RISE Working Paper 23/142
July 2023

Tackling Implementation Challenges with Information: Experimental Evidence from a School Governance Reform in Tanzania

Jacobus Cilliers and James Habyarimana

Abstract

This paper investigates the role of inter-agency coordination in policy implementation, with a focus on a nationwide roll-out of a new school governance programme in Tanzania. The programme produces a set of school- and teacher-specific diagnostics and recommendations to improve school quality. But information and managerial frictions between the ministry producing the recommendations and the ministry responsible for compliance undermine program fidelity. To address this challenge in a randomly sampled subset of schools, local bureaucrats received text messages informing them of the main recommendations and encouraging them to follow up with schools to ensure compliance. We find that the programme improved student learning and teaching practice, but only when combined with text messages. Observed gains are concentrated in regions exposed to a donor programme that provided these bureaucrats with resources to monitor. Addressing the implementation challenge places the programme in the top five most cost-effective education programmes ever evaluated.

Keywords: education, state capacity, school management, scaling

Tackling Implementation Challenges with Information: Experimental Evidence from a School Governance Reform in Tanzania

Jacobus Cilliers
Georgetown University
ejc93@georgetown.edu

James Habyarimana
Georgetown University
jph35@georgetown.edu

Acknowledgements:

We are grateful for excellent research assistance and program management support from: Shardul Oza, Anthony Mwambanga, Julia Eigner, and Surayya Masood; and feedback from participants at the Research on Improving Systems of Education conference. All errors and omissions are our own.

This is one of a series of working papers from “RISE”—the large-scale education systems research programme supported by funding from the United Kingdom’s Foreign, Commonwealth and Development Office (FCDO), the Australian Government’s Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:
Cilliers, J. and Habyarimana, J. 2023. Tackling Implementation Challenges with Information: Experimental Evidence from a School Governance Reform in Tanzania. RISE Working Paper Series. 23/142.
https://doi.org/10.35489/BSG-RISE-WP_2023/142

The findings, interpretations, and conclusions expressed in RISE Working Papers are entirely those of the author(s) and do not necessarily represent those of the RISE Programme, our funders, or the authors’ respective organisations. Copyright for RISE Working Papers remains with the author(s).

1 Introduction

Well-designed programs with prior evidence of success often fail when implemented by government at scale (Angrist and Meager, 2022; Banerjee et al., 2008; Bold et al., 2018; List, 2022). One potential driver of implementation failure is the behavior of local bureaucrats, including those tasked with "last mile" oversight and implementation. Concerns about implementation have motivated a burgeoning literature on state capacity (Brodkin, 2011; Rasul and Rogger, 2018), with some studies highlighting weak incentives of civil servants (e.g., Deserranno et al. (2020); Gulzar et al. (2017)),[1] and others focusing on efforts to improve their capacity (Azulai et al., 2020; Cilliers et al., 2020a). More recently, there has been an appreciation of bureaucratic overload—a gradual expansion of responsibilities for local bureaucrats, without a commensurate increase in resources (Dasgupta and Kapur, 2020).

This paper focuses on a related, but different source of implementation failure: lack of coordination between and within different ministries responsible for program implementation. Programs that require complementary effort from bureaucrats in different agencies may operate inside the possibility frontier due to information frictions and misaligned incentives (Hill and Lynn, 2003). This may be especially challenging for new programs that lack formal or informal institutional arrangements to share information and coordinate tasks, and before agency-specific performance metrics adjust to the cooperative optimum. In these cases, agencies may pursue independent goals, with tasks allocated to meet narrowly defined goals, and will have weak incentives to re-allocate civil servants' efforts to meet the goals of a joint program. These challenges are not unique to new programs. Coordination problems have been documented in the domains of security (Kean and Hamilton, 2004), immigration (Miles and Cox, 2014), continuity of health care, and child protection (Glisson, 1996), to name a few.

In particular, we evaluate the national roll-out of a school inspection reform in Tanzania that depends on the coordination between two agencies. The School Quality Assurance Division (SQAD), under the Ministry of Education, conducts whole school visits (WSVs) that generate diagnostic information and recommendations for enhancing school quality. This information is then passed on to the President's Office for Regional and Local Government (PO-RALG) responsible for educa-

[1] As noted by Dixit (2002) weak and misaligned incentives are inherent to most public sector organizations, due to the multiplicity of goals, tasks, principals, and tiers of management

tion policy implementation. PO-RALG managers are responsible for coordinating the activities of close-to-school supervisors, Ward Education Officers (WEOs), to ensure follow-up visits to monitor schools' compliance to the recommendations. .

However, information and managerial frictions can impede optimal coordination between these bureaucrats. WEOs do not have direct and timely access to the WSV reports, which per official guidelines are first sent to local government executives and then to the District Education Officer (DEO). Additionally, DEOs may not provide direct school-specific instructions to WEOs, given the cost of processing all of the WSV information, and/or the cost of reallocating WEO tasks to address WSV priorities and away from other agency goals.

To address these bottlenecks, we implemented a low-cost information intervention. Working with the SQAD, we summarized the main recommendations of the WSV reports for each school in our sample and sent them to the relevant WEOs by SMS. The intervention was officially approved by PO-RALG at meetings that brought SQAO officers, DEOs and WEOs together, and where the DEOs publicly endorsed the messages. To this end, messages were signed as coming from the DEO, legitimizing readily accessible, customized, and actionable information for WEOs and empowering them to act on it without additional DEO direction. The text messages thus improve the flow of information between and within agencies and also reduce the costs of adequately responding to it. Our research design is unable to rule out other potential mechanisms of the text messages intervention such as reducing the set of tasks that WEOs should prioritize and/or as regular reminders.

In our study, we used a randomized phased-in approach, selecting a nationally representative sample of about 400 schools from across Tanzania—-one from each Ward. From this group, we randomly chose 198 schools to receive an early Whole School Visit (WSV). To improve the impact of these visits, we additionally sent text summaries of the visit's main recommendations to Ward Education Officers (WEOs) in half of the selected schools. However, because not all schools fully complied with the phased-in design, we focused our final analysis on schools that were assigned an early WSV and had completed their visit by the midpoint of our study.

We produce three sets of primary findings. First, the diagnostic information and ratings produced by the program predict school value added, suggesting program fidelity. However, the WSVs alone (without follow-up prompts) showed no significant impact on overall student learning. Interestingly, when we introduced follow-up prompts via text messages (denoted as Visit&Text), we noted a moderate improvement in learning by about 0.1 standard deviations, equivalent to 0.14 Learning-Adjusted Years of Schooling. These gains are observed in both Kiswahili (0.11 SD, p-value< 0.01) and English (0.16 SD, p-value< 0.05). Given the low costs of collecting and collating the reports and sending the text messages, we estimate that the information intervention is one of the most cost-effective education interventions ever evaluated (Angrist et al., 2020).

Second, teacher effort increased at midline in Visit&Text schools both on the extensive margin (attendance) and on the intensive margin (instructional quality). These teachers, when compared to those in the control group, were approximately 8% more likely to be present in the classroom and demonstrated a higher teaching quality index (0.19 SD). At endline, we observe gains in their lesson planning and overall preparedness (p-value < 0.1). We also find improvements in a non-prespecified outcome of student reports of teaching quality, such as providing feedback on homework and providing remedial instruction (p-value < 0.01). Notably, head teachers do not increase monitoring efforts in either treatment arm.

Third, we discovered a possible behavioral shift among the WEOs who received the text messages. While these close-to-school bureaucrats do not increase the frequency or duration of their monitoring visits, their actions while in school appear distinct. Headteachers in the Visit&Text arm are 9 p.points ($p = 0.156$) more likely to report that WEOs took action to improve learning, and 10 p.points ($p = 0.137$) more likely to state that WEOs organized training for these schools. WEOs are also more likely to follow up on the implementation of the recommendations when visiting schools. Qualitative interviews indicate that WEOs valued the easy access to the information provided through the text messages, and prioritized these recommendations when visiting schools and talking to head teachers and schools.

We exploit two pre-specified margins of heterogeneity to learn more about the likely mechanisms producing the gains observed. First, access to an existing donor program that increased the capacity and coordination of WEOs. Second, a cross-randomized teacher incentive program targeting early grades, to understand the role of teacher motivation in moderating program impact. The text messages intervention increased the frequency and duration of monitoring visits by local bureaucrat in regions where they have the resources to effectively monitor. Consequently, program impacts are uniformly higher (including in Math). This evidence supports the bureaucratic monitoring channel.

On the other hand, exposure to teacher incentives does not increase program impacts.

This paper contributes to three strands of literature. First, it contributes to the literature on state capacity and scaling by identifying and addressing a potential barrier to government implementation capacity: inter-agency coordination problems. It is related to the work by Dasgupta and Kapur (2020) on bureaucratic overload, since the text message intervention alleviated the bureaucrats' workload— they might not have had the resources or time to process and act on the information produced by a new program. Second, it contributes to the literature on improving school management (Anand et al., 2023; Blimpo et al., 2011; Bloom et al., 2015). Muralidharan and Singh (2020) find that a very similar school governance program implemented at scale in India also failed to produce any student learning gains. Our study is designed to address a potential binding constraint mentioned by these authors: weak incentives to implement relevant recommendations. Third, it contributes to the literature on text message behavioral change campaigns. There is a large body of evidence that text messages can change beneficiaries' behavior and improve their health and education outcomes (Head et al., 2013; Lichand and Christen, 2021; Mo et al., 2014). Our program expands on this by targeting bureaucrats tasked to deliver these services. Interventions targeting mid-level bureaucrats can have a wider reach, and thus have the potential to be more cost-effective.

The rest of the paper is organized as follows. Section 2 provides more detail on the context and program while Section 3 describes the study design; Section 4 describes the data and Section 5 lays out our empirical strategy; Section 6 presents the main results while section 7 presents other secondary outcomes. We conclude in section 8.

2 Background and Program Description

2.1 Coordination in Public Service Delivery in Tanzania

The education service delivery system in Tanzania involves coordination between different bureaucracies, primarily the Ministry of Education, Science and Technology (MoEST) and the President's Office, Regional Administration and Local Government (PO-RALG). While MoEST establishes national curriculum and education standards, PO-RALG implements education policy regionally. The country has 185 Local Government Authorities (LGAs), each headed by a District Education

Officer (DEO) who oversees the staff responsible for implementing education policy. This includes the Ward Education Officers (WEOs) who regularly supervise schools and act as a communication link between the government and schools. As of 2017, there were 3,915 wards, each containing around 4-5 primary schools (Table 1). In turn, MoEST enforces education standards through its School Quality Assurance Division (SQAD), which employed 1,374 School Quality Assurance Officers (SQAOs) across all LGAs as of 2017.

However, coordination between MoEST and PO-RALG is hindered by their existing incentives and institutional arrangements. DEOs are required to meet performance targets set by their supervisor, the District Executive Director (DED), while WEOs are expected to execute tasks assigned by DEOs, which can be broad and varied. WEOs are often overburdened and lack resources: for example, 74 percent reported difficulty completing tasks due to different expectations, and 61 percent do not have a fuel budget to visit schools (Table 1). Implementing a new program requires reallocation of WEOs' tasks, and new management and oversight from DEOs—all costly efforts given existing tasks.

Moreover, the policy priorities of MoEST may also conflict with DEOs' performance targets. For example, recent policies prioritize early-grade foundational learning, while DEOs prioritize higher grades, to demonstrate performance in the Primary School Leaving Exam.[2] And as part of a donor-funded payment-for-results scheme, LGAs also receive funding conditional on meeting certain targets. Although this money does not directly flow to employees or managers, it is likely that the DED places pressure on DEOs to meet these targets, which cover a wide range of outcomes.[3]

2.2 The School Quality Assurance Program

A key feature of the reform evaluated in this paper is the creation of the School Quality Assurance Division (SQAD)—a rebranding of the previously punitive school inspectorate system—and the national roll-out of Whole School Visits (WSVs), which replaced the traditional school inspections with a more supportive and feedback-based system.

The WSVs consist of three steps. First, before the visit, head teachers are required to complete

[2] The most salient measure of education performance is the Primary School Leaving Exam. Government publishes rankings of school and LGA performance in this exam on an annual basis (Cilliers et al., 2020b).

[3] Some examples of the DLIs during the time of our study: (i) make available annual school-level EMIS data; (ii) meet annual targets for PTRs; improve Primary and secondary survival rates; (iii) improve girls' transition; (iv) improve their Overall School Quality Score.

a school self-evaluation form (SSEF).[4] Second, a group of three SQAOs visits a school for 2-4 days (depending on the size of the school).[5] During these visits the SQAOs interview school stakeholders (teachers, head teacher, parents, and students), assess students, inspect documents, and observe teaching. They then provide an assessment of school quality along six domains, with a large emphasis placed on teaching and learning.[6] At the end of the visit, there is an "exit meeting" where the SQAOs outline the main strengths of the school, and areas for improvement, and share concrete recommendations for improvement.

Third, after the visit, the lead SQAO writes a short 10-15 page report including their quality rating for each domain, domain-specific recommendations for improvement, and a set of 3-4 main recommendations that should be prioritized. This report is shared with the DED, who is required to pass this on to the DEO. In addition, the SQAD collates some of this information and creates a School Summary Report Card, which is sent back to the schools. The School Summary Report Card includes the overall assessment of school quality, but none of the specific recommendations.

As part of the payment-for-results program, funding was contingent on the SQAD implementing WSVs in every school in the country over a four-year period and conducting follow-up visits in a quarter of these schools.[7] Administrative delays in launching the program meant that over $21,000$ schools had to be visited over a period of three years.[8]

2.3 Coordination challenges in rolling out the SQA program

A potential bottleneck in the successful implementation of the reform is coordination between the SQAOs and WEOs in the preparation for the WSV, attending the exit meeting, and subsequent monitoring to ensure compliance with the recommendations. The head teachers and teachers do not face any explicit incentives to implement the recommendations, and the WEOs—who already visit the few schools in their ward regularly (see Table 1)—are in an ideal position to provide

[4] The form includes basic information such as student and teacher enrollment, but also subjective self-assessments on the various dimensions of school quality that align with the domains examined during the WSV.

[5] Four days for schools with student enrollment over 1,500 and two days for student enrollment below 300.

[6] The six domains are (i) learner achievement; (ii) teaching; (iii) curriculum; (iv) leadership and management; (v) school environment and its impact on welfare, health, and safety; (vi) and community engagement. The first activity SQAOs are required to do when visiting the school is "direct observation of learning and teaching in classrooms and other learning areas" (SQA Handbook, p. 19). The score ranges between 1 (unsatisfactory) and 6 (very good)

[7] 17,438 primary, 4,481 secondary schools, 131 Teacher Colleges

[8] The program was only launched in 2018 and the majority of School Quality Assurance Officers (SQAOs) only received training in July of 2018 (Table 2). All schools had to be visited by July 2021. A major reason for implementation delays was that over this same period, the SQAD planned to construct 50 regional and district offices.

an additional layer of accountability and support to make sure that the recommendations get implemented. Implementation guidelines for the roll-out of the new program were explicit about the important role played by the WEOs. However, despite these programmatic expectations, there is no formal direct flow of information and chain of command between the SQAOs and WEOs (see Figure A.1), and no budget was allocated to train the WEOs in their new role. The program relied on the DEOs to digest all the information from all the WSV reports, and re-assign WEO tasks based on the recommendations—all costly efforts. There is, therefore, a real risk that the WEOs do not receive the report from the WSV, nor receive direct school-specific instructions from the DEO to follow up to make sure that schools are implementing the recommendations.

It is these coordination challenges that our study attempts to address.

2.4 Potential Moderators of Program Impacts

2.4.1 Other Donor Support

There is a high level of donor involvement in the Tanzanian education system. Of relevance to this study is the Education Quality Improvement Plan of Tanzania (Equip-T), which operated in 9 out of the 26 regions in Tanzania over six years (2014-2020), covering 31 percent of schools in the country, with an overall budget of $113 million. One of the components of the program was the strengthening of district planning and budgeting.[9] This included management training for WEOs, introducing a practice of monthly meetings between WEOs, DEOs, and School Quality Assurance Officers, and providing motorbikes and a budget for fuel so that the WEOs could conduct monitoring visits and report to district offices.[10]

Table B.1 shows differences between Equip-T and non-program regions, in terms of student learning, teaching quality, and WEO characteristics and behavior. The most notable differences are at the WEO level: WEOs in Equip-T regions are 42 percentage points more likely to state that they have a sufficient fuel budget to complete all their tasks, and they report having visited more schools in the preceding two weeks (median of 5 vs 4 visits). They also perform different

[9] Other activities include: a) improved access to quality education; b) strengthened school leadership and management; c) stronger community participation and demand for accountability; and d) improved learning and dissemination.

[10] The WEO offices are often far from the district headquarters, and WEOs typically have to submit reports in hard copy.

activities: they are 30 percentage points more likely to meet at least monthly with their DEO and spend more time observing teaching when they visit schools. Notably, they were almost twice as likely to have received training in the new SQA framework by the beginning of 2019. In a previous paper, we found that WEOs in Equip-T regions perform more activities overall, and also perform more activities when visiting schools (Cilliers et al., 2022).

The impact of these governance reforms on coordinating the rollout of new programs is theoretically ambiguous. On the one hand, the monthly meetings should improve the information flow between the different bureaucrats. The WEOs also have more resources to expand activities in response to a new program. On the other hand, the WEOs might have a more rigid set of tasks and activities under closer oversight from the DEOs, resulting in less autonomy to reallocate tasks given new information.

2.4.2 Teacher Incentives

A teacher performance pay program, called KiuFunza, was also implemented in a randomly selected subset of 100 schools from our evaluation sample (Mbiti, Romero and Schipper 2022).[11] In this program, early-grade teachers (grades 1 to 3) were eligible for financial rewards if their students met different performance thresholds in Kiswahili literacy and numeracy. At the end of the academic year, students get assessed in Kiswahili and Mathematics by an independent evaluation team, to calculate payouts.

3 Experimental Design

3.1 Text messages to Ward Education Officers

To address the anticipated inter-agency coordination challenges, we designed and implemented a low-cost information intervention. The Chief District SQAOs were asked to send us completed WSV reports for all the schools in our evaluation sample on an ongoing basis. Upon receiving a WSV, we summarized and condensed the main recommendations, and sent them to the relevant WEO as a short SMS. The messages were unique to each school, but we sent them repeatedly over

[11] The authors excluded an additional 18 schools from their sample used for random assignment, due to perceived challenges in the implementation the teacher incentives program in these schools.

the course of the study period—roughly once every two months. Since WEOs turnover on a regular basis—more than a fifth of the WEOs surveyed at baseline had left their post by midline (Table 1)—we also regularly called the WEOs to confirm receipt and verify that they are still working in the Ward assigned to follow-up.

It was important that the WEOs knew in advance about these messages, and especially that these messages were understood as directives from the DEO. To this end, we held workshops before the start of our study, where the relevant WEOs and DEOs from our evaluation sample were informed about this intervention. In particular, during the workshops, DEOs endorsed the intervention and requested that the WEOs cooperate with us. In this direction, DEOs consented that the text messages were signed as coming from the DEO.

The text messages thus addressed both information and management frictions. First, they improved information flow to WEOs, who might never receive the WSV reports or receive them too late. They also reduced the cost of digesting the information found in a 10-page report. Second, they facilitated a re-allocation of tasks for the WEO, by circumventing the role of DEOs in (i) reading the reports, (ii) redirecting WEO efforts based on the findings of the report, and (iii) following up on WEO efforts related to the findings of the WSV report. An additional potential benefit of the program, which is not directly related to coordination challenges, is improved task prioritization for WEOs who often need to respond to multiple competing demands.

Figure 1(a) shows the distribution recommendations sent, broken down by domain in the WSV framework. Note that each text message included a set of 3-5 recommendations. Over half of the messages included recommendations for extra remedial classes for struggling students. 78 percent of the messages were related to pedagogy—typically recommending that teachers use participatory methods in the classroom and frequently assess students—and 72 percent related to teacher preparation, such as using lesson plans, creating and using learning aids, and signing the class journal. 71 percent of the recommendations also included imperatives that the school leadership should monitor teaching and review teacher records and student assessments. In terms of parental involvement, the most common recommendation sent (57 percent) was to contribute towards school lunch.

3.2 Sampling and random assignment

The program was evaluated using a cluster randomized control trial, at a ward level, with a random phased-in design. First, we performed stratified random sampling to generate an evaluation sample that is (nearly) nationally representative.[12] We randomly selected one region in each of the six zones in mainland Tanzania, and then randomly selected roughly half of the districts in each sampled region, with the probability of being selected proportional to the number of schools in a district. This process yielded a sample of 22 districts and 413 Wards.[13] We took the additional step of excluding 7.6 percent of the primary schools (and their wards) in these districts that had already received WSVs, leaving us with a sample of 397 wards.[14] We then randomly sampled one eligible school from each ward to participate in the study.

Next, we randomly assigned half the Wards in each district—198 in total—to receive a Whole School Visit at some point between April and November 2019. The remaining 199 schools—our control over the study period—were assigned to only receive WSVs after the completion of the planned endline data collection in November 2020. In addition, we randomly assigned half (99) of the early WSV schools to the implementation follow-up arm in which WEOs received text messages with information and prompts to follow up. For the remainder of the paper, the two treatments are referred to as respectively as Visit and Visit&Text.[15]

We shared this random assignment of schools with the SQAD and they agreed to comply with the proposed experimental assignment and timing of WSVs. Compliance with the randomized phase-in design was imperfect. A considerable share of schools in the control group received a WSV earlier than agreed. Most likely, this is because the government placed a high priority on particular schools requiring immediate attention. For related reasons, not all schools assigned to the early phase received a WSV before the midline. In particular, 90 percent of schools in our two

[12] For logistical reasons, the sampling frame excluded the islands of Zanzibar

[13] The sampled regions are: Kigoma, Pwani, Simiyu, Singida, Songwe, and Tanga.

[14] The WSVs were phased in, with some regions starting earlier than others. Out of a population of 1,640 schools in our selected districts, 124 (i.e. 7.6%) were excluded because they had already received the WSV. Table B.2 shows that the excluded schools performed worse on average in the Primary School Leaving Exams (PSLE) over the 2013-2016 period, relative to other schools in these districts and relative to our selected sample of schools. The worst performing schools were very likely visited first because they were prioritized for improvement.

[15] The Kiufunza teacher incentives program was randomly assigned within in a sub-set of 379 schools in our evaluation sample, stratifying by district and assignment to the Visit and Visit&Text arms. Before random assignment, the research team excluded one hard-to-reach district and an additional 10 private schools. This resulted in 100 treatment and 279 control schools.

treatment arms and 22 percent of schools in the control received a WSV by the start of midline data collection; and virtually all schools in our treatment groups and 43 percent of schools in our control group had received a WSV by endline (Panel A, Table 2) As a result of the staggered roll-out of the WSVs, there is variation in elapsed time between the date of the WSV and date of data collection, as shown in Figure A.2. The average duration of this gap is 270 days in our treatment schools at midline, with a mode of 250 days. At endline, the modal gap duration is just under 600 days.

4 Data

4.1 Administrative data

We have administrative data on the timing for all WSVs in our study sample, and the overall school quality scores. In addition to our collection of WSV reports in the treatment arms, government shared data on each WSV that was conducted by June 2021.[16] This data includes the exact date of each WSV, a weighted average of all the domain scores (with a higher weight given to teaching, learning, and leadership), and also the domain score for overall school quality, ranging between one and six. We were able to match this data to our sample, using unique school identifiers.

4.2 Primary data collection

We collected baseline data in each of the 397 schools in our sample in February/March 2019, and revisited these schools twice over two years. Midline data collection took place roughly one year after baseline, in Feb/March 2020. Our midline data collection was cut short due to school closures in the wake of Covid-19. As a result, we are missing head-teacher data from six schools, and classroom observation data from 54 schools at midline.[17] Endline data collection was complicated by the Covid-19 pandemic and presidential elections in Tanzania. In order to meet timing constraints, we implemented endline data collection over two rounds: 200 schools were visited in

[16] This data was validated by an external auditing firm that recorded all the WSVs that took place by end of June in 2019, 2020, and 2021—in lieu of conditional donor payments to government linked to the number of WSVs conducted.

[17] Two schools could not be reached because of floods, in another two schools the data was lost due to a car accident, and in another two schools, the head teacher was either not available or declined to consent to the survey. The data collection team had planned to return to these schools to conduct the head teacher survey, but this was curtailed by school closures.

November/December 2020, and the remaining 197 were visited between January and March 2021.[18] During all these school visits we conducted student assessments, classroom observations, facility inspections, and document inspections, and also surveyed WEOs, head teachers, and teachers.

We conducted curriculum-referenced **student learning assessments** on a randomly selected sample of 10 standard two and 10 standard three students at baseline. We have a total baseline sample of $6,991$ students from 393 schools for whom we were able to receive parental consent to administer assessments throughout the whole study.[19] Grade two students were assessed in Math and Kiswahili at baseline, whereas the Grade three students were assessed in Math, Kiswahili, and English (under the new curriculum English is only taught starting in grade 3). We were able to assess 94 and 92 percent of the original baseline sample at midline and endline, respectively (see Table B.3 for more details). During assessments, we also conducted a brief student survey and counted the number of pages completed in the sampled students' exercise books the previous week. The student survey asked some questions related to the pedagogical practices of their teacher. In particular, whether the teacher motivates them to work hard, gives practice exercises, explains things clearly, assigns homework, and provides written feedback on homework.

For the **teacher surveys**, we sampled all standard two and three teachers teaching the focal subjects of Kiswahili, Math, and English, and then randomly sampled additional teachers until we reached a total of ten teachers per school.[20] In addition to basic demographic questions, we also asked teachers about their beliefs of student ability and the extent of monitoring and curriculum guidance received by the school leadership. We aimed to survey the same teachers at midline and endline, and randomly sampled replacements using the same protocol if teachers were absent or had left/been transferred. We conducted **classroom observations** on two randomly selected teachers per school, teaching any of the focal subjects in grades 2 and 3, using the World Bank's *Teach* instrument (Molina et al., 2018). We observed the same teachers at midline and enldine, and replaced sampled teachers with another teacher of the same grade if they were not available or no longer teaching the same grade.

The **headteacher survey** elicited basic demographic information and key school character-

[18] We note that the possibility of differential fadeout patterns across interventions could affect the comparability of learning outcomes across periods. The magnitude and significance of our results minimize these concerns

[19] Baseline student assessment data is missing from four schools due to technical errors in data capture.

[20] If more than 10 teachers were teaching the focal subjects in standards two and three, we randomly selected 10 of those teachers

istics. In addition, we asked detailed questions about their experience of the Whole School Visit (for those that had received one) and the extent and nature of interactions with the Ward Education Officers. We also captured information about their beliefs of school quality at the start of 2019 —i.e. before most schools had received a WSV— to assess if the information generated by the WSVs shifted beliefs. For this purpose, we asked the head teachers to indicate on a scale between 1 and 4 the "room for improvement" in the school on a range of different school inputs: school leadership, teaching, school environment, and community involvement. In addition, we asked a series of vignettes to capture their beliefs about the relative importance of different inputs in the education production function. In each question there was a trade-off between two different inputs —prioritizing student performance in early v later grades, teacher training vs infrastructure, ensuring mastery of foundational skills vs completing the curriculum, and (potentially disruptive) participatory vs traditional teaching methods. We interpret head teachers' responses to these choices as reflecting their value judgment of the relative importance of the different inputs. The fieldworkers also conducted **facility inspections**, capturing measures such as the number of functional classrooms and clean toilets, and also the proportion of classrooms with students that have a teacher in them. Finally, we also **surveyed the WEO** in each Ward during each round of data collection. The survey included questions about the frequency and length of school visits, the activities they typically perform when visiting schools, their exposure to the new WSV framework, recollection of the main recommendations that were made in the WSV report, and the actions that they took in response to the report.

To minimize the risk of over-rejection of the null hypothesis due to multiple comparisons, we created indices of the main outcomes, by taking the mean of the standardized score of all the indicators relating to the same outcome (Kling et al., 2007). For some families of outcomes such as teaching practices, we are unable to construct a mean index, since outcomes within the family have different levels of observation (school, teacher, and student). In these cases, we also report the sharpened q-values that control for false discovery rate, using the two-stage procedure developed by Benjamini et al. (2006) and applied by Anderson (2008). In each case the p-values are grouped across all the dependent variables shown in a table, but with separate groupings for each treatment coefficient and round of data collection.[21]

[21] Our motivation is that each treatment coefficient tests a distinct hypothesis.

We specified all of the hypotheses and indicators that related to each outcome in a pre-analysis plan registered with the American Economic Association in April 2020, before data analysis.[22]

5 Empirical strategy

Our empirical strategy is designed to address non-compliance within a randomized phase-in research design (see section 3.2). We estimate the Local Average Treatment Effects (LATE) for both midline and endline outcomes. For all student and teacher outcomes observed at midline, the main estimating equation is:

$$y_{i,s} = \beta_0^F + \beta_1^F(\text{Visit}_s^D) + \beta_2^F(\text{Visit\&Text}_s^D) + \alpha_d + X'_{i,s}B + \epsilon_{i,s}, \tag{1}$$

where $y_{i,s}$ is the relevant outcome variable for individual i in school, s; α_d refers to strata fixed effects;[23] and $X_{i,s}$ is a vector of baseline controls included to improve precision. $(\text{Visit})_s^D$ and $(\text{Visit\&Text})_s^D$ are dummy variables indicating whether, by the time of midline data collection: (i) school s received a WSV (but no text); (ii) school s received a WSV *and* their WEO received a text message. To address the endogeneity of receiving a Whole School Visit, we instrument Visit receipt with an a set of indicators corresponding to randomized assignment—$(\text{Visit})_s^Z$ and $(\text{Visit\&Text})_s^Z$.[24] The coefficient estimate, $\hat{\beta_1^F}$, can be interpreted as the causal impact of receiving a WSV; and $\hat{\beta_2^F}$ can be interpreted as the causal impact of both a school receiving a WSV and their WEO receiving the text message.[25]

For most cases, the outcome-specific control variables are selected using the least absolute shrinkage and selection operator (LASSO), regressing the outcome variable on the full set of possible baseline control variables, after partialling out strata fixed effects. We implement this procedure separately for each outcome variable and round of data collection. The one exception is student learning, where we include the same set of control variables for each measure of student learning,

[22] AEARCTR-0005714, https://www.socialscienceregistry.org/trials/5714

[23] We stratified by both district and assignment to the teacher incentives program.

[24] We include the same vector of exogenous control variables (i.e., $X_{i,s}$ and strata fixed effects) as above.

[25] Note that in our case the LATE is not the same as the Treatment Effect on the Treated, since some schools in the control group also received a WSV. The treatment effect on these schools could plausibly have been larger, because these schools were prioritized as challenging schools that need immediate attention.

for comparability purposes. In these cases, we control separately for baseline learning in each of the three subjects.

Next, to estimate program impacts on WEO's beliefs and behavior at midline and for all endline outcomes, we restrict the sample to the 90 percent of schools in the two treatment arms (89 out of 99 schools) that had received a WSV by the time of midline data collection, and estimate the following:

$$y_{i,s} = \beta_0^R + \beta_1^R(\text{Visit\&Text}_s^D) + \alpha_d + X'_{i,s}B + \epsilon_{i,s}, \tag{2}$$

Restricting the sample in this way allows us to focus on the effects of the phone based encouragement of WEOs, which is only relevant for schools that have received a Visit. For endline outcomes, we similarly restrict the sample because comparison with the control group is complicated due to (i) differential timing in treatment uptake between the treatment and control groups—roughly a fifth of control schools received a WSV between midline and endline; and, (ii) potential dynamic treatment effects for both treatment and control schools that receive a WSV after midline would confound our estimates. Using the full sample, the coefficients $\hat{\beta_1^F}$ and $\hat{\beta_2^F}$ are therefore hard to interpret at endline.[26] The coefficient estimate, $\hat{\beta_1^R}$, can be interpreted as the *additional* effect of sending a text message to a WEO, for schools that had received a WSV by midline. Since not all WEOs serving these schools had received any text message by midline, $\text{Visit\&Text}_s^D$ is also potentially endogenous in specifications for midline WEO outcomes. As such, we instrument for $\text{Visit\&Text}_s^D$ with $\text{Visit\&Text}_s^Z$, including the same set of exogenous control variables as above.[27]

All of the analytical approaches and sample choices discussed above were pre-specified in our pre-analysis plan, although the specific number of schools in the reduced sample (for equation 2) was updated once we had information on all the schools that had received a WSV by the time of midline data collection.[28]

[26] The interested reader is referred to the appendix to see endline results using the full sample and estimated using Equation 1

[27] Note that for all endline outcomes, we have 99 percent compliance to the text messaging intervention, within the restricted sample. Only in one out of the 89 schools in the Visit&Text arm did a WEO not receive a text message by the time of endline data collection.

[28] We pre-specified that we will restrict the sample to 168 schools that had received a WSV by the end of 2019, according to our records at the time. But this sample was incomplete. We decided to expand this sample to 176 schools that had received a WSV by the start of midline data collection, to improve statistical power.

Finally, when examining heterogeneous treatment effects, we restrict the sample to the 178 schools that were visited by midline and estimate the following equation:

$$y_{i,s} = \gamma_0^R + \gamma_1^R(\text{Visit\&Text}_s^Z) + \gamma_2^R(\text{Visit\&Text}_s^Z \times G) + \gamma_3^R G + \alpha_d + X_{i,s}'\Gamma + \epsilon_{i,s}, \tag{3}$$

where G is a dummy variable indicating the pre-specified sub-group of interest.

5.1 Balance and attrition

Table B.4 shows that the sample is balanced across a range of WEO, teacher, head teacher, school, and student characteristics. Moreover, Tables B.5 and B.6 show that the sample remains balanced on the restricted samples of observations collected at midline and endline, respectively. In Table B.6 the sample is restricted to the sub-set of schools in the two treatment arms where a WSV had been conducted by the time of midline data collection.[29] Showing study balance is important since the main endline analysis will be performed on this sample.

Table B.3 reports determinants of student attrition at endline. The multi-year attrition rates are very low: 5.7 and 7.7 percent at midline and endline, respectively. Columns (1) and (4) report results of regressing attrition status at midline and endline, respectively, on indicators for treatment assignment, including strata fixed effects. We find no statistically distinguishable difference in attrition rates across the evaluation arms. In the remainder of the columns, we regress baseline characteristics on treatment assignment, attrition, as well as interaction terms between treatment and attrition. The coefficients on "Attrite" in columns (2) and (5) show that control group students who perform worse at baseline were more likely to attrite. It is not surprising that the weakest-performing students are more likely to drop out of school or be absent. But more importantly, the interaction terms show that attriters across evaluation arms do not differ in terms of baseline learning. In other words: it is not the case that worse- or better-performing students attrite in the treatment groups. The attriters in the Visit&Text arm are slightly older (0.37 years) than the attriters in the control, but the coefficients on the treatment dummies show that the sample remains balanced in terms of baseline characteristics, when restricting the sample to non-attriters. Taken together, the combination of low attrition rates, balanced attrition, and baseline balance for

[29] Baseline data on student learning is missing for two schools, one in each treatment arm.

the sample of non-attriters suggest that attrition is unlikely to bias results in this study.

Tables B.7 and B.8 repeat the attrition analysis for teacher-level data. Although teacher-level attrition is higher—23.3 and 25.5 percent in the two follow-up rounds of data collection—it is balanced and not correlated with observed teacher characteristics. Attrition levels for classroom observations are higher—56 and 63 percent at midline and endline, respectively—for a combination of reasons including the high teacher attrition reported above, grade reassignment across years,[30] premature cessation of midline data collection due to Covid, and teacher absenteeism. But the attrition rates are balanced. Given this high level of attrition, our main analysis for classroom observations includes the replacement teachers, although we also show results in the appendix for the sample of non-attriters.

6 Results

6.1 Quality of implementation and coordination

Our assessment of the current implementation of the WSVs is very encouraging, especially concerning the activities performed at the school, although coordination with WEOs was weak. Table 2, Panel A, shows that almost all (98 percent) of the surveyed SQAOs indicated that they had received training in the new framework by the time of the midline survey. The training typically lasted five days, and at least one training was provided by master trainers organized by the central division of the Directorate for SQA. This is encouraging, since it means that they mostly did not follow the typically low-quality cascade model of training of the trainers. However, the SQAOs still did not believe that the length of training was sufficient.[31]

Moreover, Figure 1(b) indicates that the DSQAOs mostly performed the appropriate activities, as specified in the training manual, when conducting the WSVs.[32] In almost all the WSVs, SQAOs observed teaching in the classroom (94 and 96 percent during midline and endline, respectively). In a large proportion of these visits, the SQAOs also talked to parents (74 and 77 percent) and

[30] Our sampling strategy prioritized observing grade 2 and 3 teachers, so we sampled a replacement even if the original teacher was still teaching in the same school, but teaching in a different grade.

[31] Figure A.3 shows that the majority received their training over July and August 2018, which was after the official start of the program, but roughly six months before the start of our study.

[32] Data is restricted to schools where the WSV took place, according to both the head teacher and our administrative records, before each survey round.

assessed students (88 and 79 percent). This is in contrast to the ‘old’ model of school inspections that typically did not involve talking to parents, assessing students, or observing teaching. Indeed, the head teachers who reported having received a WSV before July 2018—when few SQAOs had received training—were far less likely to indicate that the SQAO observed teaching, talked to students or parents, or assessed students.

Furthermore, the information produced from the WSVs was informative of school quality. Table B.9 shows that the overall school quality score, as was collected and collated by government, is positively correlated with student-value added at midline—one SD higher score is associated with 0.06 SD improvement in student value-added. I.e., students in schools with a higher quality score learned at a faster rate, compared to schools with a lower score. It is also encouraging that almost all head teachers reported that they learned something new as a result of the WSV.

However, coordination between the two ministries was weak, as shown in Panel C in Table 2. First, there was no systematic involvement of the WEOs at the inception of the program. By midline data collection, only half reported to have received any kind of training in the new framework. The trainings were conducted at a decentralized level and therefore varied substantially by region. For example, in one region only 14 percent of WEOs reported to have received training (Songwe), compared to 87 percent in another region (Simiyu). Second, the information contained in the WSVs rarely reached the WEOs. While WEOs seem to be aware that a WSV had taken place—within the restricted sample where a WSV had taken place by the time of the midline survey, over 80 percent correctly stated that a WSV had been conducted—only about a fifth can show a copy of the report. This means that the information about school improvement needs is not readily available to WEOs. Third, Figure 1(b) shows that only half of the head teachers reported having completed the School Self-Evaluation Form (SSEF) in advance of the WSV (the WEO is responsible for distributing this document), and WEO attendance at the exit meetings was low. It is thus likely that the majority of the WEOs were uninformed about the SQAO assessments and especially, the recommendations.

Turning to the text messages intervention, panel B in Table 2 indicates the potential limited reach of our intervention to address bottlenecks in the flow of WSV information. According to our communication records, 93 percent of WEOs in schools where a WSV had already taken place by baseline data collection, had received at least one text message by the time of midline data collection.[33] But only 60 percent of the WEOs in the 89 schools where a WSV had taken place indicated that they had received a text message or call reminding them to monitor specific WSV recommendations. This discrepancy is likely due to a combination of recall bias and turnover in WEO appointments (see Table 1). It may also reflect the (limited) effectiveness of the communication modality given the SMS spam that many WEOs are exposed to.

Despite this limited reach, we find evidence of a small shift in WEOs' recollection of the recommendations, as evidenced in panel D of Table 2. Teachers in the Visit&Text arm recall 20 percent more recommendations compared to teachers in the Visit arm (7.1 vs 5.9) and recall twice as many recommendations related to student learning. Figure 2 shows that the largest differences are for holding extra classes for struggling students (17 p.point difference), provision of school lunch (8 p.points), and extra training for teachers (6 p.points).

6.2 Student Learning

Table 3, panel B, shows that the sending of text messages improved learning—as measured by the average performance in Kiswahili, Math, and English—by 0.11 standard deviations (SDs) by endline, relative to schools that received a WSV but no text message. This is larger than the median effect size 0.1 SDs found in international education studies (Evans and Yuan, 2022), and equates to 14 percent of a Learning-Adjusted Year of Schooling (LAYS).[34] These gains are mostly driven by improvements in Kiswahili and English (columns (3) and (4)). The ITT results (column 5) are equivalent, because of near-perfect compliance in the restricted sample. There is a similar pattern of results at midline (Panel A), with small, but statistically insignificant improvements in Kiswahili and English in the Visit&Text arm.

For completeness, Table B.10 shows endline ITT results for the full sample. Students in schools randomly assigned to the Visit&Text arm learned 0.08SDs more, relative to students in schools assigned to the control. We can reject the null of equality between the two treatments for the overall score, Kiswahili, and English. Note that this comparison is confounded by non-compliance in the control group before the midline and the phase-in of treatment at the beginning of 2021.

[33] Since the WSVs were implemented on a staggered basis, not all WEOs received the same number of text messages: by the midline, 74 received three messages, 5 received two messages, and 11 received one message.

[34] Following the approach used by Angrist et al. (2020), we divide the effect size of the high-performing benchmark of 0.8 SDs of learning in a year.

6.3 Teaching Practices

A key objective of the SQA program is to improve the quality of teacher instruction and behavior. Overall, the results suggest that the program had modest short-term positive effects on some components of teaching practices, but only when the WEOs also received the text messages.

Table 4 panel A indicates improvements across different dimensions of teacher behavior in the Visit&Text arm at midline. First, teacher attendance in the WSV&Text arm improved by a statistically significant 10 percentage points, or 22 percent, relative to schools that did not receive a WSV (sharpened q-value of 0.12). Second, teaching practices, as measured by the Teach observation tool improved by 0.27 SD (q-value= 0.07). Table B.11 shows that this improvement is mostly driven by an increase—of 11 percentage points, or 19 percent—in the proportion of time that almost all students were on task during the observed lesson. Classroom culture and the quality of instruction also improved, by 0.11-0.14 SDs.[35] Third, teacher preparation for the classroom improved by 0.16SD, significant at the 10 percent level (q-value= 0.12).[36] There is no evidence that the programs increased the frequency of assessment or the likelihood of assigning homework, although table B.12 shows that students in the Visit&Text arm completed more exercises in Kiswahili, relative to the control.

In contrast, the midline effect sizes for the Visit arm are smaller and not statistically significant, although there is suggestive evidence of midline improvements in teaching quality, as measured in the classroom observations (0.15SDs), and teacher preparation (0.14SDs). As a result of these modest gains in the Visit arm, we cannot reject the null of equality between the two treatment arms after controlling for the false discovery rate.

Panel B shows that almost none of the gains reported in Panel A persisted to endline in the restricted sample of intervention schools that had received a WSV by midline. There are no observed improvements in teacher attendance, teaching practices, assessment, or assignment of homework. Two exceptions are the teacher preparation index—0.15SD (q-value=0.16)—and students' *own* reporting on the pedagogical practices of their teacher (column (6)), which is 0.12 SD higher. Table B.13 breaks this index down into seven teacher activities or behavior: students were more

[35] Table B.11 also shows that the effects sizes are larger still (0.37SD for the overall score) when excluding replacement teachers from the sample.

[36] This outcome is the average of four indicator variables: (i) observed and updated lesson plans; (ii) differentiated subject lesson plans; (iii) observed updated scheme of work; and, (iv) can show a class journal.

likely to report that their teacher reviews and discusses what they learned (5 p.points), assigns and provides feedback on their homework (6 and 10 p.points), and provides extra help to students who are struggling (10 p.points, or 28 percent). However, since we did not pre-specify this outcome, we treat this evidence as suggestive.

6.4 School management and community engagement

In addition to student learning and teaching practices, we also measured the quality of management of the school leadership, parental contributions, and the quality of the school environment. These were all domains in the WSV report, so it is possible the behavior of these stakeholders also changed. We also measured head teachers' beliefs, since the new information produced by the WSV could have shifted their beliefs. As reported in Figure 1, head teachers indicate in the surveys that they felt like they learned something new. Results on head teachers' beliefs and management practices are found in Tables 5 and 6, respectively.

First, we find no evidence that head teachers updated their beliefs about the (pre-treatment) quality of their school, as a result of receiving the WSV (Table 5, column (1)).[37] But head teachers in the Visit&Text arm revised upwards their beliefs about the numeracy and literacy skills of grade 2 students in their school, especially so at endline (column (2)). This is consistent with the observed improvements in student learning at endline. The final four columns in Table 5 show that head teachers' beliefs over the education production function did not change at midline: they are no more or less likely to prioritize: (i) early vs later grades; (ii) curriculum coverage vs student learning; or (iii) participatory vs traditional methods of teaching, relative to the control. Beliefs over the most important inputs into improving student learning are harder to shift, especially in the short run. However, at endline, head teachers in the Visit&Text arm are 9 percentage points (or 11 percent) more likely to indicate that they would allocate a new teacher to early grades (grades 1-3), rather than higher grades. This is consistent with the fact that many of the recommendations from the WSV report, including those sent to the WEOs, asked the leadership to prioritize basic literacy and numeracy skills.

Next, Table 6 shows that at midline schools in the Visit&Text arm were more likely to have an

[37] The dependent variable in column (1) of Table 5 is a Kling index of 11 different indicators of school quality. See section 4 for the question wording. See Table B.14 for results on each of the constituent indicators.

up-to-date Whole School Development Plan (it was a common recommendation to create a new WSDP), but there is no strong evidence of a change in the nature of monitoring or curriculum guidance provided by the head teachers, as reported by the teachers.[38] This is surprising, since so many recommendations focused on the school leadership, especially in monitoring teachers and providing curriculum guidance. We note that the indicators for teacher preparation that improved at endline, such as the use of lesson plans and teaching aids, were typically recommendations targeted at both teachers and school leadership, so those improvements can also be partly attributed to the head teacher. There is also limited evidence of a change in parental engagement (Table B.17), except for the contribution of school lunches. In the Visit arm, there is a 14 p.point (41 percent) increase in the proportion of schools where parents contribute to a school lunch. This was a common recommendation, but parents faced substantial constraints in providing resources.[39] Finally, there are no improvements in the school environment (Table B.18). This is perhaps unsurprising since the recommendations tasked the parents and broader school community to contribute towards infrastructure investments, who have limited resources themselves.[40]

6.5 Behavior of the WEOs

Lastly, we examine the behavior of the WEOs at midline. All analyses reported here compare the Visit with the Visit&Text arm only.[41] Overall, there is no evidence that addressing coordination challenges induced changes in the frequency of monitoring and interacting with the schools. However, there is suggestive evidence that WEOs in the Visit&Text arm changed their behavior along some dimensions.

First, Panel A in Table 7 shows that there is no strong evidence that the text messages induced WEOs to engage more frequently or intensively with study schools. The (self-reported) frequency of visits to program schools increased by 8 percent, and the length of visits increased by 10 percent. But these effects are not statistically significant. Importantly, we find no evidence that WEOs

[38] For completeness, Tables B.15 and B.16 show results for all the constituent indicators for monitoring and curriculum guidance, respectively. Interestingly, teachers in both treatment arms are *less* likely to report that they receive high-quality curriculum guidance feedback from the school leadership. Perhaps their expectations shifted in response to the recommendations.

[39] One WEO noted that the parents could not provide food because the harvest was bad.

[40] A common complaint from the WEOs we interviewed was insufficient resources, such as desks, classrooms, and textbooks, and too few teachers.

[41] As noted in Section 5, data is further restricted to 178 observations where a WSV was conducted by the time of midline data collection.

interact less frequently with non-study schools in the same Ward. This is encouraging from an identification perspective, since more effort exerted in program schools could have led to negative spillovers for non-study schools in the same Ward. It is important to note that the WEOs are already visiting schools regularly, roughly once every two weeks, so there might not be a lot of room for improvement. Moreover, our results in Column 5 suggest that WEOs that receive messages do not report conducting more activities when they visit schools.

Although we do not observe overall changes in effort levels, a deeper dive into the specific actions that WEOs report when visiting schools reveals some substantive differences between Visit and Visit&Text arms. The results of this analysis are shown in Figure 3. WEOs in the Visit&Text arm are 16 percentage points more likely to indicate that they check that schools are implementing a Whole School Development Plan (WSDP). This was a common recommendation from the WSV. They are also slightly more likely to report that they provide feedback to teachers (7 p.points. or 18 percent), talk to teachers (6 p.points, or 10 percent), and observe teaching (8 p.points, or 16 percent) when visiting schools, although these differences are not statistically significant.

Next, Panel B in Table 7 provides suggestive evidence that the WEOs took actions to improve learning. To overcome social desirability bias, the outcomes in this panel (except for column 5) are drawn from head teacher reports of WEO behavior. Head teachers in Visit&Text schools were 6 p.points more likely to report that a WEO followed up to make sure that the recommendations are implemented, and 9 percentage points more likely to believe that the WEO took actions that improved student learning. They are also 8 p.points (31 percent) more likely to report that the WEOs organized a teacher training workshop.[42] The observed gains are consistent with the content of the messages which focused on recommendations to improve instruction and learning. Individually these effects are large in magnitude albeit not statistically significant. In aggregate, there is an improvement in the mean index of 0.24 SDs—albeit significant only at the 10 percent level.

[42] There is consistent evidence across different data sources that WEOs were more likely to organize training for teachers. WEOs receiving text messages are 7 p.points, or 150 percent, more likely to self-report that they organized training for teachers (p=0.068). Similarly, teachers in intervention schools were 4.4 percentage points (50 percent) more likely to state that they participated in training away from school organized by a WEO. However, we did not pre-specify these analyses.

6.6 Heterogeneous Impacts: Donor Support and Teacher Incentives

One way to learn about the potential mechanisms underlying observed effects is to examine treatment effects heterogeneity. Our pre-analysis plan posits two moderators of treatment effects. First, the program may work better in environments where WEOs have more resources. Second, a teacher exposed to teacher incentives may be more likely to take up the advice and opportunities generated by the WSV. We exploit the fact that a donor-funded Equip-T program operated in three out of the six regions in our sample; and a randomly assigned teacher incentives program, called *Kiufunza*, was orthogonally implemented in 100 study schools. Tables 8 and 9 report the results of this analysis. Across both tables, we examine learning and teaching outcomes at *endline* only. In addition, the sample is restricted to schools assigned to the WSV that received a visit by midline. Finally, all results shown are intent-to-treat and reflect the impact of being assigned to the Visit&Text arm.

Panel A of Table 8 shows the moderating effects of the Equip-T program. Column (1) shows that aggregate learning gains in the Visit&Text arm are 0.21 SD larger in Equip-T regions, relative to non-Equip-T regions. We observe uniformly substantive gains across all subjects including Mathematics. The results suggest that *all* improvements in learning in the Visit&Text arm are exclusively situated in the Equip-T regions. The fact that Equip-T schools receiving the WSV but *no* WEO messages perform worse than non-Equip-T schools is further evidence that motivating and coordinating WEO action is the likely source of the gains. We can rule out that key inputs of the Equip-T program such as facilitation of WEO monitoring, coordination of SQAOs and WEOs, or in-service teacher training moderate WSV effects in the absence of WEO messages.

Panel B of Table 8 shows the moderating effects of the teacher incentives program. We observe no positive complementarities between the Visit&Text arm and teacher incentives. If anything, our point estimate of the interaction in Column (1) is negative albeit imprecisely estimated: the effect size of the Visit&Text arm on aggregate learning is 0.06 standard deviations *lower* in the schools where teachers also received financial incentives. Note that the incentives program targeted students in grades 1-3 and that at endline most students are in grades 4 and 5. However, these students were exposed to the incentive program throughout 2019 and 2020.

Turning to teacher behavior, Panel A in Table 9 provides very suggestive evidence of positive interactions between Equip-T and Visit&Text at endline. While there are no measurable differences in teacher attendance, the effect of sending a text message on the overall teaching quality index, as measured in the classroom observations, is 0.2 standard deviations larger in the Equip-T regions. While this difference is not statistically significant, it is mostly due to improvements in the quality of instruction (see Table B.19). Similarly, the impact of sending a text message on the teacher preparation index is 0.42 SD larger in the Equip-T regions, and the impact on students' perception of teaching quality is 0.1 SD larger. For these outcomes, we can reject the null that sending text messages had no impact in Equip-T regions.

Panel B in Table 9 shows weak and inconclusive evidence of interactions between the incentives program and the Visit&Text arm. Given the negative interaction results with respect to learning outcomes, it is surprising to observe some positive estimates of the interaction for classroom observations and assessment. In none of these outcomes can we reject the null that teaching outcomes in Incentives schools assigned to the Text arm are no different from non-Incentives schools assigned to the Visit only arm.

Finally, in Table 10 we turn to WEO-level outcomes, focusing exclusively on interactions with Equip-T status, since the program provided more resources/training to WEOs and changed their governance structure. In Panel A we examine interacted impacts on WEO reported outcomes. We find that the impact of sending the text messages on the number of times that a WEO visits the target school is substantially higher in the Equip-T regions. Within Equip-T schools text messages increase the median number of visits by a third, from six to eight. Interestingly, there was a commensurate increase in visits to non-study schools in the same Ward although with possibly shorter visit duration. It is possible that the increased resources available for fuel in the Equip-T regions enabled the WEOs to respond to the text messages by increasing monitoring across all of their schools. However, we don't observe any systematically higher intensity of monitoring activity in Equip-T schools.

Turning to their behavior in response to the WSV, there is suggestive evidence of positive complementarities, especially in the outcomes reported by the head teachers in Panel B of Table 10. In Equip-T regions, the head teachers are 14 p.points more likely to state that the WEO followed up on the implementation of the recommendations, and 4.6 p.points more likely to state that the WEO took actions to improve learning, and 5.1 p. points more likely to have organized a teacher training workshop. All of these coefficients are imprecisely estimated, so we only treat them as suggestive. As further evidence (not pre-specified), Panel A in Table B.22 shows that the text messages shifted the activities performed by Equip-T WEOs when visiting schools: they were more likely to interact with teachers (talking to them, observing their teaching, and providing them with feedback) and parents, but less likely to inspect documents.

The fact that the intervention was effective in Equip-T regions suggests that the improved governance structures set in place by the program—in particular, the monthly meetings between DEOs, WEOs, and SQAOs—were insufficient to overcome the coordination challenges. Perhaps these meetings did not discuss school-specific recommendations, or re-allocate WEO tasks based on the information. Evidence partially supporting this conjecture is presented in Table B.22, Panel B, Column (2). without the intervention of text messages, WEOs in Equip-T regions were more inclined to discuss the WSV reports with DEOs compared to their counterparts in other regions. Interestingly, the introduction of text messages reversed this trend. WEOs in Equip-T regions became less likely to discuss the WSV reports with DEOs when text messages were introduced. The text messages thus seemed to replace the existing communication channels set in place by Equip-T.

6.7 Qualitative interviews with local government officials

To augment our empirical findings we conducted a series of focus group discussions with the relevant local government officers—WEOs, SQAOs, DEOs, and REOs—to better understand the reasons for the success of the text messaging arm. A one-day workshop was held in each of the six regions where the study was conducted. We also conducted a phone survey of all the WEOs who participated in the program to get a better understanding of their actions in response to receiving the text messages. The focus group discussions and phone surveys highlighted three potential mechanisms related to addressing implementation challenges. The text messages produced: (i) faster access to information; (ii) reduced cost of digesting the information and translating them into actionable tasks; and (iii) induced a re-allocation of efforts away from the existing tasks required by their superiors, towards those emphasized in the WSVs.

First, the information included in the text messages reached the WEOs much earlier than the WSV report. WEOs were thus able to act on more focused recommendations shortly after the WSV. According to one WEO: “Text messages were a good reminder and they arrived on time, normal reports arrive very late and end up in lockers and sometimes the school gets the report before the WEOs". It was also convenient that this information was stored on their phones, so they could forward it to the relevant head teacher or refer to it when they visited the school. In contrast, the majority of WEOs indicated that DEOs typically do not discuss the results from the WSV reports with them. Second, WEOs reported that it was easier to act on a smaller set of recommendations, compared to the effort of extracting domain-specific and overall recommendations from the full report, which was typically 10 pages in length.[43] Third, the text messages induced the WEOs to prioritize these recommendations during their school visits.[44]

7 Cost-effectiveness

Table 11 provides a breakdown of costs and the overall cost-effectiveness of the text message intervention. All of our estimates are relative to the business-as-usual implementation of the WSV program, so we do not include the cost of implementing the Whole School Visit in our calculations. We also do not include the salaries of bureaucrats, since the text messages did not require them to perform any tasks beyond their current job description. The total cost of the text messages intervention was roughly $\$24,000$ (in 2019 prices), of which the largest cost driver was organizing the regional workshops ($\$17,000$), followed by paying a stipend for someone to manage the whole process ($\$6,000$). The costs of airtime and a data management system to automate the sending of text messages were nominal.

We include as beneficiaries all grade 1-3 students in the 92 schools whose WEOs had received at least one text message by the time of midline data collection: 34,041 in total. It is possible that all the students in the school benefited, although we only know the impacts for students who were in grades 2 and 3 at the start of 2019.

The per-pupil cost thus ranges between \$0.18 to \$0.68, depending on whether we include the costs of the workshops or not. Arguably the workshops are variable costs, since they are only required to take place once to introduce the program and clarify roles and lines of authority. But

[43] One WEO: "Text messages were very specific, so they made communication between WEOs and schools easier when regular reports were long and started with many things, such as information about the history of the school, which was not necessary or beneficial."

[44] From the qualitative report: "Most WEOs are overburdened with many roles/tasks coming from their DEOs and DEDs. So, reminding them to follow up on specific issues at specific schools ensures that they prioritize those issues and schools."

we do not know whether booster workshops might be required to sustain effects. Either way, their costs would be substantially reduced if implemented by government, since the largest cost-driver is the per diems paid to the participants.[45] We therefore consider these costs as an upper bound.

Taken together, we estimate that the program caused an improvement of between 20 and 76 Learning-Adjusted Years of Schooling per 100 dollars spent per student.[46] For comparison, Angrist et al. (2020) shows the cost-effectiveness of all known education interventions in developing countries that were rigorously evaluated and also include cost data. Compared to this list, our program is either the second- or fourth-most cost-effective intervention ever evaluated, and is 33 to 128 times more cost-effective than the median study that showed benefits.[47]

8 Discussion and Conclusion

This paper reports on a randomized evaluation of an ambitious education reform aimed at improving school governance in Tanzania. The key component of the reform was the conduct of Whole School Visits (WSVs) in schools, which provided relevant diagnostic information on poor outcomes and practices and recommendations on how to improve school management, teaching practices, and student learning. The WSVs were rolled out to all schools in the country over a period of four years. We find that the program itself had no impact, even though it was well implemented and generated information that was informative of school quality. This is consistent with the results of an evaluation of a related reform in India (Muralidharan and Singh, 2020).

A potential point of failure in this program was weak coordination between the bureaucrats who perform the inspections, and bureaucrats from a different ministry (the WEOs) who are responsible for follow-up visits to make sure that the recommendations are being implemented. With this coordination challenge in mind, we developed a low-cost program that collected WSV reports and sent text messages directly to WEOs informing them of the main recommendations and encouraging them to follow up with schools. Messages were signed as coming from their managers and workshops were held with WEOs and their managers to signal manager endorsement for the text messaging. We find that combining the WSV program with the text messages caused

[45] The per diem rate is substantially higher if paid for by external organizations

[46] See Filmer et al. (2020) for more information on this metric of learning.

[47] Median was 0.59

improvements in teaching practices and student learning. Moreover, all of the improvements in student learning were concentrated in regions with a high level of donor support, where WEOs were equipped with sufficient resources to regularly visit schools. The text messages also increased monitoring frequency in these regions.

Although this study was not designed to cleanly identify mechanisms, the combination of quantitative and qualitative evidence suggests that the text messages helped overcome both information frictions and management bandwidth constraints that hindered between- and within-agency coordination in the rollout of a new program. First, it reduced the cost of receiving the information. The WEOs got the information faster and found it valuable to have it on their phone to refer to and send directly to the schools. In the counterfactual arm, the WEOs rarely got timely access to the report, since it was sent to the District Executive Director. The reports were also relatively long and difficult to digest. Second, the text messages reduced the costs of acting on the information. Officially the program expected DEOs to read all the reports, digest the most important information, and re-assign tasks to WEOs based on this information. But both the DEOs and WEOs are already overburdened with existing tasks and responsibilities. The fact that the text messages were signed as coming from the DEOs enabled the WEOs to immediately act on the already customized recommendations, without waiting for new directives from their superior.

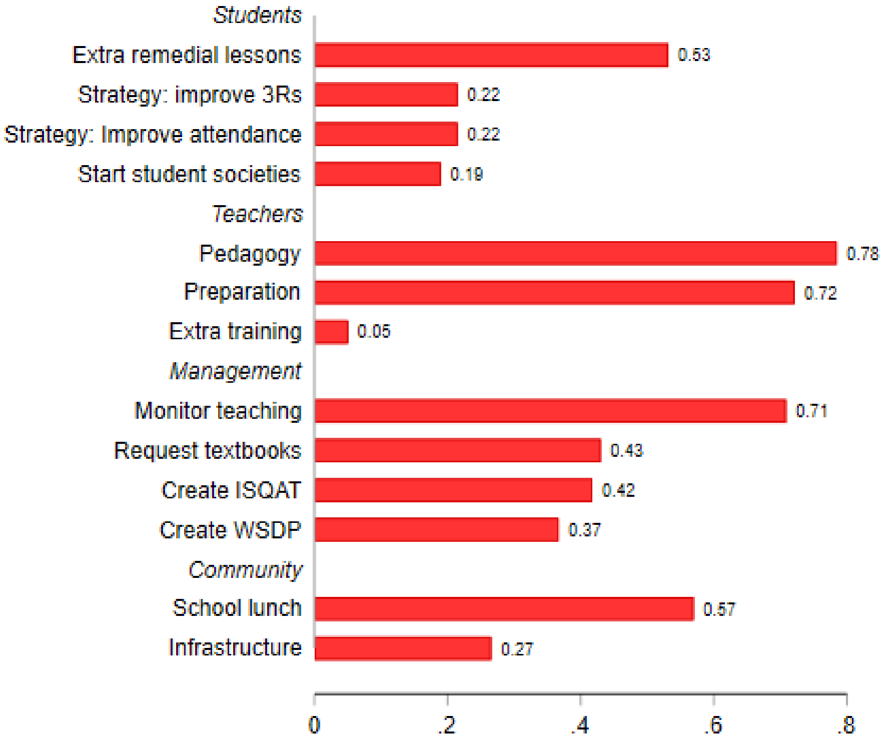
It is possible other channels not directly related to coordination challenges explain the results. First, the text messages acted as useful reminders to follow up on these recommendations when visiting schools. Second, they reduced the scope of tasks that WEOs are required to do, which allowed them to prioritize. Third, they might have motivated the WEOs, not only because it created perceptions of additional monitoring/oversight, but also because WEOs appreciated the attention and interest expressed in their job.

This study points to the potential high return of targeting additional resources to address specific bottlenecks in local service delivery. The text messages program is extremely cost-effective, because it did not require more human resources. Rather, it enabled a more efficient allocation of existing human resources by improving information flow and reducing the costs of coordination. But the program did not solve all inter-agency coordination challenges between the two relevant ministries in this sector. The institutional structure and incentives that caused weak coordination are still in place. If anything, the SMS intervention just highlights the cost of weak coordination, by showing the benefits of addressing it. As a metaphor, our booster program was the application of tape to plug a leaking pipe.

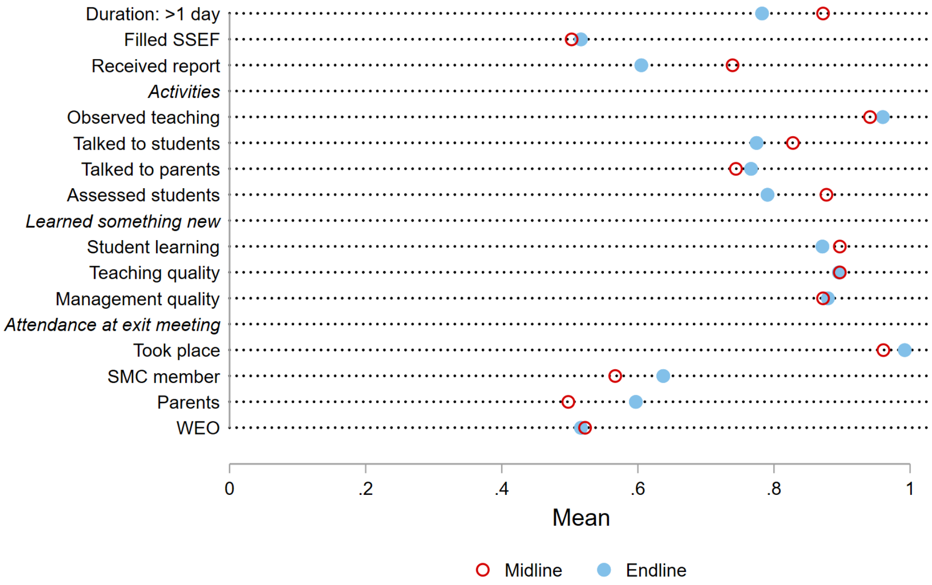
In our view, there are three broad policy recommendations from this study. First, allocate resources towards making critical implementation functions *easier* for the local bureaucrats responsible for service delivery, not harder. Many promising reforms require higher effort levels from one or more sets of bureaucrats. But interventions that reduce the cost of effort might be more sustainable and politically feasible. Second, the design of new programs should always consider the full set of existing tasks, roles, and responsibilities of the bureaucrats responsible for implementing them. Third, the design of new programs requires a deep understanding of the institutional context and the *de facto* role of all stakeholders involved in the sector. The coordination challenges addressed in this paper arise from well-meaning decentralization reforms that separated education standards and policy functions and the day-to-day supervision of schools across two agencies. Resources can be well spent on identifying and addressing bottlenecks in improving the implementation of both existing and new programs.

9 Tables and Figures

Figure 1: Implementation Quality



(a) Common Recommendations



(b) Nature of the Whole School Visits—by round of data collection

Note. Data in Figure (a) is our own categorization of the recommendations sent over text messages. Data in Figure (b) come from the head teacher survey, restricted to the sample of head teachers who indicated that a WSV had taken place in their school.

Table 1: Characteristics of the district SQAOs and the Ward Education Officers

	(1) SQAO		(2) WEO	
	Mean	Median	Mean	Median
Age	50.75	51	42.39	41
Female	0.27	0	0.24	0
University Degree	0.89	1	0.91	1
Years in current position	6.73	6	2.52	2
Experience as school teacher	1.00	1	0.97	1
Budget for fuel	0.98	1	0.61	1
Schools per Officer	14.01	13	4.29	4
Difficulty completing all tasks			0.74	1
Same position, baseline and midline			0.78	1
No. schools visited past 2 weeks			3.78	4
Observations	44		397	

Notes. We interviewed 44 District School Quality Assurance Officers (DSQAOs) at midline. We sampled 46 DSQAOs—two in each of the 23 districts in our sample—but two were unavailable because of prolonged illness. Panel A shows basic WEO and DEO characteristics.“Schools per Officer” is the number of DSQAOs per district or the number of schools per Ward in our sample. “No turnover since baseline” is a binary variable equal to one if the WEO surved at midline was the same as baseline. “Difficulty completing tasks” is a binary variable equal to one if the WEO agreed or strong agreeed with the following statement at baseline: “I find it difficult to complete all my tasks because different people expect different things from me.”

Table 2: Roll-out of program and challenges in coordination

	All districts						
	Mean	Obs					
Panel A. DSQAO exposure to training (DSQAO midline survey)							
Received training in new framework	0.98	44					
At least 5 days	0.70	43					
Central government provided	0.93	43					
At least 2 trainings	0.26	43					
Training by January 2019	0.91	43					
Sufficient length	0.12	43					
	Control		Visit		Visit&Text		Visit - Visit&Text
	Mean	Obs	Mean	Obs	Mean	Obs	Difference
Panel B. Program implementation and compliance							
WSV conducted (admin. data)							
—*Midline*	0.22	199	0.90	99	0.90	99	0.00
—*Endline*	0.43	199	0.96	99	0.98	99	-0.02
Text message sent to WEO (admin. data)							
—*Midline*†	0.00	43	0.00	89	0.93	89	-0.93***
—*Endline*†	0.00	43	0.00	89	0.99	89	-0.99***
WEO Received call or text†	0.33	43	0.22	89	0.60	89	-0.37***
Panel C. Coordination with WEOs							
Familiar with new framework	0.93	199	0.92	99	0.90	99	0.02
Received training in new framework	0.54	199	0.54	99	0.54	99	0.00
Believes that WSV conducted†	0.79	43	0.88	89	0.84	89	0.03
Can show copy of report†	0.19	43	0.21	89	0.21	89	0.00
Panel D. Number of recommendations recalled by WEO							
Student learning			0.32	78	0.61	75	-0.29**
Teaching			1.18	78	1.60	75	-0.42
Management			4.08	78	4.53	75	-0.46
Community			0.29	78	0.43	75	-0.13
Total			5.92	78	7.20	75	-1.28*

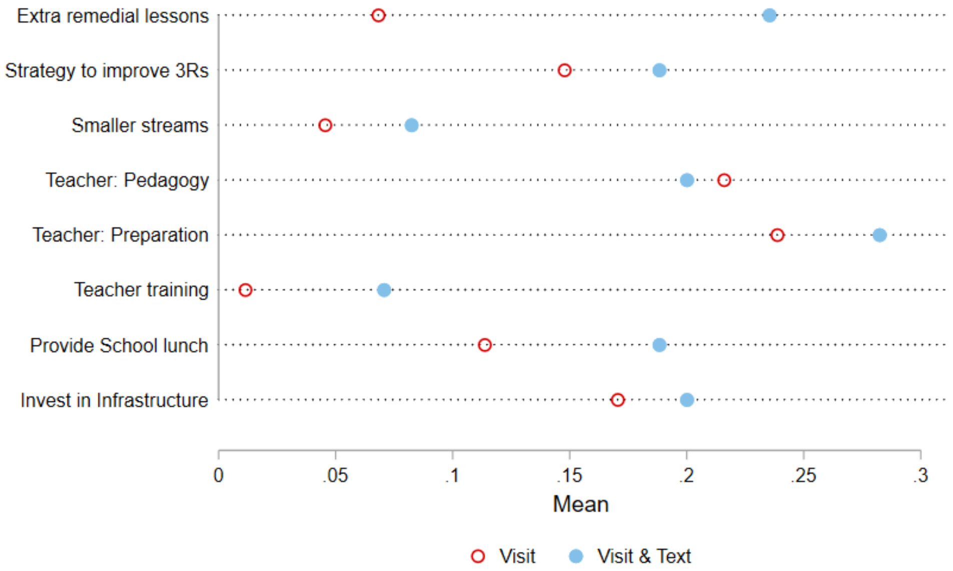
Notes. Unless otherwise stated, all data come from the midline WEO survey. All variables in Panels A, B, and C are binary. In Panel D data is restricted to observations where WEOs indicated that a WSV had taken place in the sample school. †=Data restricted to schools that received a WSV before the start of midline data collection

Table 3: Student learning

	Local Average Treatment Effects				Intent to Treat
	(1)	(2)	(3)	(4)	(5)
	Combined	Math	Kiswahili	English	Combined
Panel A. Midline (full sample)					
Visit	0.02	0.01	0.04	-0.05	0.02
	(0.04)	(0.05)	(0.04)	(0.08)	(0.03)
Visit&Text	0.04	0.02	0.05	0.05	0.03
	(0.03)	(0.04)	(0.03)	(0.07)	(0.03)
Visit=Visit&Text	0.616	0.680	0.656	0.145	0.588
Control mean	0.50	0.50	0.44	0.45	0.50
No. of schools	393	393	393	393	393
Observations	6589	6589	6589	3323	6589
First stage F-statistic	211	211	211	194	
Panel B. Endline (restricted sample)					
Visit&Text	0.11**	0.02	0.10***	0.16**	0.11**
	(0.05)	(0.05)	(0.04)	(0.07)	(0.05)
Visit mean	0.68	1.06	0.79	0.52	0.68
Observations	2896	2896	2896	2896	2896
No. of schools	176	176	176	176	176
First stage F-statistic	26530	26530	26530	26530	

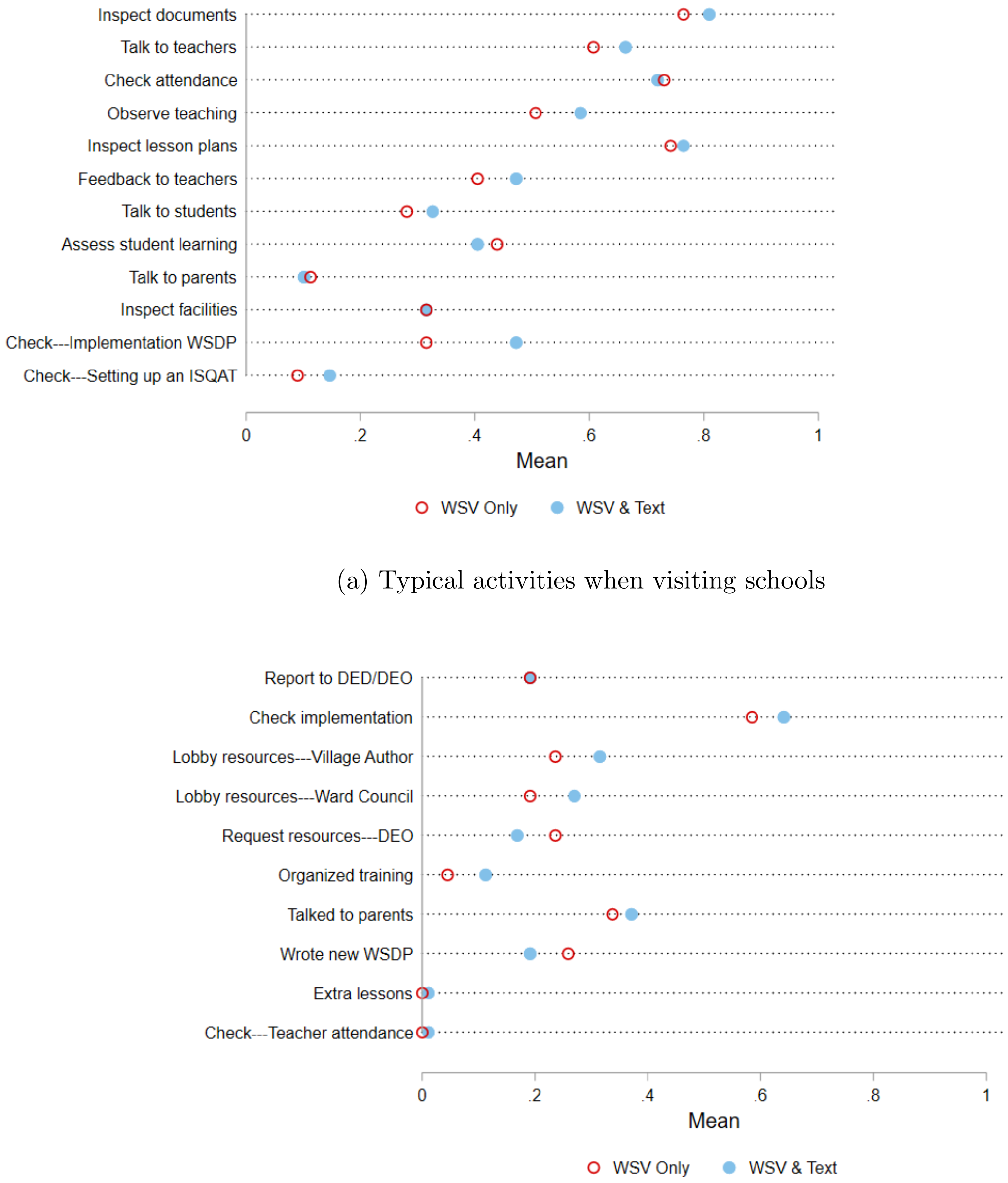
Notes: Each column represents a separate regression. Columns (1) to (4) are Local Average Treatment Effects, estimated using equations 1 (for Panel A) or 2 (for Panel B). Column (5) reports Intent to Treat Estimates. Restricted sample=schools in the treatment arms that received the WSV by the time of midline data collection. Baseline data missing for four schools—one in each treatment arm, and one in the control. Aggregate scores in Math, English, and Kiswahili are constructed using Item Response Theory, and standardized to have baseline control mean of zero and SD of one. Control variables are student gender, age and their respective baseline IRT score for each subject. Panel A, column 4, excludes the younger cohort, who were not assessed in English at baseline or midline. At midline the combined score is the average of Math and Kswahili; at endline it is the average all all three subjects. Standard errors in parentheses are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01.

Figure 2: Recommendations recalled by treatment arm



Note. Data come from the midline WEO survey, restricted to schools where the WEO indicated that a WSV had been conducted.

Figure 3: WEO activities when visiting schools and actions in response to the WSV



(a) Typical activities when visiting schools

(b) Actions in response to WSV

Note. Data come from the midline WEO survey. In Figure(b) the data is restricted to cases where the WEO indicated that a WSV had been conducted in the sample school.

Table 4: Teaching practices

	(1)	(2)	(3)	(4)	(5)	(6)
		Classroom				Student
	Attendance	Observations	Preparation	Assessment	Homework	Perceptions†
Panel A. Midline (full sample)						
Visit	-0.01	0.15	0.14	-0.14	-0.03	0.00
	(0.06)	(0.12)	(0.11)	(0.09)	(0.11)	(0.04)
	[0.98]	[0.78]	[0.78]	[0.78]	[0.98]	[0.98]
Visit&Text	0.10**	0.27***	0.16*	-0.00	0.13	0.01
	(0.05)	(0.11)	(0.08)	(0.08)	(0.11)	(0.04)
	[.12]	[.07]	[.12]	[.55]	[.22]	[.55]
Visit=Visit&Text						
P-value	0.054	0.242	0.813	0.077	0.148	0.835
Q-value	0.30	0.30	0.41	0.30	0.30	0.41
Control mean	0.46	0.00	0.00	-0.00	1.67	0.00
Observations	362	1044	2369	2626	3975	6397
No. schools	362	343	393	395	393	393
Unit of Observation	School	Teacher	Teacher	Teacher	Student	Student
Panel B. Endline (restricted sample)						
Visit&Text	0.03	-0.11	0.15*	-0.06	0.12	0.12***
	(0.04)	(0.11)	(0.08)	(0.06)	(0.10)	(0.03)
	[.37]	[.35]	[.16]	[.37]	[.35]	[<.01]
Visit mean	0.44	0.05	-0.03	0.03	1.51	-0.04
Observations	169	570	1217	1217	1776	2846
No. schools	169	178	177	177	176	176

Notes: All estimates are local average treatment effects, estimated using equating 1 (for Panel A) or 2 (for Panel B). Standard errors in parentheses are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01. Sharpened q-values that control for the false discovery rate are in square brackets (Anderson, 2008). P-values are grouped by dependent variable, with separate groupings for each treatment coefficient. The dependent variables are constructed as follows:
(1) Proportion of classes with students in them that also have a teacher present. (Schools with zero classes with students in them assigned to missing)
(2) Mean of the z-score of three indicators of teaching quality, as measured by the Teach classroom observation toolkit: instruction, classroom culture, and the proportion of times (based on three snapshots) when the majority of children were on task.
(3) Standardized mean of four binary variables: whether the teacher has (i) an updated lesson plan, (ii) a different lesson plan for each, (iii) updated scheme of work, and (iv) can show a class journal.
(4) Standardized mean of two binary variable: whether a teacher (i) assessed the skills of a child in the past 5 days; (ii) could show a record of student assessment.
(5) Counts the number of days in the last five school days that an exercise was completed. A random sample of students were selected for document inspection—some for English, some for Mathematics, and some for Kiswahili.
(6) Standardized mean of student answers on whether their teacher: (i) explains difficult concepts, (ii) motivates students, (ii) reviews and discusses content, (iii) gives practice tests, (iv) assigns homework, (v) gives feedback on homework, (vi) provides remedial teaching to struggling students.
Control variables are selected using LASSO (see Section 5). Potential school level controls are: school location (urban vs rural); respondent's position at school, the number of years they have held that position, year they started teaching at the school, and year they started teaching overall; school type; year school began operations; total number of students by gender, and all the baseline outcome indicators that are constructed at a school level. Potential teacher-level controls include: gender, position in school, year of birth, region of birth, district of birth, year they began teaching at this school and in general, education, whether they expect to teach at this school in the next year, whether they expect to teach the same grades the next year, whether they have ever taught at a private school, mode of transport used and time taken to travel to the school, whether housing is provided by the school, grade preference, and current gross total compensation per month. Potential classroom observation data controls include: baseline values for the different domains of classroom instruction, and the grade and subject of the class being taught. †=Outcome not included in pre-analysis plan

Table 5: Head teacher beliefs

	Quality		Education Production Function			
	(1) Room for Improvement	(2) Student learning	(3) Early vs Late grade	(4) Training vs Renovation	(5) Learning vs Curriculum	(6) Participatory learning
Panel A. Midline						
Visit	-0.004	0.020	0.031	-0.159*	-0.101	0.051
	(0.159)	(0.028)	(0.048)	(0.088)	(0.080)	(0.043)
	[.966]	[.858]	[.858]	[.775]	[.775]	[.775]
Visit&Text	-0.088	0.042*	0.002	-0.051	0.081	0.007
	(0.138)	(0.023)	(0.041)	(0.078)	(0.073)	(0.043)
	[1]	[.578]	[1]	[1]	[1]	[1]
Visit=Visit&Text						
P-value	0.557	0.370	0.511	0.198	0.019	0.286
Q-value	0.868	0.863	0.868	0.863	0.131	0.863
Control Mean	0.001	0.641	0.785	0.515	0.359	0.908
Observations	387	387	391	390	391	391
R-Squared	0.297	0.136	0.149	0.048	0.136	0.120
Panel B. Endline						
Visit&Text	0.072	0.064***	0.090**	0.070	0.013	-0.006
	(0.130)	(0.022)	(0.040)	(0.071)	(0.067)	(0.055)
	[1]	[.024]	[.07]	[.787]	[1]	[1]
Control Mean	0.000	0.663	0.794	0.472	0.348	0.831
Observations	177	178	178	178	178	178
R-Squared	0.308	0.345	0.218	0.146	0.135	0.221

Notes: Each column is a separate regression, estimated using equations 1 (for Panel A) or 2 (for Panel B). In Panel B, the sample is restrict to the schools that had received a WSV by the time of midline data collection. Data is at a head teacher level. The dependent variable in columns (1) is a Kling index of head teachers' responses to the following question: "Think back to the beginning of this school year (January/February 2019). How much room for improvement was there in the following areas?". Answers are categorical, ranging from 1 "A lot of room for improvement" to 4 "No improvement was necessary". See table B.14 for the 11 respective indicators. The dependent variable in column (2) is the mean of head teachers' beliefs about the share of grade 2 students in their school have grade 2-level literacy and numeracy skills, respectively. The dependent variables in columns (3) to (6) are binary variables for vignettes that illicit head teacher's preference for different education inputs. Robust standard errors in parentheses. Sharpened q-values in square brackets (grouped across the 6 dependent variables, separately for each treatment). * for p<.1; ** for p<.05; *** for p<.01. Sharpened q-values that control for the false discovery rate are in square brackets (Anderson, 2008). P-values are grouped by dependent variable, with separate groupings for each treatment coefficient. Estimates include strata fixed effects. Control variables are selected using 10-fold Lasso cross-validation.

Table 6: School leadership management practices

	(1) WSDP	(2) Monitoring	(3) Curriculum Guidance
Panel A. Midline			
Visit	0.000	-0.055	0.077
	(0.071)	(0.114)	(0.096)
	[1]	[1]	[1]
Visit&Text	0.124*	0.079	0.006
	(0.065)	(0.096)	(0.086)
	[0.218]	[0.688]	[0.1]
Visit=Visit&Text			
P-value	0.081	0.224	0.444
Q-value	.32	.32	.421
Control Mean	0.205	0.000	0.00
Observations	391	2357	2357
R-Squared	0.104	0.122	0.045
Panel B. Endline			
Visit&Text	.	0.006	-0.093
	.	(0.083)	(0.081)
	.	[0.978]	[0.978]
Visit Mean	.	0.087	0.060
Observations	.	1092	1092
R-Squared	.	0.138	0.078

Notes. Each column is a separate regression estimated using equation 1. "WSDP" is a a binary variable equal to one if the head teacher can show a Whole School Development Plan that was recently revised (since October 2019); this question was only asked at midline. "Monitoring" and "Curriculum" are Kling indeces, standardized to have a the control standard deviation of one. See Tables B.15 and B.16 for the variables that constitute each index. P-values are grouped by dependent variable, with separate groupings for each treatment coefficient. * for p<.1; ** for p<.05; *** for p<.01. Sharpened q-values that control for the false discovery rate are in square brackets (Anderson, 2008).

Table 7: WEO monitoring and behavior at midline

	(1)	(2)	(3)	(4)	(5)
Panel A. Monitoring (WEO survey)					
	Log Number of Visits†		Log Length of Visit†		Activities
	Intervention schools	Other	Intervention schools	Other	No.
Visit&Text	0.08	0.07	0.10	0.11	0.45
	(0.10)	(0.10)	(0.13)	(0.11)	(0.31)
Visit mean	2.412	2.425	4.697	4.716	5.360
Observations	178	171	174	168	178
R-Squared	0.323	0.265	0.186	0.269	0.272
Panel B. Actions (head teacher survey)					
	Overall	Followed up	Action— improve learning	Organized workshop	Meet stakeholders this year
Visit&Text	0.24*	0.06	0.09	0.08	-0.04
	(0.14)	(0.04)	(0.06)	(0.06)	(0.05)
Visit mean	0.000	0.875	0.739	0.261	0.779
Observations	176	176	176	176	178
R-Squared	0.161	0.233	0.165	0.421	0.107

Notes: Data is restricted to 89 schools in the Visit and Visit&Text treatment arms, respectively, where a WSV was conducted by the time of midline data collection. All dependent variables in Panel A and column (5) in Panel B are constructed from the WEO survey. The number of observations in columns (2) and (4), Panel A, are reduced because in some Wards there is only one school. The number of observations in columns (3) and (4) are further reduced because there are four schools where a the WEO reported to have never visited the school. The dependent variables in columns (2) to (4) in Panel B come from the midline head teacher survey, which has two missing values in this sample. "Number of Schools" is the inverse hyperbolic sin of the number of times that the WEO reported to have visited a school in the past three months. "Intervention" refers to the school in our sample, "Not" refers to the mean of a random sample of up to three other schools in the same Ward. "Length of Visit" is log of the amount of time that the WEO reported to spend at a school the last time they visited it. "Activities" is the total number of activities they report to do when visiting a school. See Figure 3 for a list of some of the most common activities. The dependent variables in columns (2) to (5), Panel B, are binary. "Overall" is the mean of these, further standardized to the mean and standard deviation in the Visit arm. Control variables that are predictive of the dependent variable are selected using Lasso. Potential controls include: baseline outcome indicators (where appropriate), WEO demographic characteristics, number of schools in the WArd, whether the WEO is a replacement or not, and the same school and head-teacher level characteristics mentioned in Table 4. †=Outcomes not include in pre-analysis plan.

Table 8: Student learning at endline, by donor involvement and teacher incentives (restricted sample)

	(1) Combined	(2) Math	(3) Kiswahili	(4) English
Panel A. Equip-T				
Visit&Text (γ_1)	0.00	-0.06	0.01	0.05
	(0.06)	(0.06)	(0.05)	(0.09)
Visit&Text × Equip-T (γ_2)	0.21*	0.17	0.16*	0.21
	(0.11)	(0.11)	(0.09)	(0.15)
Equip-T (γ_3)	-0.13	-0.01	-0.14**	-0.19*
	(0.08)	(0.08)	(0.07)	(0.11)
$\gamma_1 + \gamma_2 = 0$	0.023	0.236	0.027	0.031
Observations	2896	2896	2896	2896
No. schools	176	176	176	176
Panel B. Incentives				
Visit&Text (γ_1)	0.13**	0.03	0.12***	0.19**
	(0.05)	(0.06)	(0.04)	(0.07)
Visit&Text × Incentives (γ_2)	-0.06	0.01	-0.10	-0.05
	(0.11)	(0.10)	(0.12)	(0.14)
Incentives (γ_3)	0.02	0.09	0.04	-0.08
	(0.09)	(0.08)	(0.09)	(0.12)
$\gamma_1 + \gamma_2 = 0$	0.477	0.696	0.839	0.282
Observations	2770	2770	2770	2770
No. Schools	168	168	168	168

Notes: Each column is a separate regression estimated using equation 3, using endline student assessment data. Data restricted to 178 schools in the treatment groups where a WSV had taken place by the time of midline data collection. In Panel B the sample is further restricted to the 379 schools which were the sampling frame used to randomize assignment to the teacher incentives arm. See Table 3 for construction of dependent variables and selection of controls.

Table 9: Teacher behavior at endline, by donor involvement and teacher incentives

	(1)	(2)	(3)	(4)	(5)	(6)
	Attendance	Classroom Observations	Preparation	Assessment	Homework	Student Perceptions†
Panel A. Equip-T						
Visit&Text (γ_1)	0.05	-0.22	-0.07	-0.02	0.20	0.07*
	(0.06)	(0.17)	(0.13)	(0.11)	(0.15)	(0.04)
Visit&Text × Equip-T (γ_2)	-0.02	0.20	0.42**	-0.05	-0.15	0.10
	(0.10)	(0.25)	(0.19)	(0.14)	(0.25)	(0.08)
Equip-T (γ_3)	0.16**	0.07	-0.28*	-0.28***	0.16	-0.13*
	(0.07)	(0.16)	(0.15)	(0.10)	(0.18)	(0.07)
$\gamma_1 + \gamma_2 = 0$	0.744	0.921	0.011	0.395	0.829	0.022
Observations	169	570	1105	1217	1776	2846
No. schools	169	178	177	177	176	176
Unit of Observation	School	Teacher	Teacher	Teacher	Student	Student
Panel B. Teacher Incentives						
Visit&Text (γ_1)	0.01	-0.22	0.23**	-0.17**	0.19	0.11**
	(0.06)	(0.15)	(0.11)	(0.07)	(0.13)	(0.04)
Visit&Text × Incentives (γ_2)	0.10	0.33	-0.13	0.37**	-0.20	-0.03
	(0.11)	(0.27)	(0.19)	(0.16)	(0.29)	(0.09)
Incentives (γ_3)	-0.00	-0.30	-0.03	-0.13	0.34	0.07
	(0.09)	(0.21)	(0.15)	(0.10)	(0.21)	(0.06)
$\gamma_1 + \gamma_2 = 0$	0.175	0.573	0.504	0.143	0.972	0.248
Observations	160	536	1061	1168	1699	2720
No. schools	160	169	169	169	168	168

Notes: See Table 4 for a description of dependent variables, and Table 8 for description of empirical strategy and sample

Table 10: WEO behavior at midline–interacted with Equip-T

	(1)	(2)	(3)	(4)	(5)
Panel A. Monitoring					
	Log Number of Visits		Log Length of Visit		Activities
	(1)	(2)	(3)	(4)	(5)
	Intervention school	Other	Intervention school	Other	No.
Visit&Text (β_1)	-0.070	-0.207	0.106	0.194	0.612
	(0.128)	(0.129)	(0.153)	(0.142)	(0.451)
Visit&Text × Equip-T (β_2)	0.384*	0.627***	0.009	-0.148	-0.390
	(0.227)	(0.226)	(0.252)	(0.226)	(0.641)
Equip-T (β_3)	-0.188	-0.323	-0.347**	-0.285*	-0.717
	(0.185)	(0.196)	(0.173)	(0.151)	(0.488)
Equip-T Mean (Visit&Text=0)	2.296	2.224	4.481	4.543	5.029
Observations	178	171	174	168	178
$\beta_1 + \beta_2 = 0$	0.094	0.024	0.567	0.795	0.625
Panel B. Behavior					
	Overall	Followed up	Action— improve learning	Organized workshop	Meet stakeholders this year
Visit&Text (β_1)	0.188	-0.013	0.067	0.053	0.007
	(0.192)	(0.061)	(0.087)	(0.096)	(0.059)
Visit&Text × Equip-T (β_2)	0.064	0.141	0.046	0.051	-0.084
	(0.296)	(0.096)	(0.124)	(0.132)	(0.099)
Equip-T (β_3)	-0.183	-0.084	0.042	-0.249***	0.009
	(0.227)	(0.077)	(0.095)	(0.094)	(0.072)
Equip-T Mean (Visit&Text=0)	-0.112	0.824	0.765	0.176	0.784
Observations	176	176	176	176	178
$\beta_1 + \beta_2 = 0$	0.265	0.087	0.206	0.251	0.336

Notes: See Table 7 for a description of dependent variables.

Table 11: Cost and Cost-Effectiveness

Panel A. Costs	
Workshops	
Salary*	$1,930.31
Transport	$2,019.58
Accommodation	$1,263.47
Participant Per Diems	$10,432.81
Printing	$109.99
Catering	$1,088.55
Venue Rental	$110.94
Transfer Fees/Miscellaneous	$242.88
Subtotal	*$17,198.53*
Variable costs	
Salary†	$5,790.00
Contract with service provider	$130.24
Air time	$157.07
Subtotal	$6,077.31
Total	$23,275.84
Panel B. Cost-effectiveness	
Number of beneficiaries‡	34,041
Treatment effect	
—Standard deviations	0.11
—Learning Adjusted Years of Schooling (LAYS) §	0.14
Variable costs per student	
—Excluding costs of workshops	$0.18
—Including costs of workshops	$0.68
LAYS gains per $100 per student	
—Excluding costs of workshops	76
—Including costs of workshops	20

Notes: * 18 days of program manager's time to organize and host the workshops; † Remuneration for managing the program: collating WSV reports, summarizing the main recommendations, managing the data system, calling WEOs to update register. ‡ All grade 1 to 3 students in the 92 schools whose WEO received a WSV from us by the time of endline data collection. § Uses approach by Angrist et al. (2020), which divides the effect size by 0.8 standard deviations—the benchmark for an high-performing learning rate.

References

Anand, Gautam, Aishwarya Atluri, Lee Crawfurd, Todd Pugatch, and Ketki Sheth, "Improving School Management in Low and Middle Income Countries: A Systematic Review," 2023.

Anderson, Michael L, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American statistical Association*, 2008, *103* (484), 1481–1495.

Angrist, N. and R. Meager, "The role of implementation in generalisability: A synthesis of evidence on targeted educational instruction and a new randomised trial," Technical Report, Oxford University, Center of Excellence for Development Impact and Learning, CEDIL 2022.

Angrist, Noam, David K Evans, Deon Filmer, Rachel Glennerster, F Halsey Rogers, and Shwetlena Sabarwal, "How to improve education outcomes most efficiently? A Comparison of 150 interventions using the new Learning-Adjusted Years of Schooling metric," 2020.

Azulai, Michel, Imran Rasul, Daniel Rogger, and MJ Williams, "Can training improve organizational culture? Experimental evidence from Ghana's civil service," Technical Report, Tech. rep., University College London: 7 2020.

Banerjee, Abhijit V, Esther Duflo, and Rachel Glennerster, "Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system," *Journal of the European Economic Association*, 2008, *6* (2-3), 487–500.

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, 2006, *93* (3), 491–507.

Blimpo, Moussa P, David K Evans, and Nathalie Lahire, "School-based management and educational outcomes: Lessons from a randomized field experiment," *Unpublished manuscript*, 2011.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen, "Does Management Matter in Schools?," *The Economic Journal*, 2015, *125* (584), 647–674.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Justin Sandefur et al., "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, 2018, *168*, 1–20.

Brodkin, Evelyn Z, "Policy work: Street-level organizations under new managerialism," *Journal of Public Administration Research and Theory*, 2011, *21* (suppl_2), i253–i277.

Cilliers, Jacobous, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor, "How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching," *The Journal of Human Resources*, 2020, *55* (3), 926–962.

Cilliers, Jacobus, Eric Dunford, and James Habyarimana, "What Do Local Government Education Managers Do to Boost Learning Outcomes?," *The World Bank Economic Review*, 03 2022. lhac001.

_ , Isaac M Mbiti, and Andrew Zeitlin, "Can Public Rankings Improve School Performance? Evidence from a Nationwide Reform in Tanzania," *Journal of Human Resources*, 2020, pp. 0119–9969R1.

Dasgupta, Aditya and Devesh Kapur, "The political economy of bureaucratic overload: Evidence from rural development officials in India," *American Political Science Review*, 2020, *114* (4), 1316–1334.

Deserranno, Erika, Philipp Kastrau, and Gianmarco León-Ciliotta, "INCENTIVES IN MULTI-LAYERED ORGANIZATIONS," 2020.

Dixit, Avinash, "Incentives and organizations in the public sector: An interpretative review," *Journal of human resources*, 2002, pp. 696–727.

Evans, David K and Fei Yuan, "How big are effect sizes in international education studies?," *Educational Evaluation and Policy Analysis*, 2022, *44* (3), 532–540.

Filmer, Deon, Halsey Rogers, Noam Angrist, and Shwetlena Sabarwal, "Learning-adjusted years of schooling (LAYS): Defining a new macro measure of education," *Economics of Education Review*, 2020, *77*, 101971.

Glisson, Charles, "Judicial and service decisions for children entering state custody: The limited role of mental health," *Social Service Review*, 1996, *70* (2), 257–281.

Gulzar, Saad, Benjamin J Pasquale et al., "Politicians, bureaucrats, and development: Evidence from India," *American Political Science Review*, 2017, *111* (1), 162–183.

Head, Katharine J, Seth M Noar, Nicholas T Iannarino, and Nancy Grant Harrington, "Efficacy of text messaging-based interventions for health promotion: a meta-analysis," *Social science & medicine*, 2013, *97*, 41–48.

Hill, Carolyn and Laurence Lynn, "Producing human services Why do agencies collaborate?," *Public management review*, 2003, *5* (1), 63–81.

Kean, Thomas and Lee Hamilton, *The 9/11 commission report: Final report of the national commission on terrorist attacks upon the United States*, Vol. 3, Government Printing Office, 2004.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental analysis of neighborhood effects," *Econometrica*, 2007, *75* (1), 83–119.

Lichand, Guilherme and Julien Christen, "Behavioral nudges prevent student dropouts in the pandemic," Technical Report, Working Paper 2021.

List, John, *The Voltage Effect*, Currency, 2022.

Miles, Thomas J and Adam B Cox, "Does immigration enforcement reduce crime? evidence from secure communities," *The Journal of Law and Economics*, 2014, *57* (4), 937–973.

Mo, Di, Renfu Luo, Chengfang Liu, Huiping Zhang, Linxiu Zhang, Alexis Medina, and Scott Rozelle, "Text messaging and its impacts on the health and education of the poor: evidence from a field experiment in rural China," *World development*, 2014, *64*, 766–780.

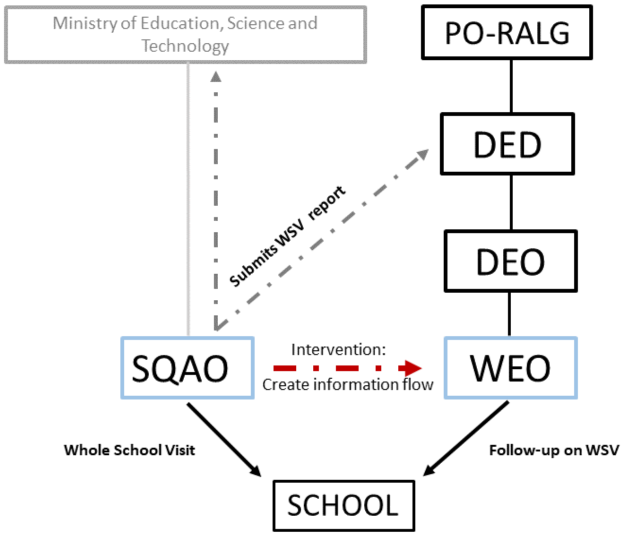
Molina, Ezequiel, Syeda Farwa Fatima, Andrew Ho, Carolina Melo Hurtado, Tracy Wilichowksi, and Adelle Pushparatnam, "Measuring Teaching Practices at Scale: Results from the Development and Validation of the Teach Classroom Observation Tool," 2018.

Muralidharan, Karthik and Abhijeet Singh, "Improving public sector management at scale? experimental evidence on school governance india," Technical Report, National Bureau of Economic Research 2020.

Rasul, Imran and Daniel Rogger, "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service," *The Economic Journal*, 2018, *128* (608), 413–446.

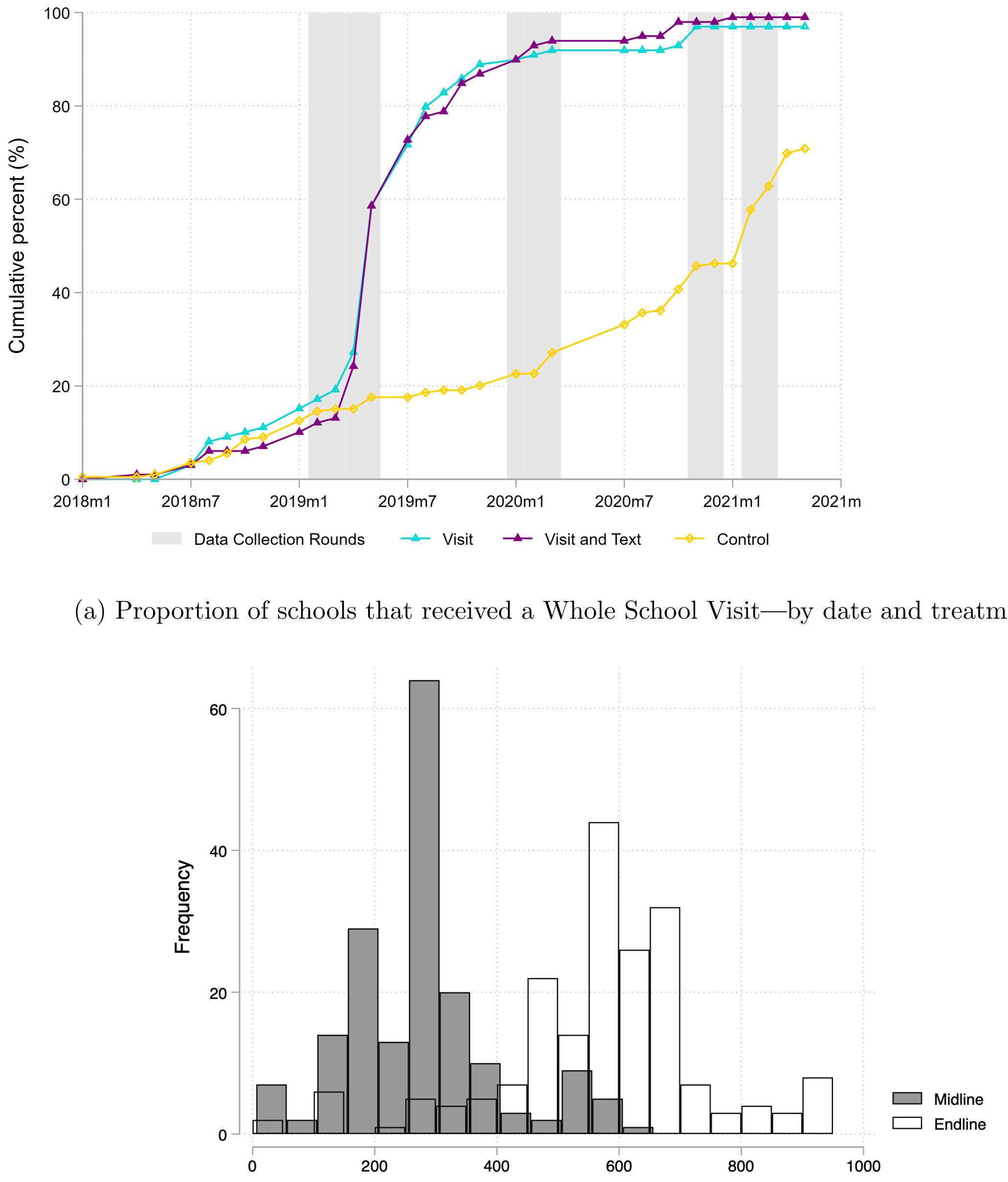
Appendix A Additional figures

Figure A.1: Organizational structure of stakeholders of Whole School Visits



Note. The School Quality Assurance Officers (SQAOs) conduct Whole School Visits (WSVs) and send the WSV reports directly to the Ministry of Education, Science, and Technology, as well as to the District Executive Director (DED). The DEO is then expected to pass it on to the District Education Officer (DEO). The DEO can assign new tasks to the WEO based on this report, and might pass on the report directly to the WEO. Our text messages intervention by-passes the long process and sends information directly to the WEOs.

Figure A.2: Whole School Visits

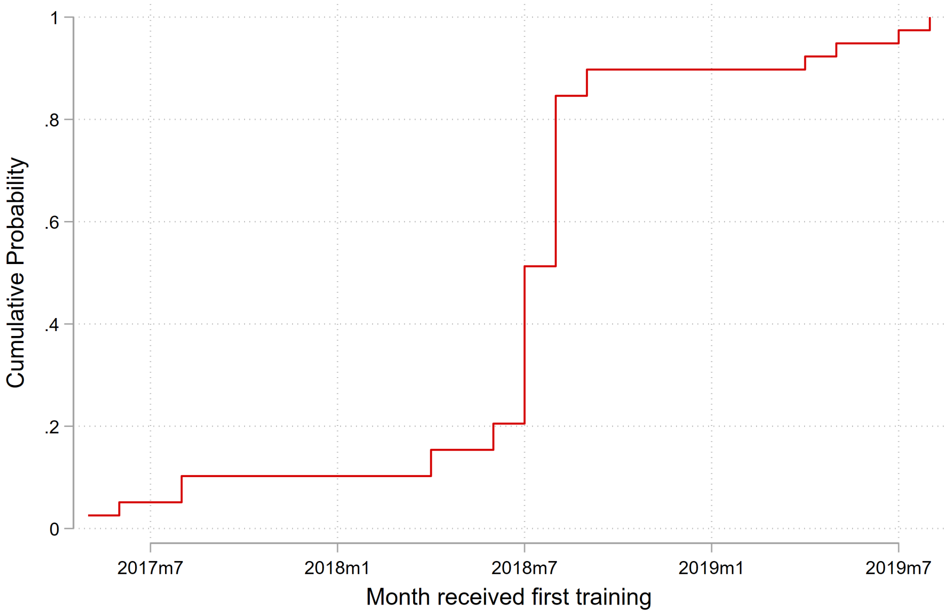


(a) Proportion of schools that received a Whole School Visit—by date and treatment arm

(b) Duration between Whole School Visit and Data Collection, days

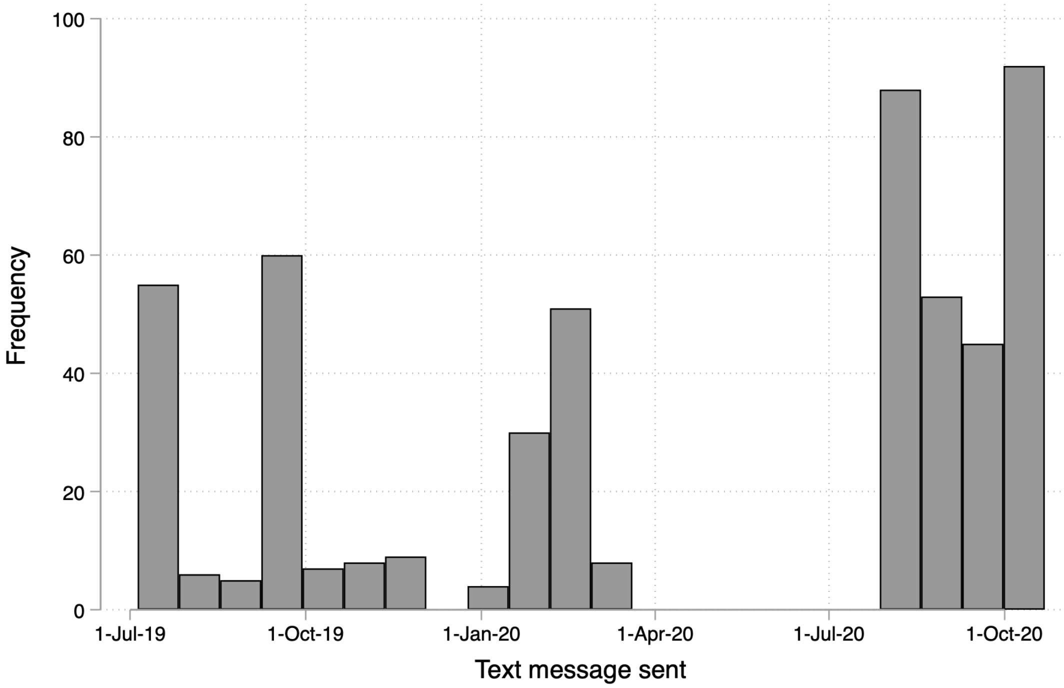
Note. Administrative data provided by government includes the date of each WSV that was conducted by June 2021.

Figure A.3: Date that DSQAOs received their first training in the new framework

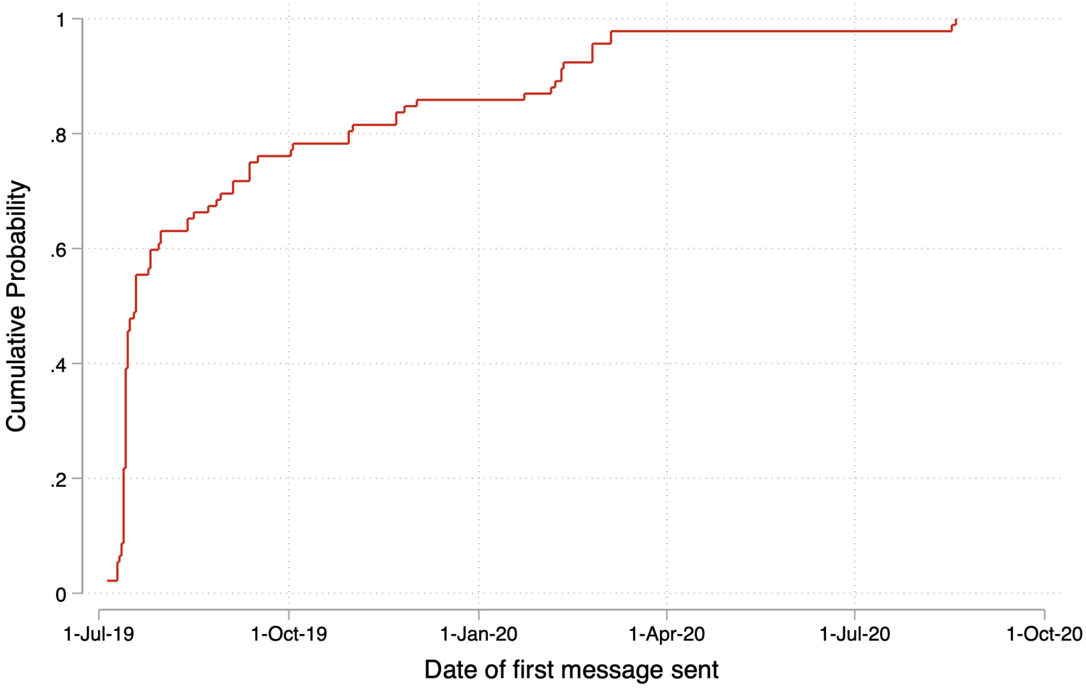


Note. Data come from our midline DSQAO survey

Figure A.4: Sending of text messages to WEOs



(a) Histogram: Dates that messages got sent.



(b) CDF: Date of first message sent

Note. Data come from the platform that managed the sending of text messages.

Appendix B Additional tables

Table B.1: Difference at baseline between Equip-T and other regions

	(1)	(2)	T-test
	Other	Equip-T	Difference
Variable	Mean/SE	Mean/SE	(1)-(2)
Panel A. Student-level			
Learning (composite)	0.042	-0.051	0.093
	(0.026)	(0.063)	
N	3773	3134	
Clusters	3	3	
Panel B. Classroom Observations			
Index	-0.033	-0.022	-0.012
	(0.057)	(0.051)	
Instruction	2.599	2.698	-0.099
	(0.076)	(0.020)	
Classroom Culture	3.330	3.288	0.043
	(0.027)	(0.013)	
Students on task	0.531	0.506	0.025
	(0.033)	(0.040)	
N	596	448	
Clusters	3	3	
Panel C. School location and WEO level			
Rural	0.793	0.711	0.082
	(0.051)	(0.040)	
Sufficient budget—fuel	0.138	0.556	-0.417**
	(0.017)	(0.149)	
Sufficient budget—maintenance	0.051	0.461	-0.410**
	(0.014)	(0.116)	
School visits in 2 weeks	3.544	3.972	-0.428***
	(0.053)	(0.077)	
Meet at least monthly	0.691	0.994	-0.303***
	(0.040)	(0.005)	
Time-observing teaching	17.115	22.394	-5.279**
	(1.693)	(1.406)	
Time-inspecting documents	22.853	20.283	2.569
	(0.533)	(1.699)	
Received training in new framework	0.350	0.600	-0.250*
	(0.106)	(0.032)	
N	217	180	
Clusters	3	3	

Notes: : The value displayed for t-tests are the differences in the means across the groups. "School visits in 2 weeks" is winsorized at the 95 percentile. The median number of visits is 4 and 3 in the Equip-T and other regions, respectively. Standard errors are clustered at variable regional level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table B.2: Comparing excluded schools with selected sample of schools

	(1) Average marks in 2016	(2) Average marks 2016-2013
Selected schools	-0.439	-0.533
	(1.301)	(1.068)
Excluded schools	-13.645***	-12.492***
	(3.153)	(2.588)
Observations	1,640	1,640
R-squared	0.188	0.208
Mean- not selected schools	119.4	113.7
F-Test: p-value	< 0.01	< 0.01

Notes: Each column represents a separate regression, including district fixed effects. Data is restricted to all primary schools in our sample of selected districts. "Selected schools" and "Excluded school" are dummy variables indicating (i) whether the school is in our evaluation sample; and (ii) whether the school was excluded prior to drawing the sample because a WSV had already taken place in that school. The outcome variable is the school's average score in the national, standardized Primary School Leaving Exam (PSLE). The reported p-value is for the null hypothesis that performance in the selected and excluded schools are the same.

Table B.3: Student-level attrition analysis

	Midline			Endline		
	(1) Attrite	(2) Age	(3) Learning	(4) Attrite	(5) Age	(6) Learning
Visit	0.007	-0.040	-0.013	-0.014	-0.010	-0.016
	(0.009)	(0.060)	(0.053)	(0.008)	(0.059)	(0.054)
Visit&Text	-0.003	0.030	-0.035	-0.005	0.020	-0.030
	(0.008)	(0.062)	(0.053)	(0.009)	(0.060)	(0.053)
Attrite		-0.030	-0.278**		0.110	-0.191**
		(0.113)	(0.087)		(0.093)	(0.065)
Attrite x Visit		0.405*	-0.118		-0.051	-0.154
		(0.175)	(0.127)		(0.179)	(0.123)
Attrite x Visit&Text		0.295	0.088		0.371*	-0.007
		(0.183)	(0.135)		(0.160)	(0.119)
Control Mean	0.057	9.017	0.014	0.077	9.017	0.014
N	6991	6991	6991	6991	6991	6991
R-squared	0.017	0.059	0.084	0.012	0.060	0.083

Notes: Learning is an aggregate score of students' baseline performance in a Mathematics and Kiswahili test, validated using Item Response Theory. Attrition is a dummy variable equal to one if the student was not assessed at midline/endline. * for p<.1; ** for p<.05; *** for p<.01. Estimates include strata fixed effects.

Table B.4: Balance Tests on Full Intervention Sample

	(1)	(2)	(3)		P-value	
	Control	Visit	Visit&Text	(1)-(2)	(1)-(3)	(2)-(3)
Panel A. School and Head Teacher Characteristics						
Rural	0.729	0.768	0.798	0.545	0.102	0.280
	(0.032)	(0.043)	(0.041)			
Pubic School	0.975	0.980	0.970	0.863	0.675	0.597
	(0.011)	(0.014)	(0.017)			
Years at school	8.497	7.808	7.697	0.332	0.213	0.890
	(0.449)	(0.551)	(0.457)			
Teaching experience (years)	18.291	18.182	16.970	0.967	0.183	0.265
	(0.654)	(0.852)	(0.709)			
Years in position	3.553	2.848	2.919	0.142	0.156	0.959
	(0.304)	(0.322)	(0.324)			
F-test of joint significance (p-value)				0.711	0.281	0.761
N	199	99	99			
Panel B. Classroom observations						
Teaching Quality Index	-0.015	-0.013	-0.048	0.795	0.532	0.430
	(0.032)	(0.043)	(0.042)			
N	383	195	197			
Clusters	198	99	99			
WEO						
Male	0.779	0.758	0.737	0.619	0.397	0.797
	(0.029)	(0.043)	(0.044)			
Age (in 2021)	43.809	44.909	45.040	0.169	0.130	0.815
	(0.458)	(0.659)	(0.702)			
University Degree	0.844	0.778	0.828	0.144	0.642	0.389
	(0.026)	(0.042)	(0.038)			
F-test of joint significance (p-value)				0.339	0.592	0.918
N	199	99	99			
Panel C. Teacher characteristics						
Age (in 2021)	38.247	38.420	38.501	0.607	0.490	0.994
	(0.281)	(0.432)	(0.405)			
Male	0.689	0.691	0.688	0.885	0.969	0.984
	(0.019)	(0.025)	(0.030)			
F-test of joint significance (p-value)				0.871	0.784	1.000
N	1352	685	676			
Clusters	199	99	99			
Panel D. Student characteristics						
Male	0.498	0.506	0.493	0.584	0.777	0.504
	(0.009)	(0.013)	(0.012)			
Age	9.017	8.995	9.066	0.819	0.526	0.364
	(0.039)	(0.060)	(0.062)			
Math	0.011	0.002	-0.024	0.930	0.643	0.712
	(0.032)	(0.050)	(0.051)			
Kiswahili	0.015	-0.034	0.003	0.497	0.918	0.523
	(0.037)	(0.048)	(0.049)			
F-test of joint significance (p-value)				0.708	0.869	0.258
N	3525	1715	1751			
Clusters	197	98	98			

Notes: The value displayed for t-tests are the differences in the means across the groups. Standard errors are clustered at the School level, where appropriate. Strata fixed effects are included in all estimations. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table B.5: Balance Tests on Reduced Sample (non-missing at Midline)

	(1)	(2)	(3)		P-value	
	Control	Visit	Visit&Text	(1)-(2)	(1)-(3)	(2)-(3)
Panel A. School and Head Teacher Characteristics						
Rural	0.723	0.765	0.796	0.540	0.117	0.295
	(0.032)	(0.043)	(0.041)			
Pubic School	0.974	0.980	0.969	0.863	0.691	0.597
	(0.011)	(0.014)	(0.017)			
Years at school	8.549	7.847	7.643	0.333	0.183	0.830
	(0.457)	(0.555)	(0.459)			
Teaching experience (years)	18.154	18.184	16.980	0.916	0.265	0.294
	(0.661)	(0.861)	(0.716)			
Years in position	3.544	2.857	2.929	0.154	0.183	0.959
	(0.309)	(0.326)	(0.328)			
F-test of joint significance (p-value)				0.699	0.337	0.786
N	195	98	98			
Panel B. Classroom observations						
Teaching Quality Index	-0.007	-0.047	-0.124	0.392	0.255	0.534
	(0.069)	(0.091)	(0.084)			
N	167	84	91			
Clusters	128	65	69			
Panel C. Student characteristics						
Male	0.498	0.502	0.486	0.800	0.469	0.416
	(0.009)	(0.013)	(0.012)			
Age	9.019	8.973	9.055	0.528	0.689	0.280
	(0.039)	(0.062)	(0.061)			
Math	0.026	0.025	-0.012	0.804	0.581	0.560
	(0.032)	(0.051)	(0.053)			
Kiswahili	0.033	-0.007	0.012	0.595	0.757	0.754
	(0.037)	(0.048)	(0.050)			
F-test of joint significance (p-value)				0.671	0.865	0.273
N	3325	1605	1659			
Clusters	197	98	98			

Notes. Balance statistics on reduced samples, based on data availability at midline (headteacher, classroom observations, and student-level data) and endline (student-level data). See Table B.4

Table B.6: Balance Tests on Reduced Sample (non-missing at Endline)

	(1)	(2)	T-test
	Visit	Visit&Text	P-value
Student			
Variable	Mean/SE	Mean/SE	(1)-(2)
Male	0.498	0.495	0.906
	(0.014)	(0.013)	
Age	8.960	9.034	0.423
	(0.061)	(0.064)	
Math	0.032	-0.009	0.556
	(0.054)	(0.054)	
Kiswahili	-0.007	0.021	0.778
	(0.051)	(0.051)	
N	1440	1456	
Clusters	88	88	
WEO			
Age	43.022	43.011	0.935
	(0.716)	(0.759)	
University Degree	0.843	0.921	0.241
	(0.039)	(0.029)	
Female	0.236	0.281	0.525
	(0.045)	(0.048)	
Years in current position	2.449	2.899	0.176
	(0.298)	(0.329)	
F-test of joint significance (p-value)			0.301
N	89	89	
Teacher			
Age (in 2021)	38.387	38.638	0.947
	(0.454)	(0.415)	
Male	0.699	0.691	0.824
	(0.027)	(0.031)	
F-test of joint significance (p-value)			0.974
N	615	602	
Clusters	89	88	
Classroom Observations			
Teaching Quality Index	0.115	0.004	0.222
	(0.109)	(0.093)	
N	58	63	
Clusters	43	50	

Notes: The value displayed for t-tests are p-values. The value displayed for F-tests are p-values. Standard errors are clustered at variable SchoolID. Fixed effects using variable DistrictID_old are included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Notes.

Table B.7: Teacher-level attrition analysis

	Midline			Endline		
	(1) Attrite	(2) Male	(3) Age	(4) Attrite	(5) Male	(6) Age
Visit	0.001	-0.002	0.836	0.018	-0.011	0.324
	(0.019)	(0.031)	(0.551)	(0.019)	(0.032)	(0.554)
Visit&Text	-0.029	-0.014	0.795	0.022	-0.006	0.311
	(0.020)	(0.033)	(0.504)	(0.019)	(0.034)	(0.513)
Attrite		0.001	0.386		0.007	1.009
		(0.028)	(0.741)		(0.027)	(0.734)
Attrite x Visit&Text		-0.036	0.451		-0.054	1.946
		(0.049)	(1.277)		(0.049)	(1.139)
Attrite x Visit		0.041	-0.845		0.067	1.066
		(0.053)	(1.234)		(0.049)	(1.239)
Control Mean	0.233	0.518	37.510	0.255	0.493	37.510
N	3039	3039	3039	3039	3039	3039
R-squared	0.042	0.065	0.052	0.028	0.066	0.059

v

Notes: Attrition is a dummy variable equal to one if a teacher was not surveyed at midline/endline. * for p<.1; ** for p<.05; *** for p<.01. Estimates include strata fixed effects.

Table B.8: Classroom observation attrition analysis

	Midline		Endline	
	(1)	(2)	(3)	(4)
	Attrite	Teaching Quality Score	Attrite	Teaching Quality Score
Visit	-0.002	-0.040	-0.009	0.055
	(0.043)	(0.128)	(0.044)	(0.136)
Visit&Text	-0.032	-0.187	-0.013	-0.007
	(0.044)	(0.134)	(0.042)	(0.135)
Attrite		-0.162		-0.012
		(0.099)		(0.098)
Attrite x Visit		0.088		-0.073
		(0.153)		(0.152)
Attrite x Visit&Text		0.178		-0.128
		(0.161)		(0.150)
Control Mean	0.564	0.000	0.629	0.000
N	775	775	775	775
No. schools	396	396	396	396
R-squared	0.066	0.185	0.066	0.183

Notes: Attrition is a dummy variable equal to one if a teacher was not observed at midline/endline. * for p<.1; ** for p<.05; *** for p<.01. Estimates include strata fixed effects.

Table B.9: WSV report score and student value-added (baseline to midline)

	(1)	(2)	(3)	(4)
Overall Quality Score1	0.143***	0.062***		
	(0.045)	(0.022)		
Weak			-0.205**	-0.149*
			(0.089)	(0.089)
Good			0.199**	0.036
			(0.092)	(0.042)
Very Good			1.116***	0.375
			(0.400)	(0.324)
Baseline Controls	No	Yes	No	Yes
R-squared	0.023	0.625	0.033	0.625
Observations	1907	1907	1907	1907

Notes: Data restricted to 113 schools that had received a Whole-School Visit after baseline data collection, but more than six months before endline data collection. The dependent variable is the combined IRT score for student performance in Kiswahili and Mathematics at midline. The control variables include baseline performance in Mathematics and Kiswahili, and student age and gender. Columns (2) and (4) thus show student value-added. The variable "Overall Quality Score" is the weighted average of the six different domain scores, which was reported by government. A higher weight (30 percent) is given to Teaching&Learning and Leadership&Management; and the lowest weight (10 percent) given to curriculum. School environment and community engagement are weighted at 15 percent. The independent variables in the second and fourth columns are dummy variables indicating differences in the overall score given by the SQAOs. This score ranges from 1 (unsatisfactory) to 5 (very good). The omitted category is 3 (Satisfactory)

Table B.10: Student learning at endline—Intent to Treat estimates, full sample

	Intent to Treat			
	(1)	(2)	(3)	(4)
	Combined	Math	Kiswahili	English
Visit	0.01	0.03	0.02	-0.02
	(0.04)	(0.04)	(0.04)	(0.05)
Visit&Text	0.08*	0.03	0.09**	0.07
	(0.04)	(0.05)	(0.04)	(0.06)
Visit=Visit&Text	0.083	0.957	0.023	0.046
Control mean	0.71	1.07	0.79	0.58
Observations	6481	6481	6481	6481
No. of schools	393	393	393	393

Notes: Each column represents a separate OLS regression on the full sample of schools, using the following estimating equation: $y_{i,s} = \beta_0^F + \beta_1^F(\text{Visit}_s^Z) + \beta_2^F(\text{Visit\&Text}_s^Z) + \alpha_d + X'_{i,s}B + \epsilon_{i,s}$. Aggregate scores in Math, English, and Kiswahili are constructed using Item Response Theory, and standardized to have control mean of zero and SD of one. Control variables are student gender, age and performance in their baseline IRT score in Math, Kiswahili, and English (if applicable). * for p<.1; ** for p<.05; *** for p<.01.

Table B.11: Classroom Observations Data

	Overall		Classroom culture		Instruction		Time on task—high	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Midline (full sample)								
Visit	0.14	0.20	0.19***	0.27***	0.06	0.01	-0.03	-0.00
	(0.12)	(0.18)	(0.07)	(0.09)	(0.09)	(0.12)	(0.05)	(0.07)
Visit&Text	0.27**	0.37***	0.11*	0.15*	0.14**	0.12	0.11**	0.16***
	(0.11)	(0.14)	(0.06)	(0.08)	(0.07)	(0.09)	(0.05)	(0.06)
Replacements included?	Y	N	Y	N	Y	N	Y	N
Visit=Visit&Text	0.19	0.19	0.13	0.13	0.28	0.25	0.01	0.00
Control Mean	0.00	0.00	3.32	3.32	2.57	2.57	0.58	0.58
Observations	1044	684	1044	684	1044	684	1020	670
No. schools	343	262	343	262	343	262	343	261
R-Squared	0.133	0.138	0.180	0.166	0.129	0.137	0.144	0.141
Panel B. Endline (reduced sample)								
Visit&Text	-0.11	-0.25	-0.09*	-0.19**	-0.00	-0.03	-0.04	-0.06
	(0.11)	(0.16)	(0.06)	(0.08)	(0.06)	(0.10)	(0.04)	(0.05)
Replacements included?	Y	N	Y	N	Y	N	Y	N
Control Mean	0.05	0.16	3.56	3.59	2.74	2.80	0.65	0.67
Observations	570	266	570	266	570	266	550	260
No. schools	178	102	178	102	178	102	178	102
R-Squared	0.126	0.176	0.121	0.179	0.146	0.188	0.083	0.176

Notes: TEACH classroom observation instrument gives a teacher a score between one and four across different dimensions of teaching quality. Classroom culture is the average of the average scores for ”supportive learning environment” and ”positive behavioral expectations”. Instruction us the average of the average scores for ”lesson facilitation”, ”check for understanding”, ”feedback”, and ”promote critical thinking”. Enumerators also reported at three different points in time how many children were on task, and coded it as high if fewer than two students are off task. We took the average of this binary indicator. Two classroom observations were performed per teacher, so data is at the teacher-by-observation level. Standard errors are clustered at a school level. In the even-numbered columns the sample is restricted to a panel of teachers who were observed at both baseline and the relevant post-treatment round (midline in Panel A, endline in Panel B).

Table B.12: Number of exercises by subject

	(1)	(2)	(3)
	English	Math	Kiswahili
Panel A. Midline (full sample)			
Visit	0.015	-0.238*	0.136
	(0.102)	(0.127)	(0.104)
Visit&Text	-0.061	0.107	0.309***
	(0.118)	(0.125)	(0.117)
Visit=Visit&Text	0.541	0.022	0.168
Control mean	1.459	2.013	1.547
Observations	1302	1326	1347
Panel A. Endline (restricted sample)			
Visit&Text	-0.041	0.321**	0.077
	(0.134)	(0.158)	(0.159)
Visit mean	1.378	1.563	1.600
Observations	598	587	591

Notes: The dependent variable is the number of days in the last five school days that an exercise was completed, by subject. A random sample of students were selected for document inspection—some for English, some for Mathematics, and some for Kiswahili. All estimates are Intent to Treat. For the final three columns (endline) the same is further restricted to schools in the two treatment arms that had received a WSV by the time of midline data collection.

Table B.13: Students' perception of teaching quality

	(1) Explain	(2) Motivate	(3) Discuss	(4) Tests	(5) Homework	(6) Feedback	(7) Remedial
Panel A. Midline							
Visit	0.003	0.003	-0.022	0.003	-0.000	-0.004	0.007
	(0.009)	(0.005)	(0.016)	(0.011)	(0.024)	(0.022)	(0.026)
Visit&Text	-0.003	0.002	-0.017	0.015	0.022	-0.012	-0.000
	(0.010)	(0.005)	(0.018)	(0.012)	(0.023)	(0.025)	(0.025)
Visit=Visit&Text	0.556	0.748	0.812	0.355	0.415	0.768	0.818
Control Mean	0.925	0.966	0.843	0.926	0.448	0.714	0.584
Observations	6397	6397	6397	6397	6397	6397	6397
R-Squared	0.028	0.013	0.044	0.029	0.101	0.084	0.061
Panel B. Endline							
Visit&Text	0.012	0.012	0.050**	-0.015	0.064**	0.098***	0.104***
	(0.010)	(0.007)	(0.020)	(0.011)	(0.030)	(0.025)	(0.030)
Visit Mean	0.925	0.966	0.843	0.926	0.448	0.714	0.584
Observations	2846	2846	2846	2846	2846	2846	2845
R-Squared	0.025	0.052	0.039	0.038	0.114	0.060	0.080

Notes: .

Table B.14: Head teacher beliefs—room for improvement

	Management			Teaching			Environment			Community	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Leadership	Monitoring	Curriculum guidance	Attendance	Preperation	Teaching	Hygiene	Functional toilets	Clean toilets	Pupil attendance	Community engagement
Panel A. Midline											
Visit	0.113	-0.062	-0.110	-0.247	-0.120	-0.031	0.081	0.088	0.032	-0.080	0.241
	(0.159)	(0.157)	(0.157)	(0.190)	(0.149)	(0.135)	(0.149)	(0.159)	(0.128)	(0.125)	(0.163)
Visit&Text	-0.044	-0.270**	-0.180	0.279*	-0.056	-0.025	-0.024	-0.115	-0.060	-0.060	-0.044
	(0.131)	(0.135)	(0.139)	(0.169)	(0.127)	(0.120)	(0.124)	(0.125)	(0.118)	(0.115)	(0.134)
Visit=Visit&Text	0.270	0.139	0.617	0.002	0.633	0.964	0.437	0.161	0.448	0.865	0.056
Control Mean	2.115	2.192	2.188	2.461	2.176	2.005	1.933	1.845	2.041	1.917	1.870
Observations	384	386	385	387	387	386	387	387	387	387	387
R-Squared	0.109	0.163	0.110	0.236	0.227	0.136	0.142	0.155	0.272	0.211	0.094
Panel B.Endline											
Visit&Text	0.068	-0.132	-0.042	-0.289*	-0.033	-0.032	0.296**	0.394***	0.112	-0.084	0.124
	(0.141)	(0.136)	(0.140)	(0.155)	(0.133)	(0.129)	(0.120)	(0.123)	(0.137)	(0.127)	(0.124)
Control Mean	2.292	2.236	2.258	2.629	2.202	2.034	1.753	1.573	1.899	1.888	2.011
Observations	177	177	177	177	177	177	177	177	177	177	177
R-Squared	0.298	0.325	0.232	0.362	0.292	0.225	0.305	0.276	0.173	0.176	0.345

Notes: See Table 5

Table B.15: Monitoring by the head teacher

	(1) Overall	(2) Inspect homework	(3) Inspect assessment	(4) Inspect Class journal
Panel A. Midline				
Visit	-0.055	-0.017	-0.041	0.014
	(0.114)	(0.043)	(0.041)	(0.057)
Visit&Text	0.079	0.024	0.003	0.059
	(0.096)	(0.039)	(0.036)	(0.048)
Visit=Visit&Text	0.224	0.337	0.257	0.404
Control Mean	0.000	0.473	0.555	0.435
Observations	2357	2357	2357	2357
R-Squared	0.122	0.118	0.072	0.148
Panel B. Endline				
Visit&Text	0.006	0.005	0.008	-0.008
	(0.083)	(0.035)	(0.031)	(0.044)
Visit Mean	0.087	0.579	0.690	0.598
Observations	1092	1092	1092	1092
R-Squared	0.138	0.099	0.084	0.158

Notes: See Table 6. All dependent variables are at a teacher level. Dependent variables are binary variables equal to one if school management inspected the following documents in the same year as the data collection: (i) student homework, (ii) student assessment, and (iii) class journal. The combined score is the mean of the binary variables standardized to have a control mean of zero and standard deviation of one.

Table B.16: Curriculum guidance by the head teacher

	(1)	(2)	(3)	(4)	(5)	(6)
		Follow-up	Follow-up	Observed	High level	High quality
	Overall	lesson plan	scheme of work	teaching	support	Feedback
Panel A. Midline						
Visit	0.081	0.072	0.063	-0.020	0.071	-0.056
	(0.097)	(0.046)	(0.050)	(0.040)	(0.063)	(0.039)
Visit&Text	0.011	0.052	0.005	-0.013	0.072	-0.068*
	(0.086)	(0.040)	(0.041)	(0.034)	(0.057)	(0.037)
Visit=Visit&Text (p-value)	0.449	0.640	0.209	0.858	0.990	0.754
Control Mean	0.000	0.360	0.358	0.365	3.822	1.827
Observations	2369	2369	2369	2369	2369	2357
R-Squared	0.045	0.046	0.077	0.033	0.044	0.057
Panel B. Endline						
Visit&Text	-0.093	-0.045	-0.068*	-0.002	0.003	-0.005
	(0.081)	(0.038)	(0.040)	(0.033)	(0.049)	(0.038)
Visit Mean	0.060	0.390	0.401	0.395	3.853	1.811
Observations	1092	1092	1092	1092	1092	1092
R-Squared	0.078	0.083	0.103	0.114	0.066	0.041

Notes: See Table 6. All dependent variables are at a teacher level. The dependent variables in columns (2)-(4) are dummy variables. Column (2): school leadership spoke to teacher about lesson plans, made recommendations in year of data collection, and followed up on recommendations. Column (3): school leadership spoke to the teacher in year of data collection about the scheme of work, made recommendations, and followed up. Column (4): school leadership observed teaching in year of data collection, made recommendations, and followed up on recommendations. Columns (5) and (6) are ordinal variables ranging from one to five. Column (5): how much a teacher agrees to the statement "school leadership provides high level of curriculum guidance, feedback and professional support". Column (6): how much a teacher disagrees with the statement: "I would like to receive more feedback about my teaching from my head teacher." The combined index is the mean of the z-scores of the respective indicators, standardized to have a control mean of zero and standard deviation of one.

Table B.17: Parental Contributions

	(1) School lunch	(2) PTA met in 2020	(3) Parent contributions	(4) SMC meetings
Panel A. Midline				
Visit	0.144**	0.045	-0.011	-0.424*
	(0.067)	(0.052)	(0.030)	(0.246)
Visit&Text	0.090	-0.010	-0.017	-0.295
	(0.061)	(0.043)	(0.026)	(0.234)
Visit=Visit&Text (p-value)	0.423	0.242	0.834	0.588
Control Mean	0.355	0.103	0.084	4.684
Observations	393	389	397	387
R-Squared	0.420	0.168	0.141	0.145
Panel B. Endline				
	(1) School lunch	(2) PTA met in 2020	(3) Parent contributions	(4) SMC meetings
Visit&Text	0.084	0.050	0.020	0.002
	(0.060)	(0.066)	(0.018)	(0.187)
Visit=Visit&Text (p-value)				
Control Mean	0.371	0.511	0.041	4.573
Observations	178	177	178	178
R-Squared	0.381	0.281	0.146	0.212

Notes:

Table B.18: School Environment

	(1) Toilet:Student Ratio	(2) Clean Toilets	(3) Good state Classrooms
Panel A. Midline			
Visit	-0.000	-0.024	0.001
	(0.001)	(0.133)	(0.001)
Visit&Text	-0.001	-0.057	-0.001
	(0.001)	(0.117)	(0.001)
Visit=Visit&Text (p-value)	0.725	0.797	0.120
Control Mean	0.024	2.734	0.017
Observations	393	397	393
R-Squared	0.913	0.227	0.755
Panel A. Endline			
Visit&Text	-0.000	-0.164	0.003
	(0.001)	(0.101)	(0.002)
Control Mean	0.026	2.843	0.015
Observations	178	178	93
R-Squared	0.888	0.253	0.786

Notes: .

Table B.19: Classroom Observations, by Equip-T

	(1)	(2)	(3)	(4)
	Overall	Classroom culture	Instruction	Time on task—high
Panel A. Midline				
Visit (β_1)	0.11	0.13*	0.01	0.00
	(0.14)	(0.07)	(0.09)	(0.05)
Visit × Equip-T (β_2)	0.03	0.06	0.03	-0.04
	(0.20)	(0.11)	(0.13)	(0.08)
Visit&Text (β_3)	0.18	0.06	0.05	0.08
	(0.12)	(0.07)	(0.08)	(0.05)
Visit&Text × Equip-T (β_4)	0.07	-0.06	0.10	0.04
	(0.18)	(0.11)	(0.12)	(0.08)
Equip-T (β_5)	-0.00	0.15**	0.01	-0.14***
	(0.11)	(0.07)	(0.07)	(0.05)
$\beta_3 + \beta_4 = 0$	0.07	0.95	0.08	0.05
Observations	1044	1044	1044	1020
R-Squared	0.019	0.032	0.011	0.054
Panel B. Endline				
Visit&Text (β_1)	-0.22	-0.18**	-0.12	0.00
	(0.17)	(0.09)	(0.11)	(0.06)
Visit&Text × Equip-T (β_2)	0.20	0.16	0.25*	-0.10
	(0.25)	(0.12)	(0.15)	(0.08)
Equip-T (β_3)	0.07	0.14*	-0.08	0.00
	(0.16)	(0.08)	(0.10)	(0.05)
$\beta_1 + \beta_2 = 0$	0.92	0.85	0.19	0.11
Observations	570	570	570	550
R-Squared	0.014	0.050	0.014	0.013

Notes: See Table B.11 for a description of dependent variables, and Table 8 for description of empirical strategy and interaction terms.

Table B.20: Teachers, enrollment, teacher-pupil-ratios

	Teachers			Pupils			Teacher-pupil ratio		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	1 to 3	4 to 6	7	1 to 3	4 to 6	7	1 to 3	4 to 6	7
Visit	-0.041	-0.734*	-0.169	-5.178	3.233	2.005	-0.002	-0.010***	-0.011*
	(0.205)	(0.401)	(0.190)	(5.583)	(4.029)	(1.239)	(0.001)	(0.003)	(0.006)
Visit&Text	-0.099	-1.071***	-0.197	-1.162	2.701	1.069	-0.004***	-0.006**	-0.004
	(0.191)	(0.384)	(0.183)	(5.505)	(3.742)	(1.157)	(0.001)	(0.003)	(0.006)
F-Test	0.798	0.471	0.900	0.524	0.895	0.480	0.219	0.212	0.313
Control Mean	6.952	14.155	4.781	321.807	286.360	61.914	0.035	0.074	0.115
Observations	372	388	353	393	393	393	372	388	352
R-Squared	0.495	0.455	0.294	0.968	0.981	0.965	0.841	0.805	0.636

Notes:

Table B.21: Teachers, enrollment, teacher-pupil-ratios (endline)

	Teachers			Pupils			Teacher-pupil ratio		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	1 to 3	4 to 6	7	1 to 3	4 to 6	7	1 to 3	4 to 6	7
Treat1: SQA	-0.053	-0.141	-0.589***	4.465	5.454	3.767**	-0.003*	-0.011**	-0.025**
	(0.204)	(0.412)	(0.205)	(11.391)	(5.994)	(1.886)	(0.002)	(0.005)	(0.010)
Treat2: SQA+	-0.196	-0.647	-0.450**	5.170	-8.333	1.335	-0.003*	-0.007	-0.006
	(0.191)	(0.406)	(0.196)	(9.843)	(5.443)	(1.805)	(0.002)	(0.004)	(0.009)
F-Test	0.518	0.298	0.563	0.957	0.063	0.254	0.879	0.438	0.108
Control Mean	6.432	12.889	4.302	322.553	271.106	59.426	0.034	0.077	0.113
Observations	397	397	397	197	197	197	197	197	196
R-Squared	0.345	0.210	0.276	0.956	0.981	0.962	0.877	0.838	0.638

Notes:

Table B.22: WEO activities when visiting schools and actions in response to the WSV—by Equip-T

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Typical activities when visiting a school							
	Total	Inspect document	Interact teachers	Interact students	Talk to parents	Check: WSDP	Check: ISQAT
Visit&Text	0.474	0.177	0.068	0.049	-0.083	0.173*	0.076
	(0.454)	(0.196)	(0.200)	(0.145)	(0.064)	(0.090)	(0.063)
Equip-T (γ_2)	-0.696	-0.033	-0.490**	-0.119	-0.134**	0.177*	0.027
	(0.482)	(0.219)	(0.223)	(0.157)	(0.059)	(0.106)	(0.063)
Visit&Text × Equip-T (γ_3)	-0.149	-0.193	0.197	-0.097	0.182**	-0.104	-0.044
	(0.646)	(0.318)	(0.304)	(0.217)	(0.089)	(0.147)	(0.094)
Equip-T Mean (Visit&Text=0)	4.795	2.455	1.227	0.545	0.045	0.409	0.068
Observations	178	178	178	178	178	178	178
$\gamma_1 + \gamma_3 = 0$	0.478	0.949	0.249	0.766	0.112	0.554	0.654
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel B. Actions in Response to WSV							
	Total	Report DED/DEO	Check Implementation	Lobby Resources	Training	Talk to parents	
Visit&Text	0.196	0.111	0.099	0.051	0.029	-0.059	
	(0.275)	(0.073)	(0.095)	(0.172)	(0.049)	(0.094)	
Equip-T (γ_2)	-0.124	0.214**	0.007	-0.121	-0.023	-0.111	
	(0.283)	(0.092)	(0.108)	(0.172)	(0.043)	(0.102)	
Visit&Text × Equip-T (γ_3)	0.114	-0.281**	-0.097	0.156	0.080	0.230	
	(0.425)	(0.124)	(0.150)	(0.260)	(0.080)	(0.145)	
Equip-T Mean (Visit&Text=0)	1.773	0.318	0.523	0.500	0.045	0.205	
Observations	178	178	178	178	178	178	
$\gamma_1 + \gamma_3 = 0$	0.341	0.092	0.986	0.291	0.085	0.123	

Notes: “Inspect document” is sum of the following four indicator variables: inspect document, inspect facilities, inspect teacher attendance, inspect lesson plans. “Interact teachers” is the sum of the following three indicator variables: Talk to teachers, observe teaching, give feedback to teachers.“Interact students” is sum of: talk to students and assess student learning. Lobby resources is the sum of: lobby the Village Authority for resources, lobby the Ward Council for resources, and request resources from the DEO.

Appendix C Qualitative Research

Appendix C.1 Phone calls to WEOs in program schools

We conducted phone surveys with all the WEOs who participated in the program to get a better idea of what actions they took in response to receiving the WSV, and get their inputs on how to improve the program. Most of the actions that they mentioned related to **discussing the recommendations** with the head teachers and teachers. Moreover the WEOs' suggestions on how to improve the program also pointed to potential mechanisms: saliency, reminders, and prioritization of a smaller range of tasks.

Appendix C.1.1 Typical actions in response to receiving the text message

Below are examples of the responses coded by the research assistant.

1. First had a meeting with the school management. Then had a meeting with the teachers to share the recommendations and how will be implemented.

2. Conducted a meeting with the teachers and school management to share and discuss the recommendations received;

3. Went to school and stayed with teachers giving directions and guidelines on what to do based on recommendations.

4. Forwarded the recommendations to teachers, then had a meeting with teachers how to discuss how to solve the problems;

5. Approached a specialist to conduct a seminar to teachers to train them on preparing lesson plans;

6. Consulted the head teacher and shared the recommendations. Then, also I had a meeting with the teachers to discuss all the recommendations together;

7. Also advised teachers not only to exert force to class IV and VII rather all classes;

8. Had a meeting first with the teachers to discuss recommendations related to teaching and learning, especially to use students participatory approach; Had a meeting with parents and

advised them to contribute on construction of toilets; emphasized teachers to make and use teaching and learning aids.

Appendix C.1.2 Recommendations for improvement

Some examples, as coded by the research assistant:

1. "They should be frequently reminded as they deal with a multiplicity of demands"

2. "He appreciates the program because it has enabled him to more quickly obtain the SQA recommendations"

3. "This should be a continuous exercise because we get connected in overcoming their challenges faced";

4. "it prompts teachers to be accountable of their responsibilities as we make follow ups"

5. "Its great we remind them on implementation";

6. "This program is very good as it reminds us of our responsibilities".

Appendix C.2 Regional Workshop Focus Group Discussions

Author: Anthony Mwabanga

Appendix C.2.1 Introduction

Since 2018, the RISE country team in Tanzania (RCT) has been evaluating Tanzania's new School Quality Assurance framework. The government adopted the new framework in 2017 as a means of replacing the old framework (School Inspection System) in order to improve educational outcomes with a more evidence-based approach. So, in March 2019, the RCT held the first round of a series of workshops in its sample regions of Pwani, Singida, Tanga, Songwe, Simiyu, and Kigoma as a preparation for the program's launch. Again from July 4th to July 28th, 2022, the RISE country team in Tanzania (RCT) conducted another series of workshops in the same sample regions with the aim of; first, disseminating the preliminary findings of the evaluation of the New School Quality Assurance Framework, and second, conducting a follow-up survey among the participants (DEOs, SQAOs, and WEOs) in order to better understand some of the preliminary findings that we have observed. So this report covers all of the events that occurred in all of the workshops in those six regions.

The workshops were generally successful, despite few issues. For example, in some regions, we were unable to obtain some of the participants, especially WEOs, owing to their participation in the national Census activity. We also had some issues with participants arriving late for various reasons. There was also a communication breakdown between the REO and the participants (as a formality was to depend on the REOs to communicate everything to the participants), and as a result, some of the participants received the invitations late, and, in some regions, invalid participants were invited, especially in Kigoma.

Appendix C.2.2 Methodology

These workshops were held for one day in each of the aforementioned regions from July 4-28. The target audience for the workshops was purposefully chosen, as we aimed at REOs, DEOs, SQAOs, and WEOs, whose districts and wards were chosen randomly at the beginning of the evaluation to represent their respective areas. The total number of participants for all workshops was 130 (Annex 1 has a list of all the participants from each region). We had 29 participants from Pwani, 15 from Singida, 12 from Songwe, 28 from Tanga, 23 from Simiyu, and 23 from Kigoma. As mentioned above, the aim was to share the research findings and also to get a better understanding of some of the preliminary findings, so we prepared a few questions that were supplied in the form of a normal group discussion and a focus group discussion with the 1-2-all technique.

The questions asked were;

1. Why did the text messages work?

2. Why were the Equip-T regions different from others?

Appendix C.2.3 Presentation and Discussion

The presentation of the research findings typically lasted 30 to 45 minutes, followed by a period of discussion in which questions were asked and various points of view were expressed. In general, the study's findings were well received, and for more than three regions, they stated that they were completely consistent with the findings of their respective regions. Singida, for example, stated that the regional mock results for the subject of mathematics were 27%, while the average for the English subject was 33%. Another finding that all participants agreed on was the difference in performance between schools in Equip-T regions and those in non-Equip-T regions.

Appendix C.2.4 Group Discussions

Participants were arranged into groups depending on their titles and the types of questions to which they were supposed to respond. In the first round of discussion, which was a focus group discussion (FGD), we had WEOs divided into two groups; those who received text messages and those who did not. Then we also had a group of SQAOs and a group of DEOs, but in some regions, SQAOs and DEOs were mixed in the same group depending on the number of them. For the second round of discussion, we had all the participants divided into either two or three groups, and it was a normal group discussion in which the participants discussed themselves and selected a member to present on their behalf. In some regions, we noticed that some participants were uncomfortable expressing their opinions, specifically when mixed with REOs or RAOs. For example, in Tanga, we had to tell the RAO in a polite way to give the participants in his group space so that they could express themselves freely.

Appendix C.2.5 Takeaways From The Three Main Questions

For almost all the questions, the answers were somewhat close to each other, but the question about the difference in performance between Equip-T and non-Equip regions had very similar answers for almost all the regions. Below are the responses of the participants to each question posed, including all answers in the sense of focus group discussion and those in the sense of normal group discussion.

1. Why did the text messages work?

Most of the responses here were the same across all the regions. For this specific question, we divided the WEOs into two groups, i.e., those who received texts and those who did not, as we wanted to get the real experience of what happened during the intervention.

For those WEOs who participated in the program their responses were as follows:

- The text was delivered to them earlier than the full report, so they were able to visit the "text school" earlier than the other schools which received the WSV. Additionally, because the text was more detailed, he was able to sit down with the h/teacher to discuss the recommendation.

- The text was putting more emphasis and it was like a push factor since it has just a few tasks to work on while other visited schools were having general recommendations
- It facilitates communication between the headteacher and the WEO and it makes it easier to follow up on issues that have been worked upon and issues that were in progress also phone calls were involved and it put more emphasis
- The text made the WEO make multiple visits to that specific school compared to other schools since he was following up on the specified tasks and someone was also calling him to ask about the progress
- A normal WSV report sits in a file at school or in my office, but with the texts, I can move with them anywhere, hence simple in following up on the recommendations
- Text worked because RISE schools were given incentives in the form of money for head teachers and sugar for teachers' tea, but they also knew that they were not alone in the project, so if they slacked off, they would be last in performance.

For those who did not, their responses were as follows:

- Communication through text messages reaches the person on time and enables him to work on time, unlike communication through letters, where sometimes it takes a long time to reach the person or sometimes it may not reach the person at all
- The text worked as a reminder, as WEOs are always busy, and it might happen that a school may be visited while a WEO is not around, so a text helped them to keep track of the recommendations, hence increasing efficiency.
- The message summarizes the recommendations so that it is easy to follow up. Also, the message increases the effectiveness of the implementation because when the WEO receives the message, he knows that there are people monitoring his performance and, therefore, he should work hard to fulfill the responsibilities.

2. Why improvement in Equip-T regions?

In our sample of six regions, we had three that are part of the Equip-T program (Singida, Simiyu, and Kigoma) and three that are not (Tanga, Pwani, and Songwe). We were trying to figure out what might have caused more improvement in those regions compared to those that did not have the program, so we basically asked a direct question on why positive improvement and another question that attempted to understand the communications aspects for all of the education officers involved. The responses are as follows:

- There were community teachers who were collaborating with normal teachers to help students, especially in schools with a small number of teachers.
- Teachers were trained in various areas of the 3Rs, including how to teach difficult topics, for example in the mathematics subject.
- The Equip-T program also supported the creation of the Parents and Teachers Union (UWW), where they worked to raise children. The union includes males and females. teachers and male and female parents work together to help children solve any challenges they face at school or at home.
- It's because of the continuous training of teachers from time to time as well as the monitoring of the presence of financial incentives for teachers who do well and the provision of tools such as motorcycles and tablets.
- With close supervision, teachers were trained in using better teaching methods, improvements in communication eg, and the use of KOBO Kollect by using tablets
- Provision of working instruments such as motorcycles, fuel, and maintenance.
- The program helped children who came from outlying areas get an education in their areas. They created centers that provided education and took children from the age of 4-5 and gave them initial education before they went to primary school. Later, they built classrooms in those centers and they became primary schools, the director assigned one teacher at each center to be able to teach children at each center.

- The program also helped to establish income sources in the school where they wrote business proposals and got 1.5 million TZS to carry out the projects and get the money from the businesses, as the benefit was used to improve the school environment and provide school needs to those who could not afford like a school uniform.

Appendix C.2.6 DEOs' Communication With The SQAOs/WEOs; Equip-T Regions Vs. Non-Equip-T Regions

Here we were trying to understand if there is proper communication flow between the DEOs, SQAOs, and WEOs in general in the implementation of the SQA framework, especially in facilitating the actioning of the recommendations provided, so we are comparing how all those government officials involved in both regions behave.

Responses from the Equip-T regions (Kigoma, Simiyu, and Singida)

From the DEOs:

- DEO did not talk to them about WSV, they usually do not talk to them about the WSV report. After WSV, they received a report from DSQAOs and they started working on the recommendation.

- DEO do not talk to them about the WSV recommendations they get the information from the log book and from the SQAOs.

- They neither talk physically nor through the phone about the reformations. If it happens that WEO did not attend the WSV, they will get all the recommendations from the school and not from DEO.

- Sometimes it is even hard to get the recommendations from the SQAO; they only get them from the school head teachers or from the log book. WEOs are close to schools, and they work with schools more closely than they work with DEOs.

- The DEO usually knows that the WEO has the ability to work on the WSV report, so there is no need to follow up and discuss the WSV recommendations unless otherwise. They are those recommendations that the WEO cannot work on without the help of the DEO, but if they are in his power, then WEOs do not need to follow up with the DEO.

From the SQAOs:

- Yes, SQAOs communicate directly with the WEOs, and they do so through phone calls, council meetings, and school visits.
- SQAOs also communicate using letters, especially in the matter of implementation or sending information to the schools.
- Yes, they mostly communicate after WSV. Often, SQAOs talk about what they observed on WSV, and communication is usually done by phone.
- Others said communication is very rare and they only communicate during the WSV period and after the WSV.
- SQAOs send their calendars to WEOs whose schools are on the visit schedule so that they can prepare for the visit. And there is also communication during the special visit for the preparation and that visit as well as sharing of the visit report.
- Now they use self-elevation SQAO to communicate with WEO and share a self-evaluation form so they can take the form to school and the schools to evaluate themselves before the whole school visit. They will also communicate with WEO to share a copy of the WSV reports with schools and the DED.

Responses from the Non-Equip-T regions (Pwani, Songwe, and Tanga) DEOs

- The answer is no, for many of the WEOs though they claimed that DEOs do talk to them about school improvement in general.
- Some of the WEOs responded. Yes, the DEO is talking to the WEO in relation to WSV, and among the things they discuss include students who do not know the 3Rs and strategies to improve the situation. Talking to parents about their children's absenteeism; lesson notes prepared by the teacher for teaching are not enough; evaluation of teaching and learning in schools, and maintenance of the school environment.
- Some WEOs responded that NO, the DEO did not talk to them regarding the WSV reports.

- Communicated to WEOs orally by phone, by writing letters to WEOs, by conducting monitoring visits, and through meetings with WEOs.

- The oral communications were always targeted at the burning issues that simply could not wait for a written letter to the WEOs e.g. a collapsing school toilet.

- They communicate about all the recommendations that are discussed in the respective schools and have a summation meeting where the WEO will present the feedback from the meeting to the DEO and also discuss with the DEO all the recommendations made and how to work on them.

SQAOs

- Yes, though it is not frequently to some WEOs, it mostly happens during or after the WSV in their schools, and it is mostly for collecting the WSVs reports or about the School Self Evaluation forms

- Someone said in the past 6 months she talked to the SQAOs once and they went to her schools to insist on teachers making sure that they cover the syllabuses

- Yes, though it is not frequently to some WEOs, it mostly happens during or after the WSV in their schools, and it is mostly for collecting the WSVs reports or about the School Self Evaluation forms

- Someone said in the past 6 months she talked to the SQAOs once and they went to her schools to insist on teachers making sure that they cover the syllabuses

- They all responded, ”YES.” SQAOs have direct communication with WEOs, and they mostly communicate when they need to come to the school for the WSV. They also communicate when they issue a visit letter or visit certificates in schools. When there are shortcomings in the school, they communicate to discuss how to improve and remove the limitations. Sometimes when SQAOs want to give congratulations to schools that have done well by giving out congratulations cards, notebooks, and pens, they contact the relevant WEO of those schools before sending the congratulations to the school.

Appendix C.3 Conclusions and Recommendations

In conclusion, despite a few flaws, we had successful workshops, and in the end, we were able to get some consistent and fruitful responses to some of the questions, such as why the text messages worked, why there was more improvement in Equip-T regions compared to non-Equip-T regions, and so on. Unfortunately, we were unable to obtain convincing responses that told us what might have caused the decrease in English performance because the responses provided were insistent that it is a persistent and well-known problem, but since the decrease was observed after the WSV, the causes may not be far away from it.