervention and control groups, respectively.

In Table 3, we present characteristics of these randomized children and their caregivers, separately by intervention group. Following recommended best practice (Bruhn and McKenzie, 2009; Campbell et al., 2012; Moher et al., 2010) we do not conduct statistical tests for differences in baseline characteristics, as any differences would necessarily have arisen by chance. We see similar gender ratios, identities of the child's caregiver, caregiver education, caregiver literacy, and child age across intervention and control groups. In both groups, roughly three-quarters of parents had never gone to school. By chance, the ethnicity of children varies somewhat by randomization group, with control children somewhat more likely to be Wolof than intervention children. Such a chance imbalance is not unexpected given that randomization is by village cluster, with ethnicity variations being pronounced across villages. We chose not to collect baseline learning levels from children because, at the time they were enrolled in the study, our participants were exclusively children between age 6 and 8 who had not yet begun formal schooling. Our assumption was that all but a trivially small number of children would have registered zero scores, particularly because we selected study areas that were remote rural regions with very low baseline levels of literacy and numeracy among adults. The large number of zero scores among control group children at the endline test corroborates this assumption.

3.5. Intervention implementation

In 78 of the 82 villages assigned to receive the intervention, our after-school classes comprised 2 hours of teaching, given six times per week. In four villages, the schedule was adjusted slightly because a large proportion of students had to attend both traditional school and Qur'anic school during the week. Two of these villages held classes five times a week (four classes at 2 hours per class, and one 4-hour class) and two other villages held classes four times a week (two classes at 2 hours per class, and two at 4 hours per class).

We provide details on take-up of the intervention in Table 4. The first column shows that on a week-by-week basis there was no deviation from the schedule (this masks a substantial amount of on-the-ground rescheduling of classes, for instance, when teachers were sick). The second and third columns show that, on average, intervention children attended our after-school classes slightly more than 75 percent of the intended time.

3.6. Migration and attrition

Next, we describe the retention of participants in our study

Table 3
Baseline characteristics of children and their caregivers.

Variable	Intervention	Control
Number of children in the trial at baseline	2060	2458
Gender:		
Male	51.9% (1070)	50.4% (1239)
Female	48.0% (989)	49.5% (1217)
Not Known	0.0% (1)	0.1% (2)
Ethnic group:		
Mandinka	40.9% (842)	42.3% (1040)
Wolof	16.2% (334)	24.7% (608)
Fula	25.0% (516)	19.7% (485)
Other	15.1% (312)	11.4% (279)
Missing	2.7% (56)	1.9% (46)
Child's main caregiver:		
Biological mother	73.3% (1511)	74.6% (1833)
Biological father	3.3% (69)	5.2% (128)
Grandmother	11.0% (227)	10.1% (249)
Other	9.7% (200)	8.5% (208)
Missing	2.6% (53)	1.6% (40)
Caregiver's education:		
No education	73.8% (1520)	75.7% (1861)
Pre-school or primary	15.4% (318)	14.3% (352)
Junior secondary	5.3% (110)	5.8% (142)
Senior secondary or higher	2.6% (54)	2.3% (57)
Missing**	2.8% (58)	1.9% (46)
Child's age in September 2015	6.87 (SD = 0.55)	6.88 (SD = 0.55)
Language spoken in home:		
Mandinka	42.2% (869)	44.5% (1093)
Wolof	18.4% (379)	28.1% (691)
Fula	24.6% (506)	17.9% (440)
Other	12.1% (250)	7.6% (188)
Missing	2.7% (56)	1.9% (46)
Caregiver literacy at baseline:		
Can't read	75.6% (1557)	77.9% (1915)
Can read ≥*one letter, < one full word*	9.9% (204)	8.5% (209)
Can read ≥*one word, not entire card*	3.9% (80)	3.8% (93)
Read entire card slowly	2.8% (57)	3.6% (89)
Read entire card fluently	5.1% (106)	4.3% (105)
Refused	0.0% (0)	0.0% (1)
Missing	2.7% (56)	1.9% (46)
Wealth*:		
Category 1	7.3% (151)	4.6% (113)
Category 2	66.7% (1373)	66.5% (1635)
Category 3	23.3% (480)	27.0% (664)
Missing	2.7% (56)	1.9% (46)
Has an older sibling:	84.6% (1742)	85.4% (2098)
Has a younger sibling:	87.6% (1805)	90.3% (2220)

Note: Except when labeled otherwise, this table presents the group-specific proportion of children holding each characteristic with the number of observations in parentheses. We chose not to collect baseline learning levels from study children based on the assumption that, prior to entering formal schooling, all but a trivially small number of children in these regions would have registered zero scores. * We define wealth as a categorical measure defined by the materials of the roof, walls, and floor of the child's home at baseline. Category 1 is that all materials are natural (e.g., a thatched roof, mud walls, and an earthen floor); category 2 is that some but not all materials are synthetic (e.g., a steel roof, but natural walls and floor); category 3 is that all materials are synthetic (e.g., a steel roof, brick walls, and a tile or concrete floor). ** Two caregivers in the intervention group did not know their education level at baseline. They are grouped among the "missing." Results correspond to Table 2 of analysis plan, see Appendix.

Table 4
Adherence to the intervention.

	Percent of regularly scheduled classes actually held (village-level) N = 82	Percent of children attending each regularly scheduled class (village-level) N = 82	Percent of regularly scheduled classes the child attends (child-level) N = 2060
Mean (SD)	100% (0%)	78.9% (8.7%)	76.8% (27.8%)
Distribution:			
0%	–	–	0.7% (15)
>0 to 25%	–	–	10.1% (209)
>25% to 50%	–	–	3.7% (77)
>50% to 75%	–	30.5% (25)	10.2% (210)
>75% to 100%	100% (82)	69.5% (57)	74.6% (1537)
Missing	–	–	0.6% (12)

Note: The table shows the proportion of children and villages meeting pre-specified adherence targets in terms of the proportion of regularly scheduled intervention classes held at the village level (column 1), the proportion of children attending these classes (column 2) and the proportion of regularly scheduled classes children attended (column 3). We give the number of observations for each adherence level in parentheses next to the proportion. Results correspond to Table 3 of analysis plan, see Appendix B.

Table 5
Child presence in study village (migration) over course of trial.

Variable	Intervention (N = 2060)			Control (N = 2458)		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Present in village	93.2% (1920)	89.3% (1839)	89.6% (1845)	93.0% (2286)	87.7% (2156)	83.9% (2062)
Not present in village	6.6% (135)	8.4% (174)	9.6% (197)	6.7% (164)	10.4% (256)	15.1% (372)
Data missing	0.2% (5)	2.3% (47)	0.9% (18)	0.3% (8)	1.9% (46)	1.0% (24)

Note: This table provides the proportion of children who were present in the village at the time of our annual visits, which took place at the end of each academic year, e.g., May-July 2016, May-July 2017, and February-May 2018, for Year 1, Year 2, and Year 3, respectively. The relevant number of observations is given in parentheses under the proportion. Results correspond to Table 4 of analysis plan, see Appendix B.

throughout the course of the trial. At the end of each academic year, we visited all trial villages with the goal of locating each child who was randomized in the trial, asking her/his caregiver simple information about their activities in school the previous year, and recording whether or not the child was present in the village. In Table 5, we tabulate children's presence in the village across intervention and control groups and years of the study. In the first year of the study, less than seven percent of children are absent in either group. This increases to 8.4 and then 9.6 percent in the intervention villages in years two and three respectively, and to 10.4 and then to 15.1 percent in the control villages.

At endline, attrition is 5.7 percentage points higher in control villages. Should higher-ability children be more likely to leave a village in the absence of the intervention, this could bias our estimates upwards. Nonetheless, given the small magnitude of this difference, no more than a small fraction of the large differences in performance we measure between intervention and control children could be explained by differential attrition. In the next section we quantify the magnitude of this effect by computing Lee bounds on the estimated effect of the intervention. In the first column of Table 7, we also show attrition by treatment group and baseline trait, accompanying relevant heterogeneity analysis.

4. Main results

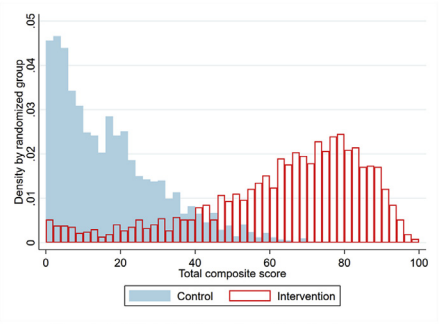
In this section we present our main results. We begin by showing the primary outcome – children's performance on the endline test, calculated as the arithmetic mean of their scores on the reading test and the mathematics test. We then discuss performance on the various subsections of the test, which vary in difficulty, followed by comparisons of the primary outcome across pre-specified subgroups. We also present results for our pre-specified secondary outcomes: the child's enrollment and attendance in school, the household's education-related expenditure on the child, and time use of the child and caregiver. We conclude this section with a brief discussion of the costs of the intervention.

4.1. Primary outcome

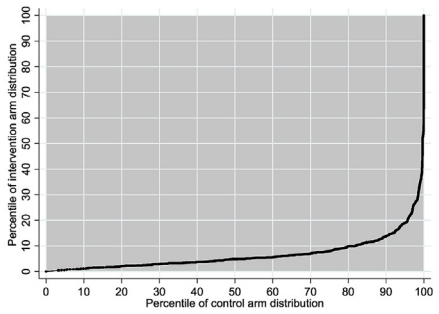
We show the distribution of our primary outcome, the composite test score, in Panel A of Fig. 2, separately for intervention and control children. We show group-specific means of the primary outcome and adjusted differences between them in the first row of Table 6. Intervention children score 46.0 percentage points higher on the test than control children (95% confidence interval: [43.3, 48.8]). This difference in means is highly statistically significant ($p < 0.0001$). In SD terms, this comprises a 3.2 SD difference; however, the SD measure is hard to interpret since the distributions of test scores in the two groups are far from Gaussian – with the consequence that the standard deviation is not a comprehensive description of dispersion – and markedly different between the two randomized groups. Accordingly, we also present the comparison as a percentile-to-percentile mapping of the two distributions in Panel B of Fig. 2. The 10th percentile of the intervention group distribution corresponds closely to the 80th percentile of the intervention group distribution. This implies that the effect of the intervention would be equivalent to moving a child at the 10th percentile of the control group distribution to the 80th percentile; further up the distribution, it suggests that the intervention would move a child at the 20th percentile of the control group distribution to the 95th percentile. The second and third rows of Table 6 provide analogous results for the overall reading and mathematics test scores, separately; these show differences similar to the overall difference in scores. In Panels A and B of Fig. 3, we present a bar chart showing mean scores, by intervention and control groups, for each component (also called a "subtask") of the mathematics and reading tests, respectively. As the number label on the subtask increases (e.g., from subtask 1 to subtask 2), so does its difficulty. The nature of each subtask is described in Table 1, and the full test papers are provided in Appendix D. The patterns in this figure show dramatic differences between control and intervention groups on all subtasks, from the easiest (letter recognition in reading and number recognition in mathematics) to the most difficult (reading comprehension in reading, and two- and three-digit subtraction with borrowing in mathematics). They also reveal very low learning levels in the control group. For the addition and subtraction subtasks on the mathematics test, the mean control group score is less than 20 percent correct for simple addition and less than 10 percent correct for advanced addition, simple subtraction, and advanced subtraction, respectively.[29] For the reading subtasks, the control mean score never exceeds 30 percent correct answers; for the five most difficult subtasks, it does not exceed 6 percent correct answers. In the intervention group, the mean mathematics subtask scores are between 80 and 95 percent correct answers for easier subtasks, and between 50 and 65 for the more difficult subtasks. For reading, these mean scores are between 47 and 69 percent correct.

We conduct ancillary analyses of the primary outcome to explore whether the effect of the intervention was greater for those intervention group children who attended classes more regularly than others. These comprise a series of pre-specified per-protocol analyses restricting

[29] The higher control group scores from subtask 6 likely reflect the fact that the question is spoken to the child and relies less on school-based knowledge than the previous subtask, 5b, which requires parsing written, two-digit subtraction problems. Subtask 6 questions are also computationally simpler (e.g., single digit adding or subtraction) than those in 5b.



Panel A: Distribution of endline test scores



Panel B: Percentile-to-percentile plot of endline test scores

Note: Panel A shows the distribution of child endline test scores, separately by randomization status. Panel B shows the percentile-to-percentile plots of these scores. The endline score is the arithmetic mean of the child's score on EGRA and EGMA tests. There are 1,815 children's scores in the intervention group and 2,071 in the control group.

Fig. 2. Distribution of primary outcome for intervention and control children.

Table 6
EGRA and EGMA test results.

Variable	Intervention	Control	Adjusted difference [95% CI]	P-value
Composite test score	63.3 (22.3)	17.1 (14.2)	46.0 [43.3, 48.8]	$p < 0.0001$
Composite test score difference in SD units	–	–	3.23 [2.89, 3.63]*	–
Mathematics test score, overall	68.2 (21.8)	24.7 (19.7)	43.4 [40.2, 46.5]	$p < 0.0001$
Reading test score, overall	58.3 (25.3)	9.5 (11.2)	48.7 [46.1, 51.4]	$p < 0.0001$
N taking test/N randomized (%)**	1815/2060 (88.0%)	2071/2458 (84.3%)	–	–

Note: Except where otherwise indicated, columns 1 and 2 show group-specific test score means, with standard deviations in brackets. In column 3, we show the estimated difference between column 1 and 2 adjusted for the randomization stratification factors (by including these factors as dummy variables in the regression model) with a 95% confidence interval (that takes into account the clustered design) in brackets below. In column 4 we present the p-value from the corresponding hypothesis test. * Bootstrap confidence interval, bias corrected and accelerated, based on 2000 bootstrap samples of clusters with stratification by intervention/control. ** We have one additional observation in reading for the control group (a child who took the reading test, but not the mathematics test). Results correspond to Table 5 of analysis plan, see Appendix B.

attention to only villages and children which held or attended classes, respectively, at a prespecified minimum adherence level. The estimates (reported in Appendix G) were slightly greater than those seen in the primary intention to treat analysis, particularly in the analysis restricted to the 74% (1525/2060) of intervention group children with at least 75% attendance.

In an exploratory analysis we calculate Lee bounds (Lee 2009) on the difference in mean composite test scores between groups, with 95% bootstrap confidence intervals that take account of clustering, in order to address the potential for differential attrition to affect our estimate of our primary outcome. This creates bounds on the treatment effect in the subset of the population who attended the examination in the control villages under the assumption that all of these children (along with some others) would have attended the examination were they in the intervention group. The "untightened" bounds on the difference in means between groups is [44.7, 48.8] with relatively tight 95% confidence intervals. We report various other estimates of the Lee bounds and 95% confidence intervals in Appendix H using different specifications, all of which yield results similar to those we present here. These suggest that differential attrition does not make a material contribution to the difference in endline test scores that we observe.

These tests target skills for which, according to the national curriculum, students should achieve mastery by the end of the second grade. The only questions which are not on the official national curriculum guidance for grade 2 are four questions we added to the EGMA tests (questions 5 and 6 on the addition level 2 and subtraction level 2 subtasks, respectively). Removing these questions changes the estimated difference in mean score between the randomized groups by only 0.2 percentage points.

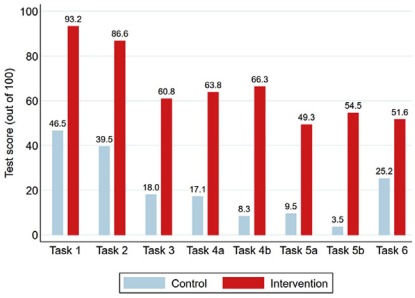
4.2. International comparisons

We next describe these learning gains in the context of other administrations of EGRA and EGMA tests in developing countries, both nearby and further afield. The reading speed and comprehension of intervention children was particularly remarkable. The transition from learning to read to reading to learn – that is, reading with comprehension to acquire new knowledge – is key to any early-grade reading program. In

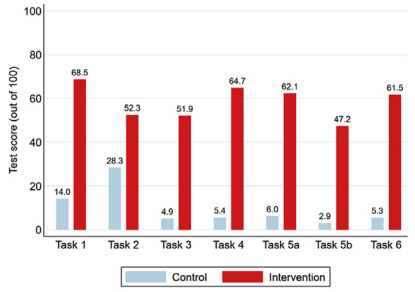
Table 7
Composite test scores by subgroup, with interaction tests.

Subgroup By group: N taking test/N randomized (%)	Intervention	Control	Adjusted difference [95% CI]	P-value for interaction
Gender				
Male	62.4	16.4	45.9	
(N: I=935/1070 (87.4%), C=1022/1239 (82.5%))	(23.1)	(14.1)	[42.9, 49.0]	p = 0.86
Female	64.1	17.9	46.2	
(N: I=879/989 (88.9%), C=1047/1217 (86.0%))	(21.4)	(14.2)	[43.2, 49.2]	
Wealth				
Category 1	60.9	15.2	45.6	
(N: I=138/151 (91.4%), C=93/113 (82.3%))	(24.9)	(12.5)	[38.5, 52.7]	p = 0.53
Category 2	63.5	16.7	46.7	
(N: I=1227/1373 (89.4%), C=1396/1635 (85.4%))	(21.8)	(13.8)	[43.9, 49.6]	
Category 3	63.7	18.6	45.1	
(N: I=429/480 (89.4%), C=570/664 (85.8%))	(22.1)	(15.3)	[41.5, 48.7]	
Ethnicity				
Mandinka	63.4	16.7	46.5	
(N: I=751/842 (89.2%), C=901/1040 (86.6%))	(21.6)	(13.6)	[42.3, 50.6]	p = 0.20
Wolof	63.8	16.4	47.4	
(N: I=295/334 (88.3%), C=504/608 (82.9%))	(24.6)	(14.2)	[43.0, 51.8]	
Fula	63.5	20.8	42.6	
(N: I=467/516 (90.5%), C=411/485 (84.7%))	(21.9)	(14.9)	[38.8, 46.3]	
Other	62.7	14.2	49.1	
(N: I=281/312 (90.1%), C=243/279 (87.1%))	(21.6)	(13.6)	[43.2, 55.0]	
Region				
Lower River	63.4	19.5	43.9	
(N: I=677/774 (87.5%), C=688/829 (83.0%))	(21.9)	(14.9)	[38.8, 48.9]	p = 0.27
North Bank	63.2	15.9	47.2	
(N: I=1138/1286 (88.5%), C=1383/1629 (84.9%))	(22.5)	(13.6)	[44.0, 50.4]	
Distance to road				
>median	63.9	16.2	47.6	
(N: I=731/823 (88.8%), C=836/987 (84.7%))	(22.8)	(13.4)	[43.3, 52.0]	p = 0.34
<median	62.8	17.8	45.0	
(N: I=1084/1237 (87.6%), C=1235/1471 (84.0%))	(21.9)	(14.6)	[41.6, 48.3]	
Caregiver education				
None	63.0	16.5	46.5	
(N: I=1364/1520 (89.7%), C=1579/1861 (84.8%))	(22.4)	(13.9)	[43.6, 49.3]	p = 0.11
Pre-schoolor primary	64.0	17.6	46.4	
(N: I=286/318 (89.9%), C=311/352 (88.4%))	(21.4)	(14.6)	[42.7, 50.1]	
Junior secondary	66.6	22.6	44.0	
(N: I=100/110 (90.9%), C=123/142 (86.6%))	(21.0)	(14.0)	[39.9, 48.1]	
Senior secondary or higher	61.4	23.3	38.1	
(N: I=39/54 (72.2%), C=43/57 (75.4%))	(21.9)	(15.4)	[30.4, 45.8]	
Exploratory Analysis, not pre-specified				
School in Village				
No	66.4	16.1	50.0	
(N: I=516/582 (88.7%), C=525/616 (82.5%))	(21.2)	(13.6)	[46.6, 53.5]	p = 0.008
Yes	62.0	17.5	44.5	
(N: I=1299/1478 (87.9%), C=1546/1842 (83.9%))	(22.6)	(14.3)	[41.3, 47.6]	

Note: This table shows a series of pre-specified tests for heterogeneity in our primary outcome. As in Table 6, columns 1 and 2 show the mean test score for the group (e.g., in the first row, males in the intervention and control group, respectively) with standard deviations in parentheses below. In column 3, we show the difference between column 1 and 2 adjusted for the randomization stratification factors with a 95% confidence interval (that takes into account the clustered design) in brackets below. Column 4 shows the p-value for a test of the hypothesis that there is no interaction between the categorical heterogeneity variable and receipt of the intervention. In the first column, under each predetermined characteristic we show, by randomized group, the number of children who took the endline test relative to the number of children randomized. Here "I" signifies the number for the intervention group and C the number for the controls. Results correspond to Table 6 of analysis plan, see Appendix B.



Panel A: Mathematics subtask test scores for intervention and control children



Panel B: Reading subtask test scores for intervention and control children

Note: Panel A shows the mean test score of children in control and intervention groups on each of the mathematics subtasks described in Table 1, and Panel B shows the corresponding means for each of the reading subtasks described in Table 1. Results correspond to Table 5 of analysis plan, see Appendix B.

Fig. 3. Test performance, by subtask.

the analysis of EGRA-like assessments, it is generally accepted that children are considered proficient readers when they read "with good fluency" and can correctly answer 80% or more of the reading comprehension questions associated with the text read (Dubeck and Gove, 2015). When asked to read a grade 2 level reading passage, around 43% of the intervention children were reading at least 45 words correct per minute – i.e., "with good fluency" – while less than 1% of the control children could do so. Intervention children were also much more likely to read with comprehension: 29% were able to correctly answer 80% of the questions related to the text, while only 0.1% of the control children could do so. Among other countries in the region that conducted similar EGRA assessments, 12% of Tanzanian grade 3 children and 7% of Ghanaian children could do so. In Iraq and in Jordan, when children were asked grade 2 level reading comprehension questions, 8% and 24% of the children, respectively, could answer 80% or more questions accurately.[30]

Another useful comparison is the percentage of children unable to read a single word. In the intervention group, this was around 8%. Among second and third graders in nearby Niger and northern Nigeria, this percentage is more than 80%. Zero scores in the intervention group are also notably less than grade 3 national figures in other sub-Saharan African countries (Uganda, 35%; Liberia, 11%), as well as other, more developed countries (Egypt, 35%; Morocco, 18%).[30]

The numeracy assessment results suggest that the majority of the intervention children had acquired basic mathematics skills and procedures essential for mastering higher-level mathematics concepts. Here, the most noteworthy differences in performance between intervention and control children were in tasks measuring conceptual knowledge and problem-solving skills. These show students' ability to make meaning, as opposed to simply memorizing arithmetic tables. In the tasks that require children to recognize number pattern and extension, for example, more than half of the intervention children answered at least 70% of the items correct, while less than two percent of the control children did so. When comparing with other countries that had assessed children with similar tasks, intervention children in our study were able to correctly answer twice as many conceptual questions as grade 2 and 3 children in Ghana and Tanzania. They also did better than grade 3 children assessed in Morocco and Iraq.[30]

4.3. Subgroup analyses

In Table 7, we present subgroup analyses of the primary outcome across pre-specified characteristics of the child and village. We find no evidence of a differential impact of the intervention by any of the following variables: gender; wealth; ethnicity; the region in which the village is located (Lower River or North Bank); and distance of the village to the road. There is some evidence that the intervention may be marginally more beneficial for children with less-educated caregivers, though the pattern we observe is not statistically significant. We also present one exploratory (i.e., non-pre-specified) analysis of subgroup heterogeneity, estimating the intervention's effect by whether or not there was a school in the village at baseline. This provides some evidence that the intervention's impact is somewhat larger for villages without schools (a 50.0 percentage point intervention/control difference) than for those with schools (a 44.5 percentage point intervention/control difference). This difference is statistically significant ($p = 0.008$), though its magnitude is small relative to that of the overall intervention/control difference and interpretation should be done with caution since multiple subgroupings have been considered here and this was not a pre-specified analysis.

[30] References for these comparisons: Jordan: https://earlygradereadingbarometer.org/files/EGRA%20in%20Jordan.pdf and https://www.globalreadingnetwork.net/sites/default/files/eddata/01_Jordan_Intervention_Final_Report_07_April_2015_English1.pdf; Kenya: https://ierc-publicfiles.s3.amazonaws.com/public/resources/Tusome%20Midline%20evaluation%202017%20final%20report%20from%20DEC.pdf; Liberia: https://earlygradereadingbarometer.org/files/EGRA%20in%20Liberia.pdf; Morocco: https://ierc-publicfiles.s3.amazonaws.com/public/resources/2-Morocco%20Final%20Report_ENGLISH_WITH%20INSTRUMENTS_26Apr2012.pdf; Iraq: https://earlygradereadingbarometer.org/files/EdDataII%20Maharat_ExeSummary_Iraq.pdf; Tanzania: https://www.globalreadingnetwork.net/sites/default/files/resource_files/2016%20TZ%20EGRA%20Findings%20Report.pdf; all accessed April 19, 2020.

4.4. Pre-specified secondary analyses

In this section, we present a series of pre-specified secondary analyses, including estimation of the impact of the intervention on children's enrollment in school, attendance in school, time use, and family resource expenditure on the child's schooling.

Enrollment and attendance in school: We estimate how the child's enrollment in school and attendance at school over time varies by treatment status. Enrollment data come from caregiver surveys administered at the end of each academic year. Attendance data, as described below, come in two forms: from these caregiver surveys and from school-level administrative data. In Panel A, we show children's enrollment in school over the three years of the trial. We see that in years two and three of the trial, the odds of enrollment in school are 56% and 92% higher for intervention children than for control group children. These correspond to absolute differences, after controlling for baseline stratification variables, of 9.6 and 11.3 percent, respectively, in the proportion of children who are enrolled in grade 1 or above.[31] For attendance, we have two measures: parents' report of the number of days in the past two weeks the child missed school (higher numbers indicate a greater number of days missed), and administrative data from the child's school on the percent of regularly scheduled school days the child attended (higher numbers mean a greater percent of days attended).[32] Overall, conditional on being enrolled in school, children in intervention villages are less likely to miss school than children in control villages, though the statistical significance of these differences varies across different measures of attendance.

Expenditure and time use: we next present differences between intervention and control groups in terms of three key family inputs into education: financial expenditure on education, the proportion of the child's waking time spent on education, and the amount of time the child's caregiver spends helping the child with homework. These data were collected from the child's caregiver during the endline survey. We present results in Table 9. Families in the control group may spend slightly more money than those in intervention villages on school-related expenditures, but this difference is not significant at traditional levels.[33] As expected given the time-intensive nature of our intervention, caregivers of children in intervention villages report the child spending significantly more time in school-related activities. We find no evidence of a difference between intervention and control children in the amount of time the caregiver spends helping the child with schoolwork. In Appendix F, we report pre-specified secondary analyses of whether the learning gains for children in the intervention group had any spillover effects on literacy and numeracy of their siblings and parents. We find little evidence of any such effects.

Finally, we estimate whether the intervention had any impact on the activity of school management committees, which are local organizations that help with school operations. This is a test for whether the intervention increased demand for schooling in these communities. We see that all schools attended by children in both intervention and control villages have active committees, consistent with a high level of demand for schooling in these areas. These results are presented in Appendix B, Table 12.

4.5. Costs and cost-effectiveness

In this section, we discuss the cost of the intervention as it was implemented during the course of the study, how this may vary in future implementation, and the likely cost under government implementation.[34] Our main costs comprise the following components: teacher, monitor, and other staff salaries, benefits, and incidentals; design, piloting, printing, binding, and shipping of teaching and learning materials; purchasing, importing, maintaining, and fueling vehicles; the construction, renovation, and maintenance of a main office and a field office; staff training expenses (food, lodging, transport, per diems, and training materials); and various administrative costs (such as accountancy, taxes, insurance, HR) that come with running a stand-alone organization of roughly 150 employees. We capitalize vehicle costs and, separately, office construction and other capital expenses, over an expected lifespan of 10 and 20 years, respectively. We express our figures in 2015 dollars and use an annual discount rate equivalent to the US Consumer Price Index for that year.[35]

Total expenditure for running the intervention was 1.493 million US dollars. We calculate the per-child cost of implementing this intervention by dividing the total expenditure by the 2060 children in the study at baseline in villages assigned to receive the intervention. This generates a per-child cost of $724.77, or $241.59 per child per year. In this implementation, we focused primarily on maximizing the fidelity of the intervention and not on cost minimization. In future implementation at larger scale, we estimate that our per-child cost may decrease by up to 30 to 40 percent, under the assumption that we could substantially reduce the expenses we bore for administrative costs (e.g., senior staff costs, rent of the office in the capital, field office construction, vehicle purchase and maintenance, and other related expenditures).

Using the "additional SD per $100" metric from Kremer et al. (2013), and using only the 1815 children who took the endline test as the denominator, we estimate a 0.389 SD increase per $100 spent with the existing costs. In Appendix I, we conduct a rudimentary cost-benefit analysis which is constrained by the unfortunate paucity of labor market data from The Gambia.

5. Discussion

In this section, we address two questions. First, why are our impact estimates so large? Second, what mechanisms may be behind these large effect estimates?

5.1. Magnitude of effect estimates

Here we discuss the large magnitude of our effect estimates. We think this is likely to stem from two factors: one, the comprehensive nature of the intervention and possible complementarities between its component parts; and two, the particularly low learning levels in this context, which make such large gains possible.

The intervention combines at least three strategies shown in prior research to be effective in similar settings. First, the short-term nature of para teacher contracts has been shown to increase teacher performance, incentivizing greater effort (Banerjee et al., 2007; Duflo et al., 2015;

[31] Note that this parameter is estimated using only children who are enrolled in school. Since, as the trial progresses, the intervention induces some children to stay enrolled in school who otherwise might not have, its interpretation comprises both the attendance and enrollment effects of the intervention.

[32] There are fewer observations in the administrative data because in some cases we were unable either to find or to uniquely identify the child's name in the school register.

[33] While this difference is not significant, its sign is consistent with evidence from other developing country contexts of substitution behavior by parents in response to education interventions (Das et al., 2013).

[34] Our implementation involved entering the country and establishing a service delivery apparatus entirely independent of the government. Our initial calculation includes all the attendant costs that this incurred. While our strategy here may be optimal in some country contexts, in many other contexts the government should be able to implement a version of this intervention using current staff, either after-school or during the school day, which would imply a substantial reduction in costs. This is substantiated by the results of Piper et al. (2018), who show sizeable gains from a similarly multipronged intervention implemented within the Kenyan government system.

[35] Taken from www.usinflationcalculator.com/inflation/current-inflation-rates/on May 15, 2019.

Table 8
Enrollment and attendance in school.

Panel A: Enrollment in School (grade 1 or above) in each academic year (AY)

Enrollment in school *(grade 1 or above)*	Intervention	Control	Adjusted Difference [95% CI]	Adjusted Odds ratio [95% CI]	P-value
Year 1 (AY, 2015–16)	47.5% (978)	42.6% (1046)	4.7% [–2.1, 11.4%]	1.21 [0.92, 1.59]	p = 0.179
Year 2 (AY, 2016–17)	73.0% (1503)	63.1% (1551)	9.6% [3.7, 15.4%]	1.56 [1.18, 2.07]	p = 0.002
Year 3 (AY, 2017–18)	82.8% (1706)	71.4% (1756)	11.3% [6.9, 15.7%]	1.92 [1.50, 2.45]	p < 0.001
Number of observations	2060	2458	–	–	–

Panel B: Grade child is enrolled in during each academic year

	Intervention (N = 2060)	Control (N = 2458)
Year 1: AY 2015-16		
Not in school	17.8% (367)	15.3% (377)
ECD*/Nursery	26.1% (538)	33.5% (823)
1	45.3% (934)	40.2% (987)
2	1.7% (34)	2.0% (50)
3	0.3% (7)	0.3% (8)
4 or 5	0.1% (3)	0.0% (1)
Don't know	1.8% (37)	1.6% (40)
Missing	6.8% (140)	7.0% (172)
Year 2: AY 2016-17		
Not in School	11.7% (240)	15.3% (377)
ECD/Nursery	4.7% (96)	9.3% (228)
1	32.0% (659)	35.8% (880)
2	39.4% (811)	26.4% (648)
3	1.4% (29)	0.9% (23)
4 or 5	0.2% (4)	0.0% (0)
Don't know	0.0% (0)	0.0% (0)
Missing	10.7% (221)	12.3% (302)
Year 3: AY 2017-18		
Not in School	2.0% (42)	2.4% (59)
ECD/Nursery	0.7% (15)	3.3% (81)
1	7.6% (156)	17.1% (421)
2	33.1% (681)	32.6% (802)
3	40.2% (829)	20.9% (514)
4 or 5	1.9% (40)	0.8% (19)
Don't know	0.0% (0)	0.0% (0)
In school, but grade missing	4.0% (82)	6.8% (166)
Missing	10.4% (215)	16.1% (396)

Panel C: Attendance (conditional on enrollment in grade 1 or above)

	Intervention	Control	Adjusted difference [95% CI]	P-value
Caregiver report of the number of days the child was absent from school in the two weeks prior to being surveyed				
Year 1: AY 2015-16	0.56	0.75	−0.20 [−0.39, −0.00]	p = 0.047
(N: I=977, C=1044)	(1.65)	(2.02)		
Year 2: AY 2016-17	0.42	0.52	−0.09 [−0.24, 0.06]	p = 0.247
(N: I=1500, C=1550)	(1.42)	(1.71)		
Year 3: AY 2016-17	0.52	0.56	−0.04 [−0.16, 0.08]	p = 0.514
(N: I=1701, C=1748)	(1.54)	(1.70)		
School's record of child's attendance throughout study				
Percent of regularly scheduled classes child attended during the study	81.1	75.1	6.0 [1.2, 10.8]	p = 0.016
(N: I=1565, C=1589)	(21.6)	(26.2)		

Panel A note: This shows the proportion of children enrolled in grade one or above in each academic year, by intervention status, with the number of enrolled children given below in parentheses. Columns 1, 2, 4, and 5 follow the convention of Table 6. To aid interpretation, we also present adjusted differences in Column 3. Results correspond to Table 7 of analysis plan, see Appendix B.

Panel B note: We chose not to conduct hypothesis tests for equivalence/treatment effects for this secondary outcome in order to avoid the multiple comparison problem (i.e., to minimize the risk of Type 1 error). While we recognize this is less customary among economists than conducting tests with post-hoc adjustment for multiple comparisons, this paper was a collaborative effort between economists and medical statisticians, and the decision was made in keeping with usual practice in medical statistics (c.f., Moher et al., 2010; Campbell et al., 2012). As before, the number of observations for a given cell is given in parentheses next to the proportion. ECD stands for the (often informal) early child development classes held in some villages. Results correspond to Table 7 of analysis plan, see Appendix B.

Panel C note: Columns 1, 2, 3, and 4 follow the convention of Table 6. The number of observations are given, by treatment group, under the description of each variable. These vary with the enrollment of children in grade 1 or above across years and the missingness of data in different surveys. In AY 2015–16, there was one child from the intervention group and two from the control group who reported being enrolled in school and missing some school, but for whom days missed were not recorded. In AY 2016–17, there was one intervention child and three control children with missing data for both of these questions. In AY 2017–18, there were four intervention children and eight control children with missing data for these questions. There was also one intervention child whose guardian reported the child missing school, but the number of days missed was not recorded. We pre-specified that we would also estimate bootstrap confidence intervals, bias corrected and accelerated, based on 2000 bootstrap samples of clusters with stratification by randomized group. These are included in the appendix tables, but because they are so similar to the conventional confidence intervals, we do not include them here. Results correspond to Table 7 of analysis plan, see Appendix B.

Table 9
Caregiver spending on education, and school-related time use of caregiver and child.

Variable	Intervention Mean (SD)	Control Mean (SD)	Adjusted difference [95% CI]	P-value
Total parental spend in past year (Gambian Dalasis*) (N: I = 1803, C = 2003)	591 (438)	659 (528)	−66 [−147, 14]	p = 0.106
School-related time use of child (proportion of child's waking hours)** (N: I = 1845, C = 2062)	0.683 (0.123)	0.553 (0.140)	0.130 [0.113, 0.147]	p < 0.001
Number of hours caregiver spends helping child with homework per week (N: I = 1803, C = 2003)	3.08 (4.27)	2.99 (4.29)	0.09 [−0.34, 0.53]	p = 0.678

Note: Columns 1, 2, 3, and 4 follow the convention of Table 6. We pre-specified that we would also estimate bootstrap confidence intervals, bias corrected and accelerated, based on 2000 bootstrap samples of clusters with stratification by randomized group. These are included in the appendix tables, but because they are so similar to the conventional confidence intervals, we do not include them here. * Gambian Dalasis, average exchange rate over period of trial: 43.72 Dalasis per 1 US dollar. ** School-related time use of child measured as proportion of non-sleeping hours spent in school or on homework. Note that these data come from *all* children in our sample, not just those enrolled in school (as is the case for Table 8, Panel C). Results correspond to Table 8 of analysis plan, see Appendix B.

Muralidharan and Sundararaman, 2013). Second, we conducted high-frequency monitoring of our para teachers. The main purpose of this monitoring was to improve teacher effectiveness through providing regular feedback on teaching methods and practice, also called "coaching" in recent research (Kraft et al., 2018; Muralidharan et al., 2017; Piper et al., 2018a, 2018b). Third, we used a curriculum, shown to be effective elsewhere, that included scripted daily lesson plans and greater teacher-student interaction (Banerjee et al., 2017; Lakshminarayana et al., 2013; Shalem et al., 2016). In light of the substantially smaller estimated effects of the prongs when implemented individually (e.g., Banerjee et al., 2017; Muralidharan et al., 2017; Muralidharan and Sundararaman, 2013), we argue that our results provide some suggestive evidence of complementarity between these interventions, though our study was not designed to test this hypothesis. This is consistent with recent work from Tanzania showing complementarities between programs that change teacher incentives and those which provide additional resources to schools (Mbiti et al., 2019).

The other likely main contributor to these large relative gains is the extremely low learning levels in rural regions of The Gambia. The area in which we work is remote, difficult to reach, and very poor. Learning levels in these areas have been consistently low: government assessments of third grade children in villages similar to those in our trial show performance roughly similar to that of our control children. Other studies have found large learning gains from delivering interventions to similar settings and populations (Burde and Linden, 2013; Lakshminarayana et al., 2013).

There are a few important alternative explanations for the large effects we estimate. One is the potential for teaching to the test by our intervention team. The intervention used materials adapted from the materials used in Lakshminarayana et al. (2013) to follow the Gambian curriculum. As described in Section 2, the Gambian Ministry of Basic and Secondary Education regularly uses EGRA and EGMA tests to assess both students and teachers. As a result, government teachers were incentivized to teach the specific skills assessed in the EGRA and EGMA tests. In addition, the higher-level skills tested here are quite general – word recognition, reading comprehension, and arithmetic. The performance gap between intervention and control children is greater for these skills than for the more basic skills, such as letter naming or number sequences, that might be considered more test-specific. This pattern runs contrary to what we would expect to see should there have been teaching to the test.

The second is potential enumerator bias or leakage of the test paper. We took great care to minimize the risk of these two potential issues. With regards to enumerator bias, EGRA and EGMA have rigid rules for implementation. Our training and supervision of the enumerators who administered the tests emphasized close adherence to these rules. Enumerators were recruited independently of our other research activities, and were not told of a village's assignment to either intervention or control. Furthermore, two authors (Eble and Hsieh) travelled to The Gambia to supervise the assessments in order to guard against such bias. We also took great lengths to ensure our test paper was not leaked. While we conducted an initial pilot and adaptation session in The Gambia as per the EGRA and EGMA guidelines, the final items on the endline test were selected by Hsieh after the pilot and brought to The Gambia for printing only in May 2018, a few days prior to the training of enumerators. During the test administration, Eble and Hsieh conducted occasional interviews with children to test for possible cheating on the test. To do so, we identified high performers on the test and informally interviewed them to see whether they could answer spontaneous questions of similar difficulty to higher level tasks on the test, but with different content. In all cases, they performed in line with their performance on the test itself, further suggesting no evidence that the test paper was leaked. Finally, in the data analysis, we did not notice any unusual outcomes that would suggest the test leaked to any specific clusters/classes, nor did we notice unusual results by question, e.g., that children performed, on average, more poorly on the easier questions than the more difficult ones.

Finally, we explain why we think the impact of this intervention in The Gambia is so much larger than the impact of a similar intervention, implemented in India and reported in Lakshminarayana et al. (2013). The main reason for this pattern, we believe, is the match between the strengths of the intervention and the lower levels of learning in The Gambia than in India. The Indian children who were participants in that study had much higher mastery of basic skills than the Gambian children in this study. We believe that this intervention is most effective at teaching low level skills, as it relies upon previously untrained teachers implementing scripted lessons. We found that the baseline levels of learning of Gambian CEs limited their capacity to teach higher-level skills without more extensive training than we were able to provide. In other words, it was more difficult to train our Gambian CEs to teach the content in the third grade curriculum than either that of the first or second grade curricula. Another likely contributor is the duration of the program. The intervention in The Gambia lasted a full academic year longer than in India, allowing for potentially greater compounding of effects.[36]

5.2. Mechanisms

In this section, we conduct exploratory analyses to examine potential mechanisms behind our results. We use mediation analysis to estimate the relative contribution of three potential mediators to the endline difference in test scores: getting children to enroll in school, getting them to attend school, and getting them to follow normal grade progression in school. We use a similar analysis to show how important attendance in our intervention classes is for driving endline test score differences. At the end of the section, we briefly discuss the interaction between local

[36] The India study was also much less expensive. For India, we estimate 1.40 SD per $100, whereas in Gambia we estimate only 0.389 SD per $100. This suggests that efforts to reap larger learning gains, particularly in more difficult to reach areas, may face exponentially higher costs.

Table 10
Mediation analysis of enrollment, attendance, and grade progression.

Mediator	*Intervention endline scores*		*Control endline scores*		Decrease in primary outcome if intervention mediator distribution set to control mediator distribution (95% CI*)
	As-is	Like control	As-is	Like intervention	
Government schools as mediator					
Enrollment	63.4	60.2	17.2	18.4	6.9% (4.2, 10.0%)
Attendance	63.4	61.0	17.2	18.6	5.2% (1.8, 9.1%)
Grade progression	63.4	56.8	17.2	21.3	14.2% (9.5, 18.7%)
Intervention as mediator					
Attendance	63.7	19.2	17.2	–	95.6% (85.8, 104.1%)

Note: This table shows mediation analysis for potential mediators. Rows 1 to 4 show mean predicted endline test scores for all children under the four possible scenarios defined in the column headings. For example, in column 1, the first row shows the mean predicted endline test scores for all participants were they in the intervention group with school enrollment as in that group (in other words, "as-is") and in column 2, it shows their mean predicted score were they in the intervention group but their enrollment in school at endline mirrored that of the control group. In column 3 the first row shows the mean predicted endline test scores for all participants were they in the control group with school enrollment as in that group (in other words, "as-is") and in column 4, it shows their mean predicted score were they in the control group but their enrollment in school at endline mirrored that of the intervention group. Column 5 shows, in percentage terms, the decrease in the primary outcome that we would expect if the intervention value of the mediator would be set to that of the controls, e.g., the value in column 1, minus that in column 2, divided by the value in column 1, minus that in column 3. *Bootstrap confidence interval, bias corrected and accelerated, based on 2000 bootstrap samples of clusters with stratification by intervention/control. This is exploratory (non-prespecified) analysis.

schools and our intervention.

Mediation analysis is a tool from statistics, also occasionally used in economics (c.f. Imai et al., 2010; Imai et al., 2011, and Heckman and Pinto, 2015) which attempts to estimate how much of an intervention's effect can be attributed to a given mediator.

This is done by evaluating, in expectation, how much the composite score would be expected to change if the distribution of that mediator (e.g. school enrollment) were to change from that in the control group to that in the intervention group, ceteris paribus. The process of evaluating this change, termed an indirect effect, builds on the so-called mediation formula (Pearl, 2012). In a randomized trial there are two distinct indirect effects. The first, termed the "total indirect effect", estimates the effect of changing the mediator distribution were all children randomized to receive the intervention, and is our primary focus here. The second, termed the "pure indirect effect" estimates the effect of changing the mediator distribution were all children randomized to receive the control. A key concept in their calculation is that each child in the trial is hypothesized to have several potential outcomes, dependent on which group they were randomized to ("on intervention" or "on control") and mediator level.

A two-step process is used to calculate the total indirect effect. First, we estimate the probability distribution of the mediator (e.g., school enrollment) for each child. We estimate this separately for each randomized group, allowing the estimated probabilities to depend on covariates.[37] Second, for the intervention group we estimate mediator-specific expected composite scores for each child using a linear regression model with the mediator and the same covariates as predictor variables. We use the second step model to create one prediction of the composite score for each child corresponding to each possible value the mediator can take, fixing the covariates at their observed values. We then take a weighted average of these multiple predictions per child (one for each level of the mediator), weighting by the probabilities of these mediator values in the control group for that child, as obtained from the first step model. The resulting mean estimates the expected composite score that would have been observed on intervention if, for each child, the mediator took on the value it would take on control (column 2 in Table 10). Weighting instead by the probabilities of these mediator values in the intervention group yields an estimate of the expected composite score on intervention (column 1 in Table 10). An exactly analogous approach, but using mediator-specific expected composite scores for each child in the control (rather than the intervention) group at the second step, allows prediction of the expected composite score that would have been observed on control if, for each child, the mediator took on the value it would take on control (column 3) or on intervention (column 4 in Table 10); the difference between them being the pure indirect effect. Dividing the total indirect effect by the total intervention effect (i.e., the difference between columns 3 and 1 in Table 10) expresses the total intervention effect as a percentage.

We conduct this for four possible mediators. The first three have to do with the interaction between government schools and our intervention: one, enrollment in school as a mediator; two, attendance in school as a mediator; three, grade progression in school as a mediator. The fourth expresses how frequently the intervention classes were attended. Since children in the control group did not attend intervention classes, these children were assigned to the 0–50% attendance stratum. Here, the expected composite score that would be observed in the control group if attendance of the intervention classes was as in the intervention group is ill-defined; it is therefore not shown in the table.[38]

Our results, presented in Table 10, show that the intervention's effects on enrollment and attendance play a small role in the large difference in test outcomes (5–7% of the overall effect). The role of grade progression, however, is larger – the intervention/control difference in endline test score would decrease by almost 15% if intervention children progressed through grades in the same way that control children did. Finally, we see that attendance in the intervention classes is crucial for the success of the program. Endline scores of intervention children who did not attend, or did not regularly attend, intervention classes are very close to those of control children.

These results should be considered with some caution for several reasons. First, we consider each mediation variable in isolation, yet in reality they are inter-related. For example, a child had to be enrolled to progress through grades, and the effect of the intervention on a child's ability likely affected grade progression. Second, because the relationship between the mediator and the outcome is not a randomized comparison, confounders of this relationship need to be included in the analysis. We have included several potential confounders,[39] but there remains possibility of residual confounding. Finally, another relevant

[37] In this case, our covariates are region (binary), distance to road (binary), gender (binary), wealth category (three levels), ethnicity (four levels), caregiver education (four categories) and whether or not there was a school in the village (binary).

[38] We give the full tables of calculated expected outcomes, along with further description of how they were calculated, in Appendix K.

[39] The confounders we included were geographic location, distance to road, gender, wealth category, ethnicity, caregiver education and whether or not there was a school in the village.

caveat is that some children in both groups are missing outcome data[40] and this likely impacts our results somewhat.

We close this section with a discussion of interactions between intervention children, community educators, and government schools. We were not able to collect operations data from government schools throughout the study, but we briefly describe the interactions we observed between students from the trial intervention villages, teachers providing the trial interventions, and teachers in the government school system. As the intervention progressed, these students often rose to the top of their school classes, and, anecdotally, reported receiving performance awards from their teachers in their government schools. Teacher absenteeism in government schools in The Gambia is similar to or less severe than in other countries (Blimpo et al., 2015). Nonetheless, we were told that the punctuality and work ethic of the community educators was noticed by the local government school teacher. In some cases, this may have led to greater effort on the part of the teacher; in other cases, it was reported to have led to negative feelings and perhaps deliberate obstruction of our work by the local teacher. Overall, we note that our estimates comprise the direct impact of this intervention and any interaction effects it may have had with local government schools.

6. Conclusion

We find that a para teacher intervention combining frequent monitoring with an improved curriculum, shown to raise literacy and numeracy levels dramatically in rural India among primary-aged children, had even larger effects on literacy and numeracy levels among such children in rural parts of The Gambia.

The first policy implication of our work is that there exists a demonstrated way to reach the learning gains that many have called for in particularly disadvantaged areas. The intervention we study combines several best practices from the literature[41] and has dramatically raised learning outcomes in two diverse contexts: central regions of The Gambia and southern Telangana, India. This work has definitively shown that contracting this type of intervention out to an NGO *can* lead to dramatic improvements in educational outcomes in such areas. Our approach is similar to the "graduation" model studied in Banerjee et al. (2015) which uses a multifaceted intervention to establish sustainable self-employment and generate lasting well-being improvements among the extremely poor. Together, these results suggest that greater expenditure may be necessary to reap such large learning gains in other particularly poor and remote areas. We believe these interventions are also scalable. In The Gambia, India, and in Guinea Bissau, we have expanded our projects based on our estimates of their impact. In this work, we have been successful at recruiting a sufficient number of qualified para teachers (hundreds in both Gambia and India) and teachers (dozens in Guinea Bissau) to meet our needs.

The second implication is that the choice facing governments and donors hoping to reap such gains in similar areas is whether to attempt to operationalize this within the government system, or contract it out. We argue that large gains may be possible within the confines of existing systems. While operationalizing this type of intervention is likely to be both logistically and politically challenging for government (Bold et al., 2018), the component pieces of our intervention – a change in contract structure; improved, scripted curricular materials; and more frequent monitoring with a focus on pedagogic improvement – are all hypothetically implementable within the existing school hours and the existing school budgets of such systems. Recent work has shown that an intervention which focuses on raising literacy levels, and which combines similar components, was scalable in government schools in Kenya (Piper et al., 2014, Piper et al., 2018a, 2018b). A key challenge is that such changes may be demotivating or otherwise threatening to existing teachers. These teachers are important stakeholders in the education system who may exert political pressure to oppose these changes. The main limitation of our study is that it cannot speak to whether such an intervention would yield similar gains if implemented by the government, nor even if this is possible in settings with strong political pressure in favor of the status quo arrangement for delivering primary education.

Our findings suggest two possible paths for governments of countries facing similar problems. Obviously, government implementation of such a program within the existing funding system and school day would be ideal, and is worth aiming at. However, this may not be possible in nations where either 1) there exists strong political opposition to changes in teacher contracts and supervision, 2) government capacity, including that for financing, is low, or 3) both. Indeed, some of our collaborators in The Gambia have argued that it would be politically difficult to implement the type of restructuring this would imply for the Gambian education system. When such a change in bureaucracy is not possible, our results show that there remains room for large gains in learning levels through contracting these services out to NGOs. Even in this case, however, financing issues and perhaps others will remain.

Finally, our results suggest that large, supply-side interventions may be necessary to raise learning levels in particularly disadvantaged areas similar to the one we study. Several data points suggest high demand for education in our trial area. These include the high levels of enrollment in school, both in the data we collected and in the national gross enrollment data, and the common presence of school management committees, parent-run groups which aid local schools' functioning. Despite this, learning levels in these areas remain tragically low. Our study introduced experimental variation in the supply of education, and we find that this corresponds to massive learning gains. Together, we interpret these findings as evidence of the primary importance of the supply side in raising basic learning levels in remote areas of extreme poverty.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jdeveco.2020.102539.

References

Bandiera, Oriana, Buehren, Niklas, Burgess, Robin, Goldstein, Markus, Gulesci, Selim, Rasul, Imran, Sulaiman, Munshi, 2020. "Women's empowerment in action: evidence from a randomized control trial in Africa. Am. Econ. J. Appl. Econ. 12 (1), 210–259. https://doi.org/10.1257/app.20170416.

Bandiera, Oriana, Burgess, Robin, Deserranno, Erika, Morel, Ricardo, Rasul, Imran, Sulaiman, Munshi, 2018. Social Ties and the Delivery of Development Programs. Working Paper.

Banerjee, Abhijit, Banerji, Rukmini, Berry, James, Duflo, Esther, Kannan, Harini, Mukerji, Shobhini, Shotland, Marc, Walton, Michael, 2017. From proof of concept to scalable policies: challenges and solutions, with an application. J. Econ. Perspect. 31 (4), 73–102.

Banerjee, Abhijit, Cole, Shawn, Duflo, Esther, Linden, Leigh, 2007. Remedying education: evidence from two randomized experiments in India. Q. J. Econ. 122 (3), 1235–1264. https://doi.org/10.1162/qjec.122.3.1235.

Banerjee, Abhijit, Duflo, Esther, Goldberg, Nathanael, Dean, Karlan, Osei, Robert, Parienté, William, Shapiro, Jeremy, Thuysbaert, Bram, Udry, Christopher, 2015. A multifaceted program causes lasting progress for the very poor: evidence from six countries. Science 348 (6236).

Blimpo, Moussa P., Evans, David K., Lahire, Nathalie, 2011. School-Based Management and Educational Outcomes: Lessons from a Randomized Field Experiment (Unpublished Manuscript).

Blimpo, Moussa P., Evans, David K., Lahire, Nathalie, 2015. Parental Human Capital and Effective School Management: Evidence from the Gambia. World Bank Policy Research Working Paper 7238.

Bold, Tessa, Filmer, Deon, Martin, Gayle, Molina, Ezequiel, Stacy, Brian, Rockmore, Christophe, Svensson, Jakob, Wane, Waly, 2017. Enrollment without learning: teacher effort, knowledge, and skill in primary schools in Africa. J. Econ. Perspect. 31 (4), 185–204.

[40] In the mediation analysis we additionally exclude some children missing covariate data, but the number of additional exclusions is small, see Appendix J for details.

[41] Altering teachers' contract structures; using improved curricular materials, including scripted lessons; and monitoring teachers frequently with the goal of improving their teaching practice.

Bold, Tessa, Kimenyi, Mwangi, Mwabu, Germano, Ng'ang'a, Alice, Sandefur, Justin, 2018. Experimental evidence on scaling up education reforms in Kenya. J. Publ. Econ. 168 (December), 1–20. https://doi.org/10.1016/j.jpubeco.2018.08.007.

Boone, Peter, Camara, Alpha, Eble, Alex, Elbourne, Diana, Fernandes, Samory, Frost, Chris, Jayanty, Chitra, Lenin, Maitri, Silva, Ana Filipa, 2015. Remedial after-school support classes offered in rural Gambia (the SCORE trial): study protocol for a cluster randomized controlled trial. Trials 16 (1), 574.

Brudevold-Newman, Andrew, Honorati, Maddalena, Jakiela, Pamela, Owen, Ozier, 2017. A Firm of One's Own: Experimental Evidence on Credit Constraints and Occupational Choice (Working Paper).

Bruhn, Miriam, McKenzie, David, 2009. In pursuit of balance: randomization in practice in development field experiments. Am. Econ. J. Appl. Econ. 200–232.

Burde, Dana, Linden, Leigh L., 2013. Bringing education to Afghan girls: a randomized controlled trial of village-based schools. Am. Econ. J. Appl. Econ. 5 (3), 27–40.

Campbell, Marion K., Piaggio, Gilda, Elbourne, Diana R., Altman, Douglas G., 2012. Consort 2010 statement: extension to cluster randomised trials. BMJ 345, e5661.

Chaudhury, Nazmul, Hammer, Jeffrey, Kremer, Michael, Muralidharan, Karthik, Rogers, F. Halsey, 2006. Missing in action: teacher and health worker absence in developing countries. J. Econ. Perspect. 20 (1), 91–116.

Das, Jishnu, Dercon, Stefan, James, Habyarimana, Krishnan, Pramila, Muralidharan, Karthik, Venkatesh, Sundararaman, 2013. School inputs, household substitution, and test scores. Am. Econ. J. Appl. Econ. 5 (2), 29–57.

Dubeck, Margaret M., Amber, Gove, 2015. The early grade reading assessment (EGRA): its theoretical foundation, purpose, and limitations. Int. J. Educ. Dev. 40, 315–322.

Duflo, Esther, Dupas, Pascaline, Kremer, Michael, 2015. "School governance, teacher incentives, and pupil–teacher ratios: experimental evidence from Kenyan primary schools. J. Publ. Econ. 123 (March), 92–110. https://doi.org/10.1016/j.jpubeco.2014.11.008.

Evans, David K., Anna, Popova, 2016. What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. World Bank Res. Obs. 31 (2), 242–270. https://doi.org/10.1093/wbro/lkw004.

Fazzio, Ila, Eble, Alex, Lumsdaine, Robin L., Peter, Boone, Baboucarr, Bouy, Hsieh, Pei-Tseng Jenny, Chitra, Jayanty, Simon, Johnson, Ana Filipa, Silva, 2020. Large Learning Gains in Pockets of Extreme Poverty: Experimental Evidence from Guinea Bissau. National Bureau of Economic Research. NBER Working Paper Number 27799.

Ganimian, Alejandro, Murnane, Richard, 2016. Improving education in developing countries: lessons from rigorous impact evaluations. Rev. Educ. Res. 86 (3), 719–755.

Glewwe, Paul, 2002. Schools and skills in developing countries: education policies and socioeconomic outcomes. J. Econ. Lit. 40 (2), 436–482.

Glewwe, Paul, Muralidharan, Karthik, 2016. Improving education outcomes in developing countries: evidence, knowledge gaps, and policy implications. In: Handbook of the Economics of Education, vol. 5. Elsevier, pp. 653–743.

Heckman, James J., Rodrigo, Pinto, 2015. Econometric mediation analyses: identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. Econom. Rev. 34 (1–2), 6–31.

Imai, Kosuke, Keele, Luke, Tingley, Dustin, Yamamoto, Teppei, 2011. Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. Am. Polit. Sci. Rev. 105 (4), 765–789.

Imai, Kosuke, Keele, Luke, Yamamoto, Teppei, 2010. Identification, inference and sensitivity analysis for causal mediation effects. Stat. Sci. 51–71.

Kraft, Matthew A., Blazar, David, Hogan, Dylan, 2018. The effect of teacher coaching on instruction and achievement: a meta-analysis of the causal evidence. Rev. Educ. Res. 88 (4), 547–588.

Kremer, Michael, Brannen, Conner, Glennerster, Rachel, 2013. The challenge of education and learning in the developing world. Science 340 (6130), 297–300.

Lakshminarayana, Rashmi, Eble, Alex, Bhakta, Preetha, Frost, Chris, Peter, Boone, Elbourne, Diana, Mann, Vera, 2013. "The support to rural India's public education system (STRIPES) trial: a cluster randomised controlled trial of supplementary teaching, learning material and material support. PloS One 8 (7), e65775.

Lucas, Adrienne M., McEwan, Patrick J., Moses, Ngware, Oketch, Moses, 2014. Improving early-grade literacy in east Africa: experimental evidence from Kenya and Uganda. J. Pol. Anal. Manag. 33 (4), 950–976.

Mbiti, Isaac, Muralidharan, Karthik, Romero, Mauricio, Schipper, Youdi, Manda, Constantine, Rajani, Rakesh, 2019. Inputs, incentives, and complementarities in education: experimental evidence from Tanzania*. Q. J. Econ. 134 (3), 1627–1673. https://doi.org/10.1093/qje/qjz010.

McEwan, Patrick J., 2015. Improving learning in primary schools of developing countries: a meta-analysis of randomized experiments. Rev. Educ. Res. 85 (3), 353–394.

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P.J., Elbourne, D., Egger, M., Altman, D.G., 2010. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 340 (c869).

Muralidharan, Karthik, Das, Jishnu, Holla, Alaka, Mohpal, Aakash, 2017. The fiscal cost of weak governance: evidence from teacher absence in India. J. Publ. Econ. 145, 116–135.

Muralidharan, Karthik, Singh, Abhijeet, Ganimian, Alejandro, 2019. Disrupting education? Experimental evidence on technology-led education in India. Am. Econ. Rev. 109 (4), 1426–1460.

Muralidharan, Karthik, Venkatesh, Sundararaman, 2011. Teacher performance pay: experimental evidence from India. J. Polit. Econ. 119 (1), 39–77.

Muralidharan, Karthik, Sundararaman, Venkatesh, 2013. Contract Teachers: Experimental Evidence from India. National Bureau of Economic Research. NBER Working Paper Number 19440.

Pearl, Judea, 2012. "The causal mediation formula—a guide to the assessment of pathways and mechanisms. Prev. Sci. 13 (4), 426–436.

Piper, Benjamin, Sitabkhan, Yasmin, Mejía, Jessica, Betts, Kellie, 2018b. Effectiveness of Teachers' Guides in the Global South: Scripting, Learning Outcomes, and Classroom Utilization. Occasional Paper. RTI Press Publication OP-0053-1805. RTI International.

Piper, Benjamin, Destefano, Joseph, Kinyanjui, Esther M., Ong'ele, Salome, 2018a. Scaling up successfully: lessons from Kenya's tusome national literacy program. J. Educ. Change 19 (3), 293–321.

Piper, Benjamin, Simmons Zuilkowski, Stephanie, Abel, Mugenda, 2014. Improving reading outcomes in Kenya: first-year effects of the PRIMR initiative. Int. J. Educ. Dev. 37, 11–21.

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

Pratham, 2010. Annual Status of Education Report (Rural) 2010. http://www.pratham.org/aser08/ASER_2010_Report.pdf.

Pritchett, Lant, 2013. The Rebirth of Education: Schooling Ain't Learning. CGD Books. http://books.google.com/books?hl=en&lr=&id=PQ72AAAAQBAJ&oi=fnd&pg=PR1&dq=pritchett+schooling+aint+learning&ots=uvSg4RtJhA&sig=1jSzmH3E1acmSrT3eRBDQCyjXwA.

Rajani, Rakesh, 2010. Are Our Children Learning? Annual Learning Assessment Report Tanzania 2010. Twaweza.

Romero, Mauricio, Sandefur, Justin, Sandholtz, Wayne Aaron, 2020. Outsourcing education: experimental evidence from Liberia. Am. Econ. Rev. 110 (2), 364–400.

Salehi, Ahmad Shah, Abdul Tawab, Kawa Saljuqi, Akseer, Nadia, Rao, Krishna, Coe, Kathryn, 2018. Factors influencing performance by contracted non-state providers implementing a basic package of health services in Afghanistan. Int. J. Equity Health 17 (1), 128.

Shalem, Yael, Steinberg, Carola, Koornhof, Hannchen, De Clercq, Francine, 2016. The what and how in scripted lesson plans: the case of the gauteng primary language and mathematics strategy. J. Educ. 66, 13–36.

Sprenger-Charolles, Liliane, 2008. The Gambia : Early Grade Reading Assessment. World Bank Policy Report. World Bank. https://openknowledge.worldbank.org/handle/10986/12972.

The World Bank, 2018a. "Economic data on the Gambia" website. https://data.worldbank.org/country/gambia-the?view=chart. (Accessed 16 December 2018).

The World Bank, 2018b. "Education data on the Gambia" website. http://datatopics.worldbank.org/education/country/gambia,-the. (Accessed 18 December 2018).

Wood, Lesley, Egger, Matthias, Schulz, Kenneth F., Peter, Jüni, Altman, Douglas G., Gluud, Christian, Martin, Richard M., Wood, Anthony JG., Sterne, Jonathan AC., Gluud, Lise Lotte, 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. Br. Med. J. (Clin. Res. Ed.) 336 (7644), 601.