School Governance Reform at Scale—Experimental Evidence from Tanzania

Jacobus Cilliers and James Habyarimana

June 12, 2021

We report on a randomized evaluation of a school management reform program in Tanzania, rolled out to all schools in the country over a period of two years. Government officers (previously known as school inspectors) visit schools and produce a set of diagnostics and recommendations to improve school quality that are shared with all stakeholders. We evaluate the program in a nationally representative sample of 397 schools, with 199 schools assigned to the control. In a subset of treated schools we encourage additional monitoring by the local government front-line education officers by short-circuiting the information flow between the two separate ministries. We document three main findings from the midline survey. First, head teachers exposed to the additional monitoring changes their beliefs: they revised downwards their beliefs about the quality of school leadership. There were no commensurate changes in beliefs about the quality of teaching or extent of community engagement. Second, teacher presence increased by 7.9 percentage points and teaching practice improved. We find no evidence of improvements in school management, community involvement, nor the overall quality of the school environment. Third, there was a modest improvement in student learning of 0.05SD in Kiswahili, but no improvements in mathematics.

1 Introduction

Learning outcomes in a large number of education systems in developing countries are abysmally low. For example, 98% of a sample of 15 year olds in Zambia cannot correctly answer a simple mathematical problem such as converting a price from one currency to another (OECD (2019)). In part, these very low learning levels are a result of a remarkable expansion in access to schooling with many systems doubling in size over the last two decades (Pritchett (2013)). The growth of other inputs, including teachers and instructional materials, has not kept up leading to larger class sizes and a difficult learning environment (Duflo et al. (2015)). Other potentially related features of these systems include an ambitious curriculum, low teacher competence and effort (Bold et al. (2017)). While these challenges can be addressed in the long run through the selection, pre-service training and deployment of a sufficient number of effective teachers, a key question is how the transition from a focus on access to quality can be supported in the short run. This paper documents the short term impacts of a reformed school inspection program in Tanzania targeting management and teaching practices. Most school systems across the developed and developing world rely on school inspectors to provide quality control, often focusing on the quality of school management. Theoretically school inspectors produce information, prompts and facilitate incentives to shift school management and teacher behavior. Given the observed low levels of teacher and head teacher competence, there is potential scope for a school inspector to share new information about management and teaching practices. Even when the information is not new, a school inspector can help provide useful prompts to coordinate and focus teacher and head teacher expectations around system-wide objectives. This is particularly useful during a transition from one set of system goals to another. Finally, and perhaps most important, a school inspector facilitates incentives for head teachers and teachers to change behavior. Recommendations about classroom practice focus the monitoring actions of head teachers and inform target setting, and the recruitment and retention of teaching talent. Similarly, recommendations about the management of other non-classroom inputs can be used by parents and other stakeholders to hold head teachers accountable. Recommendations can also strengthen the claims of head teachers for greater support from higher government levels.

However, there are a number of reasons why school inspectors might not work at improving school quality. Fundamentally, inducing desirable behavior change relies on the quality of the information generated during the inspection – i.e. inspectors can (i) accurately diagnose the problem, (ii) set obtainable goals, and (iii) make feasible recommendations that address binding constraints – and the extent to which head teachers and teachers have incentives to actually take recommended actions. The inspectors' willingness and ability to improve school quality clearly interacts with school stakeholders' ability to make effective changes that will improve learning.[1]

This research is related to a growing literature examining the role of school based management, teacher feedback, and beyond education, the role of information and audits on bureaucrat and politician behavior. Bloom et al. (2015) found large variation in the management quality of schools, both between and within countries, which correlates with student performance. Blimpo et al. (2011) found that intensive training in school management (10-20 days), combined with grants to schools, led to a reduction in student and pupil absenteeism in Gambia. And other studies have found that changing school governance to allow more parental voice can improve student learning (Barr et al., 2012; Duflo et al., 2015).

A growing body of research finds that interventions targeting teacher pedagogy including coaching (Cilliers et al. (2020)), feedback (Muralidharan and Sundararaman (2010)) and training (Piper et al. (2018) Evans and Popova (2016) Conn (2017)) can boost learning outcomes. Beyond education, other researchers have documented mixed evidence of the impacts of information and community monitoring (Bjorkman and Svensson (2009), Björkman Nyqvist et al. (2017) and Raffler et al. (2018), or the role of audits in shaping the behavior of politicians, bureaucrats or firms (Avis et al. (2018), Olken (2007) and Duflo et al. (2013)).

Closer to the questions asked in this paper, there is evidence to suggest that inspections by themselves are not sufficient in improving performance. An evaluation of a School Quality Assurance program in India found that the program provided relevant and accurate information on school quality, but had no impact on metrics of school performance (Muralidharan and Singh, 2018). One possible reason is that there was no monitoring and follow up after the inspections. Outside of education, Callen et al. (2020) evaluate a program that reduced the cost of information flow between health clinic inspectors and senior policy makers in Pakistan. The intervention temporarily increased effort levels of inspectors but did not change the behavior of the inspected. Similarly,

[1] Stigler (1971) points out that a corrupt, incompetent or low effort inspector will not induce the required changes in behavior of the inspected.

in Nigeria, an intensive health management training, which included 9 months of implementation support, improved adoption of good practices, but only over the period of training. However, a less intensive version, which had limited oversight follow-up monitoring, had no impact (Dunsch et al., 2017). All of these studies reveal that monitoring and oversight are key to ensuring the adoption of improved practices. For example, there is evidence that local bureaucrats can improve performance through regular monitoring of schools. In India, an increase in the frequency of monitoring is correlated with lower teacher absenteeism (Muralidharan et al., 2016). Lavy and Boiko (2017) identified the contribution of local education officers (superintendents), who are responsible for a cluster of schools, on learning in Israel and concluded that one SD improvement in the bureaucrat improves student test scores by 0.04 SD. [2]

In Tanzania, the government has recently reformed its school inspection program, reframing the formerly punitive inclinations as a friendly source of support to schools now called the School Quality Assurance program. Under this new framework, headteachers complete a self-assessment form that informs a three day visit by School Quality Assurance Officers (SQAOs) to conduct interviews, document inspection, and observe teaching. SQAOs then provide an assessment of the overall quality of the school, as well as its performance in the key domains, which include: pupil learning, leadership and management, and teaching. This assessment is provided in the presence of other stakeholders including parents, local politicians and the local school administrator (Ward Education Officer (WEO)). The assessment provides a focal point for the WEO to follow up on mutually agreed actions during his regular visits to the school.

In this paper we report the mid-line results of a randomized phased-in of the School Quality Assurance program and a randomly assigned program to encourage increased monitoring and oversight. In particular, a random sample of study schools receive the a Whole School Visit by SQAOs. We refer to this group as *Visit*. A random sample of study schools, which we refer to as *Visit&Text*, receive both the whole school visit and the WEOs are encouraged through text-based reminders to engage head teachers on how they are addressing the recommendations of the SQAOs. Text messages include key recommendations and guidance to regularly follow up with schools and discuss the most important areas for improvement.

[2] Other studies including Banerjee et al. (2008) and Dhaliwal and Hanna (2017) find limited or no effects of increased monitoring in health facilities across two states in India

We document three main findings. First, head teachers in Visit&Text arm revised downwards their beliefs about the quality of school leadership at the start of 2019, prior to when most schools received the WSV. There were no commensurate changes in beliefs about the quality of teaching or extent of community engagement. Second, teacher behavior changed in the Visit&Text arm: teacher presence increased by 7.9 percentage points and teaching practice, as measured using the *Teach* classroom observation toolkit, improved. We find no evidence of improvements in school management, community involvement, nor the overall quality of the school environment. Third, there was a modest improvement in student learning of 0.05SD in Kiswahili, but no improvements in mathematics.

The rest of paper is organized as follows. Section 2 provides an overview of the program, section 3 discusses the sampling strategy and experimental design, section 4 describes the data collection, section 5 outlines the empirical strategy, section 6 summarizes the results, before concluding in section 7.

2 The Reform: School Quality Assurance

The Government of Tanzania has recently reformed its school inspection process, now called School Quality Assurance (SQA). The broader motivation for the reform is to shift the emphasis of the inspections away from being seen as a form of accountability, towards being a source of diagnostic feedback and support to schools. There is also a shift away from inspecting traditional inputs towards a focus on student learning, and improving teaching and management practices.

A key part of the reform is to replace the traditional school inspections with Whole School Visits (WSVs), which consist of the following steps. First, prior to the visit, schools are required to fill in a school self-evaluation form (SSEF). The form includes basic information such as enrollment, but also subjective self-assessments on the school quality. Second, a group of 3-4 School Quality Assurance Officers (SQAOs) visit a school for 2-3 days (depending on the size of the school). 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: (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. Together with this assessment they provide recommendations for each domain, as well as 3-4 main recommendations. The recommendations can apply to a wide set of school stakeholders: school leadership, teachers, parents, and community leaders. But the focus of the WSV is on school leadership and teaching quality.[3] At the end of the visit, there is an "exit meeting" where the SQAOs outline the main strengths of the school, areas for improvement, and make some concrete recommendations for improvement. The school leadership and staff attend the exit meeting, and parents and community leaders (e.g. Ward Councilor and religious leaders) sometimes also attend.

Third, subsequent to the visit the lead SQAO writes a short 4-5 page report including their quality rating for each domain, as well as recommendations for improvement. This report is shared with the head of local government business, the District Executive Director (DED), who again shares this with the officer who manages public education services in the district, the District Education Officer (DEO). In some cases, a very high level summary of these reports is produced by the Directorate of School Quality Assurance and shared with the supervisors of the DED and DEOs. In addition, the central government collates some of this information and creates a School Summary Report Card, which is sent back to the schools. The School Summary Report Card also includes information such as performance on the national exams and the availability and quality of facilities. Finally, according to the new SQA framework, the SQAOs are also required to perform follow-up visits in a subset of the schools, to make sure that the schools are implementing the recommendations. This rarely happened during the period of the evaluation because they were under pressure to meet their target of conducting WSVs in half the schools by July 2019.

2.1 Text messages to Ward Education Officers

An additional source of follow-up and ongoing support after completion of the WSV could be provided by the Ward Education Officer (WEO). There are roughly $4,000$ WEOs in the country who are typically responsible for only $4-5$ primary schools in their ward. The WEOs already visit schools in their ward on a regular basis, and according to the new SQA framework the WEOs are also supposed to follow up to see that schools are implementing the recommendations made during

[3] For example, the first activity they 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 WSVs. However, there is not a direct flow of information and chain of command between the SQAOs and WEOs. The WEOs report to the DEOs within the President's Office of Regional and Local Government (PO-RALG), and not the Department of School Quality Assurance, which falls under the Ministry of Education, Science and Technology (MoEST). The SQAOs send the WSV report to the DED, who then sends it down to the DEO. The DEO, in turn, decides what information to share with the WEOs. (see figure A.1). There is therefore a risk that the WEOs do not receive the report from the WSV, nor do they receive direct orders from the DEO to follow up with schools to make sure that they are implementing the recommendations.

With the purpose of addressing the above-mentioned institutional constraint, we implemented a low-cost program aimed at improving the frequency and focus of follow-up visits from WEOs in half the schools randomly assigned to receive a Whole School Visit. The Chief District SQAOs in each of the 23 districts in our evaluation sample would send us the WSV reports for all the schools in our evaluation sample on an ongoing basis. We then summarized and shortened the most important recommendations, and sent this summary to the WEOs over text messages. The WEOs also received multiple reminders over the course of the years, roughly once every two months. Since the WSVs were implemented at a staggered basis, not all WEOs received the same number of text messages: 74 received three messages, 5 received two messages, and 11 received one message.

It was important that the WEOs knew in advance about these messages, and that these messages were interpreted as directives from the DEO. We therefore held workshops, where we invited both the relevant WEOs and the DEOs from our evaluation sample. During these workshops we informed them of the program, and the DEOs expressed their official support and requested that the WEOs cooperate with us. The text messages were signed as coming from the District Office. In addition to the text messages, we endeavoured to survey all the WEOs participating in the program to learn from their experience in participating in the program. We surveyed 83 of the 90 possible WEOs.

2.2 Theoretical framework

Figure A.2 provides a schematic overview of the theory of change for this program, and Figure A.3 shows this in terms of a results framework. Whole School Visits could induce a change in behavior from three possible stakeholders: (i) the school leadership, who improve their management practices, (ii) teachers, who improve the quality of their teaching, and (iii) the school community, who contribute resources to the school. Behavioral change of these stakeholders could, in turn, improve student learning through the following channels:

1. The feedback to school leaders improves the management practices and curriculum guidance provided by the school leadership, which in turn improves teacher effort and teaching practice, which improves student learning.

2. The feedback to teachers improves teaching effort and teaching practices directly (i.e. not due to improved management practices), which improves student learning.

3. The feedback provided to parents induces parents to:

   (a) provide school lunches, which improves student learning.

   (b) encourage their children to attend school, which increases student attendance and thus student learning.

   (c) provide resources for school renovation and school construction. Improved classroom facilities leads to less disruption in the classroom and ultimately improves learning.

But there are severe binding constraints for a Whole School Visit to induce behavioral change. First, the SQAOs need to (i) accurately diagnose problems faced by the school, (ii) set realistic goals, and (iii) identify actions that school stakeholders can feasibly take to reach these goals. Second, the information garnered from the WSVs need to be new or become more salient to the stakeholders, and they need agree with the diagnosis/recommendation. Third, the stakeholders need to have sufficient *capacity* and *motivation* to change their behavior in response to the updated beliefs.

It is possible that regular monitoring by the WEOs, who follow up to see if school stakeholders are implementing the recommendations, can address some constraints to behavioral change. But this relies crucially on the actions taken by the WEO, and how responsive head teachers are to monitoring and feedback from the WEO. Figure A.4 provides a schematic summary of these necessary conditions.

3 Sampling and experimental design

The program is evaluated using a cluster randomized control trial, with a random phased-in design and randomization taking place at the Ward level. Prior to random assignment we performed stratified random sampling to make sure that our evaluation sample is representative of the country as a whole. We randomly selected one region in each of the six zones in the country, and then randomly selected roughly half of the districts in each of those regions, weighted by the number of schools in a district, yielding a sample of 22 districts and 413 Wards.[4] We took the additional step of excluding all primary schools in these wards that have already received WSVs, yielding a sample of 397 wards.[5] We then randomly sampled one school in each ward to participate in the study.

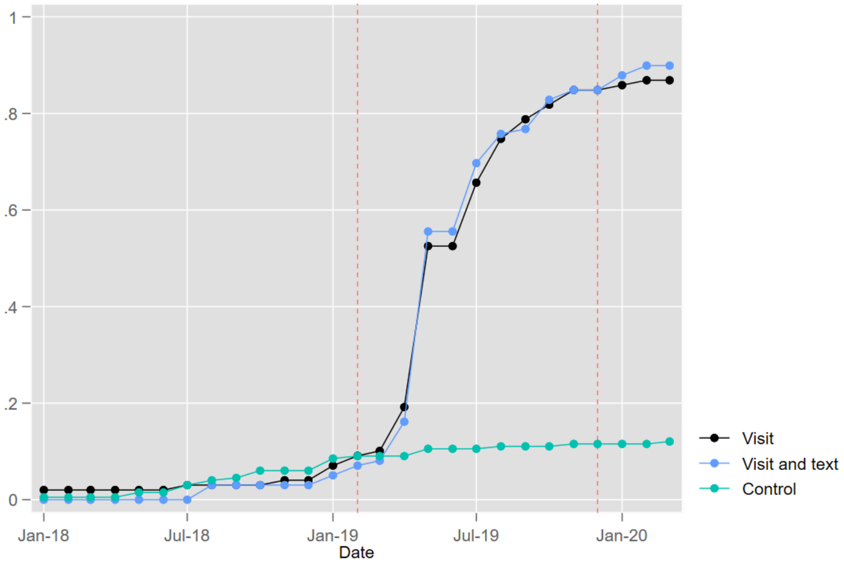
Out of this nationally representative sample of 397 wards, roughly half the Wards in each district were randomly assigned to receive a Whole School Visit at some point between April and November 2019, yielding a total sample of 198. The remaining 199 schools are assigned to only receive WSVs after the completion of the planned endline data collection in November 2020. In addition, we randomly assigned half (99) of the treated schools to the booster program of sending text reminders to WEOs. For the remainder of the paper, the two treatments are referred to respectively as *Visit* and *Visit&Text*.

We shared this sample of schools with the School Quality Assurance Division (SQAD) and they agreed to comply with the treatment assignment. However, compliance to the randomized phase-in design was imperfect. Figure 1 shows the proportion of schools in our sample that received the Whole School Visits over time, broken down by evaluation arm, according to the data shared with us by the SQAD. By the end of December 2019, 85% of schools in our treatment arms received a WSV, compared to 12% in the control. The majority of visits took place over May and July 2019. The two red dotted lined indicate the dated that baseline and midline data collection started (February 2019).

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

[5] 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.7 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. It is thus possible that the worse-performing schools were visited first, because they faced the greatest need for improvement.

Figure 1: Proportion of schools that received a Whole School Visit - by treatment arm



4 Data

We collected baseline data in each of the 397 schools in our sample in February/March 2019, and revisited these schools roughly one year later to perform midline data collection in Feb/March 2020. During these school visits we conducted student assessments, classroom observations, facility inspections, document inspections, and also surveyed 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 for whom we were able to receive parental consent to do the assessments. The grade two students were assessed in Math and Kiswahili, whereas the standard three students were assessed in Math, Kiswahili, and English (English is only taught in grade 3). At endline we aimed to assess the same students, the majority of whom had been promoted to Grades 3 and 4 in 2020. We were able to assess 95 percent of the original baseline sample at midline, yielding a sample of 6,626 students.[6] During assessments we also conducted a brief survey and counted the number of pages completed in the sampled students' exercise books the previous week.

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

[6] 6,623 observations for Kiswahili and 6,596 for Mathematics.

reached a total of ten teachers per school.[7] 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 randomly sampled replacements using the same protocol if we could not find them. We performed **classroom observations** on two randomly selected teachers per school, one for each focal grade, using the World Bank's TEACH instrument (Molina et al., 2018).

The **head teacher survey** included basic demographic information and information about the school. In addition, we asked the head teachers detailed questions about their experience of the Whole School Visit (provided that they had received one), and the extent and nature of interaction with the Ward Education Officers. We also captured information about the beliefs of the quality of the school that they held at the start of 2019 —i.e. before the majority of schools had received a WSV— to see 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 related to the main domains that are the target of the WSV: 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 into the education production function. In each question there was a trade-off between two different inputs —prioritizing early v later grades, teacher training vs infrastructure, learning vs completing the curriculum, and (potentially disruptive) participatory vs traditional teaching methods— and so the head teachers' responses capture their value judgement 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 that have a teacher in them. Finally, we also **surveyed the WEO** in each Ward, and also two School Quality Assurance Officers (SQAOs) in each of the 23 districts.

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.[8] To minimize risk of over-rejection of the null hypothesis due to multiple comparisons, we

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

[8] Two schools could not be reached because of floods, in another two schools the data got lost due to a car accident, and in another two schools, the head teacher was either not available or was unwilling to talk. The data collection team was planning on returning to these schools to conduct the head teacher survey, but this was cut short by school

created Kling 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). We specified all of the hypotheses and indicators that related to each outcome in a pre-analysis, which we registered with the American Economic Association in April 2013, prior to data analysis.[9]

4.1 Balance and attrition

Tables B.1 to B.3 show that the sample is balanced across a range of WEO, teacher, head teacher, school, and student characteristics. The tables also show that the sample remains balanced on the reduced sample of students who we were able to assess at midline, the reduced sample of teachers were able to observe for the classroom observations, and the reduced sample of schools where we were able to interview the head teachers at midline. The bottom row in table B.1 also shows that the proportion of head teachers who reported to have received a Whole School Visit in 2019 or 2020 is 36, 88 and 91 percent in the Control, Visit and Visit&Text arms respectively, leading to a treatment effect on up-take of just over 50 percentage points.

Table B.4 reports attrition analysis for the student-level data. Column one reports results of regressing attrition status on the treatment dummies, including strata fixed effects. It shows that there are no statistically distinguishable difference in attrition rates across the evaluation arms. In the remainder of the columns, we regress a baseline characteristics on treatment assignment, attrition, as well as interaction terms between treatment and attrition. The coefficients on “Attrite” in columns (2) and (3) show that the students who perform worse at baseline were most likely to attrite. This is not surprising: the weakest performing students are most likely to drop out of school or not attend. But more importantly, the coefficients on the interaction terms show that there are no differences in terms of the types of attriters in the evaluation arm. In other words: it is not the case that worse- or better-performing students attrite in the treatment groups. Taken together, the combination of low attrition rates, balance attrition, and no systematic differences in attrition patterns across treatment arm suggests that attrition is unlikely to bias results in this study.

closures.

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

5 Empirical strategy

Our main estimating equation is an Intent-to-Treat (ITT) specification that estimates the effect of being assigned to any of the two treatment arms:

$$y_{i,s,b} = \beta_0 + \beta_1(\text{Visit})_s + \beta_2(\text{Visit\&Text})_s + \gamma_d + X'_{i,s}\Gamma + \epsilon_{i,s,b}, \tag{1}$$

where the dummy variables, $(\text{Visit})_s$ and $(\text{Visit\&Text})_s$, indicate the two treatment arms, γ_d refers to strata fixed effects, $X_{i,s}$ is a vector of controls (measured at baseline) included to improve precision, and $y_{i,s,b}$ is the relevant outcome variable for individual i in school, b.[10] Depending on the outcome, analysis is either at the student, teacher or school level. The error term, $\epsilon_{i,s,b}$, is clustered at the school level, when analysis is at a teacher or student level.

The ITT compares the average outcomes of those assigned to the treatment group with those assigned to the control. However, since we have two-sided non-compliance (12 percent of schools in the control sample received a WSV; and only 85 percent of schools in the treatment group received a WSV by the time of data collection), this does not measure the impact of actually receiving a WSV. In order to calculate the Local Average Treatment Effect— i.e. the treatment effect on the schools that were induced to receive a WSV as a result of treatment assignment (the compliers)— one can simply divide the ITT estimates by the proportion of compliers: $0.85 - 0.12 = 0.73$. In other words, the treatment effect on the compliers is roughly 37 percent larger than the ITT.

For some outcomes the relevant comparisons are only between the two treatments, and so we restrict data to the 84 schools in each of the treatment arms that had received a Whole School Visit by the end of 2019, and estimate the following equation:

$$y_{t,s,b} = \beta_0 + \beta_1(\text{Visit\&Text})_s + \gamma_b + X'_{t,s}\Gamma + \epsilon_{s,b} \tag{2}$$

[10] We stratified by both district and assignment to the teacher incentives program. When possible, we control for the baseline measure of the same outcome. When analysis is at a student level, we control separately for the students' baseline performance in both the mathematics and Kiswahili tests. For the classroom observations data, our control variables include the baseline classroom observation scores. If any observations in the control variables are missing, we impute the control mean and include a dummy variable equal to one if the observation is missing.

6 Results

6.1 Implementation

As a starting point, we investigate the quality of implementation. Table 1 provides some statistics on the nature of the Whole School Visits, as reported by the head teachers. Data is restricted to the two treatment arms, Visit and Visit&Text. The first row in Table 1 shows that 89 percent of head teachers reported that they received a Whole School Visit in 2019 or 2020. Of these Whole School Visits, 12 percent lasted only one day, and thus did not meet the required minimum length of two days. The WEOs were present in 73 percent of these visits, and the average number of days present is just over one. In almost all the Whole School Visits (95 percent) the SQAOs observed teaching in the classroom. In a large proportion of these visits the SQAOs also talked to parents (74 percent) and assessed student (87 percent). This is in contrast to the 'old' model of school inspections that did not involve talking to parents, assessing students, or observing teaching. However, less than half (49 percent) filled in the School Self-Evaluation Form (SSEF) in advance of those visit, although those who did fill it in found it very helpful and easy to understand. The WEO is responsible for distributing the SSEF in advance of the WSV. An exit meeting almost always took place (96 percent of the Whole School Visits), but attendance of parents and the community was low in these exit meetings. The WEO attended in just over half (54 percent) of the exit meetings. Almost all head teachers (90 percent) reported that they learnt something new about student learning and teaching quality, and a slightly smaller fraction reported that they learnt something new about management quality (88 percent) and the quality of community engagement (96 percent),

To summarize, there are some aspects of current implementation of the WSVs that are very encouraging, such as the near universal observation of teaching in the classroom and high proportion of visits including student assessment and talking to parents. It is also highly encouraging that such a high proportion of head teachers believed that they learnt something new. However, the coordination with the WEOs and school community is not perfect. If only half of WEOs attended the exit meeting, they might not be well-informed of the recommendations made by the SQAOs. Moreover, some aspects of the program, such as distributing the SSEFs and displaying the score cards, were not well implemented.

Tables B.5 and B.6 compare the attributes of the WSVs between the treatment arms. Surprisingly, there are strong differences: the WSVs conducted in the control were shorter (only 56 percent were two days or longer), they were less likely to have a WEO present, and the SQAOs were less likely to observe teaching or talk to parents. They were also far less likely to have a SMC member attend the exit meeting, and to receive a school report card. This result has two interpretations. Perhaps the SQAOs were performing better visits in the treatment schools, since they were aware of the evaluation. Or perhaps the head teachers in the control school mistakenly interpreted a regular visit by a SQAO or another government official as a Whole School Visit. Since far fewer schools received the WSV in the control school according to the data collected from the SQAO, relative to the number of head teachers who reported to have received a WSV, the latter interpretation is most likely. Either way, this sheds doubt on using head teacher's report on whether a WSV took place as the measure of treatment uptake when estimating the Local Average Treatment Effect using 2SLS.

6.2 Impacts on beliefs and behavior

Next, we test if the information provided by the Whole School Visits lead to a change in beliefs. Table 2 shows that head teachers in the Visit&Text arm believed that there was *more* room for improvement in the quality of school leadership at the start of the year, relative to the control. There were no differences between treatment arms in head teachers' beliefs of the quality of teaching, community involvement, not belief in the quality of the school environment. The school leadership therefore shifted their beliefs of their *own* behavior, but not in the room for improvement for other school stakeholder. It is possible that head teachers already have accurate (and relatively low) beliefs over the state of community involvement and the school environment. Indeed, the control means reported in the bottom row in Table B.8 show that their beliefs over different indicators of the school environment and community involvement are lower than beliefs over the quality of school management and teaching. The fact that this effect is only statistically significant in the Visit&Text arm suggests that regular reminders from the WEOs about the changes needed for improvement was necessary for these beliefs over quality to really "sink in". Table B.8 further unpacks the leadership domain, and shows that it is beliefs over the need to improve the monitoring of teachers that drives the changes in beliefs over the quality of leadership.

The final four columns in table 2 show that head teachers' beliefs over the education production function did not change: 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 clearly harder to shift. Head teachers in the Visit arm are slightly more likely to prefer investing in infrastructure rather than teacher training, perhaps due to the fact that infrastructure improvements is a common recommendation made in the WSV report.

Next we investigate any changes in behavior of school stakeholder targeted by the intervention: school leadership, teachers, or parents. Table 3 shows that schools are more likely to have an up-to-date Whole School Development Plan (it was a common recommendation to create a new WSDP), but there is no statistically discernible positive impact on management practices, as measured by an overall index of monitoring and curriculum guidance. Monitoring improved by 0.12 SD in the Visit&Text arm, but this is impact is not statistically significant. Table B.10 unpacks the monitoring index and shows that in both treatment arms the teachers are more likely to have their class journal inspected, but they are no more likely to have had homework or student assessment observed, nor to be observed teaching in the classroom.

Table 4 indicates that teacher behavior changed as a result of the program. Column one shows that teacher attendance, measured as the proportion of classrooms in the school that had a teacher in during data collection, is 7.9 percentage points higher in the Visit&Text arm, relative to the control— a 17 percent increase relative to the control mean of 45.7 percent. Teaching practices, as measured by the *Teach* observation tool, also improved by 0.26 standard deviations. There is no evidence of improvements in teacher preparation, assessment, or assigning homework. Moreover, with a p-value of 0.012, we can reject the null that teaching quality did not improve in the Visit&Text arm, even after adjusting the p-value up to account for multiple comparisons.[11] Table 5 further investigates which measures of teaching quality changed in the classroom observation data. There was an improvement in instructional quality, and also the proportion of time that a large number of students were on task during the lesson.

Table 6 show that there is no evidence of improvement in any indicators of parental involvement, such as PTA meetings, SMC meetings, parent contributions, or pupil attendance. There is a 10

[11] The unadjusted p-value for the treatment effect on teaching quality is 0.012, which is the lowest out of the five outcomes in this hypothesis. The new critical value is thus $0.1/5 = 0.2 < 0.012$, regardless of the method of correction (Bonferroni correction, Holm-Bonferroni correction, or the Benjamin Hochberg False Discovery Rate procedure).

percentage point improvement in the Visit arm in the probability of having a school lunch program, which was a common suggestion. Table B.15 further shows that there are no improvements in the quality of the school environment. This is not surprising, given the large resource constraints under which these schools operate. In fact, during our qualitative interviews with WEOs they often note that schools are unable to implement any of the recommendations that require additional resources for infrastructure investments.

6.3 Impacts on student learning

Consistent with the observed positive impact of Visit&Text on teaching practice, table 7 shows suggestive evidence of a modest impact in learning. Student performance in the Kiswahili test improved by 0.05 standard deviations. There was no impact on student performance in the mathematics assessment. The result on the Kiswahili is only marginally significant (p=0.071), however, and the impact on the combined test score (both Kiswahili and Math) is not statistically significant.

6.4 Impacts on WEO knowledge and behavior

Finally, we look at the knowledge and behavior of the Ward Education Officers (WEOs), restricting the sample to the 168 schools in our treatment groups that had received a Whole School Visit by the end of 2019, according to our own monitoring data. Two of these observations are unfortunately missing from the head teacher data due to the issues discussed in Section 4. Moreover, only in 160 of these schools did the head teachers also state that they had received a WSVs, and for some outcomes we only asked the head teacher about the WEOs' role if a WSV had in fact taken place. Statistical power is thus weaker in this smaller sample. Nonetheless, some strong trends merge.

Table 8 reports results on our respective mean indices for WEO knowledge, monitoring, and behavior (Tables B.12 to B.14 show the results to each individual indicator that constitutes the index). There is no evidence that WEOs in the Visit&Text treatment are more knowledgeable about the SQA program: WEOs do not remember more recommendations, nor are they more likely to correctly state that a WSV had taken place in this school. They also did not change the frequency and nature of monitoring, measured by the frequency of visiting schools, and the type of activities performed by the WEO when visiting these schools. However, column (3) shows that there is a sizable and statistically significant increase of 0.27 SD in the in other actions taken by the

WEO in the Visit&Text arm. Table B.14 unpacks the indicators for WEO behavior and shows that head teachers in schools that had a received a WSV are 5.4 p.points more likely to state that the WEO followed up on the recommendations made by WSV report, and 7.9 percentage points more likely to state that the WEO had taken actions that improved learning. They were no more likely to talk to other stakeholders in the community, however.[12] Taken together, it does not seem that WEOs changed their effort levels as a result of the receiving the text messages, but they did change their focus when they visited schools, and were more likely to follow up on the recommendations.

7 Conclusion

This paper reports results of a major government reform by the government of Tanzania. We evaluate the national roll-out of a revised school inspection program, now called School Quality Assurance. Over a period of 12 months, half the schools in the country —roughly 10,000 schools— received a Whole School Visit (WSV) using this new framework. We evaluate the impact of this program using a nationally representative sample of 397 schools, and find that the WSV changed head teacher beliefs, improved teaching practices, and also had a modest positive impact on student literacy. It is impressive that such an ambitious program implemented over such a short time-frame had a positive impact on teacher behavior (1,74 SQAOs and over 4,000 WEOs, spread across the whole country, had to be trained on the new framework), but the impacts on student learning are modest.

8 Tables

[12] Since a common recommendation is involvement of the community to help with infrastructure investments in the school, a possible action from the WEOs would be to visit members of the Village Authority or Village Council in order to lobby for more resources or ask in support in lobbying parents for more resources.

Table 1: Implementation quality

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Whole School Visit					
Conducted	0.893	0.31	0	1	196
At least two days	0.88	0.326	0	1	175
Ward Education Officer					
Present	0.726	0.447	0	1	175
Days present	1.193	0.939	0	3	161
Activities performed					
Observed teaching	0.949	0.222	0	1	175
Talked to teachers	0.903	0.297	0	1	175
Talked to students	0.823	0.383	0	1	175
Talked to parents	0.737	0.441	0	1	175
Assessed students	0.869	0.339	0	1	175
Reviewed documents	0.943	0.233	0	1	175
School Self-Evaluation Form					
Filled SSEF	0.491	0.501	0	1	175
Very helpful	0.942	0.235	0	1	86
Easy to understand	0.744	0.439	0	1	86
Exit meeting					
Took place?	0.96	0.197	0	1	175
Teachers attended	0.931	0.253	0	1	175
SMC member attended	0.571	0.496	0	1	175
Parents attended	0.497	0.501	0	1	175
Community leader attended	0.343	0.476	0	1	175
Students attended	0.269	0.444	0	1	175
WEO attended	0.537	0.5	0	1	175
DEO attended	0.006	0.076	0	1	175
No. parents attended	23.115	24.09	0	110	87
Report card					
Received	0.737	0.441	0	1	175
Publicly displayed	0.535	0.501	0	1	129
Did you learn something you did not know?					
Student learning	0.903	0.297	0	1	175
Teaching quality	0.903	0.297	0	1	175
Curriculum	0.857	0.351	0	1	175
Management quality	0.88	0.326	0	1	175
School environment quality	0.874	0.332	0	1	175
Community engagement	0.857	0.351	0	1	175

Notes: Data is restricted to schools in the Visit and Visit&Text arms.

Table 2: Head teacher beliefs over school quality and education production function

	Room for Improvement				Production Function			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Leadership	Teaching	School environment	Community Involvement	Early vs Late grade	Training vs Renovation	Learning vs Curriculum	Participatory learning
Visit	-0.034	-0.098	0.055	0.163	0.036	-0.105*	-0.069	0.035
	(0.123)	(0.113)	(0.097)	(0.116)	(0.034)	(0.062)	(0.056)	(0.030)
Visit&Text	-0.203*	0.049	-0.099	-0.064	0.019	-0.014	0.083	-0.002
	(0.116)	(0.112)	(0.093)	(0.107)	(0.035)	(0.063)	(0.059)	(0.035)
F-Test	0.210	0.256	0.156	0.078	0.656	0.207	0.025	0.310
Control Mean	0.000	0.000	0.000	1.870	0.785	0.515	0.359	0.908
Observations	387	387	387	387	391	390	391	391
R-Squared	0.105	0.222	0.117	0.086	0.085	0.044	0.138	0.117

Notes: Each column represents a separate regression, estimated using equation 1. Data is at a head teacher level. Columns (1) to (4) show responses to the 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". Columns (1) to (3) are Kling indices, standardized to have a control mean of zero and control standard deviation of one. See table B.8 respective indicators. The dependent variables in columns (5) to (8) are binary variables for vignettes that illicit head teacher's preference for different education inputs. Standard errors are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01; Estimates include strata fixed effects.

Table 3: Management practices of the school leadership.

	(1) WSDP	(2) Monitor	(3) Curriculum
Visit	-0.010	0.021	0.052
	(0.049)	(0.078)	(0.059)
Visit&Text	0.104*	0.112	0.032
	(0.053)	(0.076)	(0.060)
F-Test	0.061	0.327	0.779
Control Mean	0.205	-0.000	0.000
Observations	391	2369	2369
R-Squared	0.116	0.117	0.030

Notes: Each column represents a separate regression, estimated using equation 1. Column (1) is a binary variable equal to one if the school has a Whole School Development Plan that has been updated since November 2019. The dependent variables in columns (2) and (3) are Kling indices standardized to have the control mean zero and standard deviation of one. Data is at a teacher level. See table B.10 and B.11 for the respective indicators. Data from teacher surveys is restricted to teacher who are not also head teachers.

Table 4: Teacher behavior

	(1) Attendance	(2) Teaching practice	(3) Preparation	(4) Assessment	(5) Homework
Visit	-0.006	0.125	0.049	-0.055	-0.031
	(0.043)	(0.112)	(0.035)	(0.039)	(0.081)
Visit&Text	0.079**	0.260**	0.047	0.016	0.123
	(0.039)	(0.103)	(0.034)	(0.038)	(0.089)
F-Test	0.073	0.243	0.947	0.100	0.113
Control Mean	0.457	0.000	0.000	-0.000	1.698
Observations	362	521	2626	2626	3973
R-Squared	0.202	0.181	0.043	0.073	0.124

Notes: Each column represents a separate regression, estimated using equation 1. The dependent variable in column (1) is the proportion of classrooms with students in them that also have a teacher present. Columns (2) to (4) are Kling indices standardized to have the control mean zero and standard deviation of one. Data for column (2) are from the classroom observations, data for columns (3) to (4) are from the teacher survey, and data for column (5) are from the student assessment. The dependent variable in column (5) is the number of exercises completed by a student the week before data collection.

Table 5: Classroom observations

	(1)	(2)	(3)	(4)
	Overall	Culture	Instruction	Time on task
Visit	0.125	0.138^{***}	0.021	-1.706
	(0.112)	(0.047)	(0.063)	(3.625)
Visit&Text	0.260^{**}	0.074	0.091^{*}	7.699^{**}
	(0.103)	(0.047)	(0.055)	(3.669)
F-Test	0.243	0.192	0.293	0.021
Control Mean	0.000	3.320	2.564	55.867
Observations	521	521	521	520
R-Squared	0.181	0.277	0.139	0.224

Notes: Each column represents a separate regression, estimated using equation 1, baseline classroom observation scores for each domain (classroom culture, instructional quality, and time on task). Data is restricted to the 343 schools where we conducted classroom observations at midline. The dependent variables in columns (2) and (3) can range from 1 to 5. See (Molina et al., 2018) for how these measures are constructed. The dependent variable in column (4) is the proportion of time that a high number of students are on task.

Table 6: Parental involvement

	(1)	(2)	(3)	(4)	(5)
	School lunch	PTA met in 2020	Parent contributions	SMC meetings	Student Attendance
Visit	0.097^{**}	0.027	-0.007	-0.287^{*}	0.007
	(0.047)	(0.037)	(0.022)	(0.174)	(0.024)
Visit&Text	0.058	-0.018	-0.006	-0.208	0.019
	(0.051)	(0.035)	(0.023)	(0.194)	(0.021)
F-Test	0.511	0.269	0.977	0.708	0.638
Control Mean	0.355	0.103	0.085	4.684	0.786
Observations	393	389	393	387	393
R-Squared	0.414	0.150	0.070	0.121	0.214

Notes: Each column represents a separate regression, estimated using equation 1. Moving from column (1) to (5), the dependent variables are: (1) a binary variable equal to one if the school has a lunch program; (2) a binary variable equal to one if the PTA met in 2020; (3) the proportion of activities (construction, examinations, and instruction) where parents have helped; (4) the number of SMC meetings held in the past 6 months; and (6) pupil attendance rate

Table 7: Student learning

	(1) Combined	(2) Math	(3) Kiswahili
Visit	0.009	-0.005	0.019
	(0.025)	(0.029)	(0.027)
Visit&Text	0.037	0.017	0.050*
	(0.027)	(0.031)	(0.028)
F-Test	0.354	0.517	0.320
Control Mean	-0.001	-0.004	0.000
Observations	6626	6623	6596
R-Squared	0.619	0.553	0.548

Notes: Each column represents a separate regression, estimated using equation 1, controlling for student baseline performance in Math and Kiswahili. Aggregate scores in Math and Kiswahili are constructed using Item Response Theory, and standardized to have control mean of zero and SD of one.

Table 8: WEO-level outcomes

	(1) Knowledge	(2) Monitoring	(3) Actions
Visit&Text	0.048	0.004	0.272**
	(0.139)	(0.165)	(0.135)
Control Mean	-0.000	0.904	-0.029
Observations	168	166	166
R-Squared	0.202	0.119	0.257

Notes: Each column represents a separate regression, estimated using equation 1. Data is restricted to the schools in the Visit and Visit&Text arms where a Whole School Visit had taken place by December 2019. Each outcome is a Kling index standardized to a control mean of zero and SD of one. See tables B.12 to B.14 for the indicators underlying each index.

References

Avis, Eric, Claudio Ferraz, and Frederico Finan, "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians," *Journal of Political Economy*, 2018, *126* (5), 1912–1964.

Banerjee, Abhijit, 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), 487–511.

Barr, Abigail, Lawrence Bategeka, Madina Guloba, Ibrahim Kasirye, Frederick Mugisha, Pieter Serneels, and Andrew Zeitlin, "Management and motivation in Ugandan primary schools: an impact evaluation report," Technical Report 2012.

Bjorkman, Martina and Jakob Svensson, "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Ugnada," *Quarterly Journal of Economics*, 05 2009, *124*, 735–769.

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, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane, "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa," *Journal of Economic Perspectives*, November 2017, *31* (4), 185–204.

Callen, Michael, Saad Gulzar, Ali Hasanain, Muhammad Yasir Khan, and Arman Rezaee, "Data and policy decisions: Experimental evidence from Pakistan," *Journal of Development Economics*, 2020, *146.*

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.

Conn, Katherine, “Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations,” *Review of Educational Research*, 2017, *87* (5), 863–898.

Dhaliwal, Iqbal and Rema Hanna, “The devil is in the details: The successes and limitations of bureaucratic reform in India,” *Journal of Development Economics*, 2017, *124*, 1 – 21.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, “What Does Reputation Buy? Differentiation in a Market for Third-Party Auditors,” *American Economic Review*, 2013, *103* (3).

_ , Pascaline Dupas, and Michael Kremer, “School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools,” *Journal of public Economics*, 2015, *123*, 92–110.

Dunsch, Felipe A, David K Evans, Ezinne Eze-Ajoku, and Mario Macis, “Management, supervision, and health care: a field experiment,” Technical Report, National Bureau of Economic Research 2017.

Evans, David and Ana Popova, “What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews,” *World Bank Research Observer*, 2016, *31* (2), 242–270.

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

Lavy, Victor and Adi Boiko, “Management quality in public education: Superintendent value-added, student outcomes and mechanisms,” Technical Report, National Bureau of Economic Research 2017.

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?," 2018.

_ **and Venkatesh Sundararaman**, "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India," *The Economic Journal*, 2010, *120* (546), F187–F203.

_ **, Jishnu Das, Alaka Holla, and Aakash Mohpal**, *The fiscal cost of weak governance: Evidence from teacher absence in India*, The World Bank, 2016.

Nyqvist, Martina Björkman, Damien de Walque, and Jakob Svensson, "Experimental Evidence on the Long-Run Impact of Community-Based Monitoring," *American Economic Journal: Applied Economics*, January 2017, *9* (1), 33–69.

OECD, *PISA 2018 Results (Volume I)* 2019.

Olken, Benjamin, "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, 2007, *115* (2), 200–249.

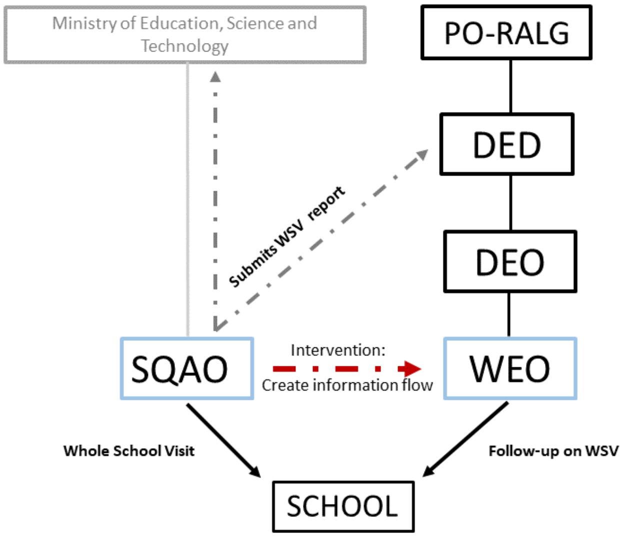
Piper, Ben, Joseph Destefano, Esther Kinyanjui, and Salome Ong'ele, "Scaling up successfully: Lessons from Kenya's Tusome national literacy program," *Journal of Education Change*, 2018, *19*, 293–321.

Pritchett, Lant, *The Rebirth of Education: Schooling Ain't Learning*, The Center for Global Development, 2013.

Raffler, Pia, Dan Posner, and Doug Parkeson, "The Weakness of Bottom-Up Accountability: Experimental Evidence from the Ugandan Health Sector," 2018.

Stigler, George J., "The Theory of Economic Regulation," *The Bell Journal of Economics and Management Science*, 1971, *2* (1), 3–21.

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



Appendix A Additional figures

Appendix B Additional tables

Figure A.2: Theory of Change—Whole School Visits

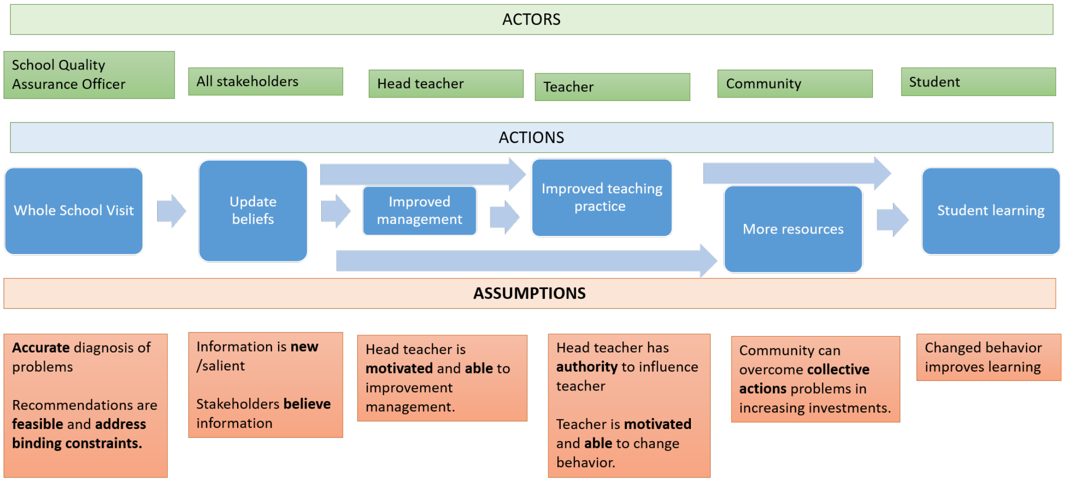

ACTORS
School Quality Assurance Officer
All stakeholders
Head teacher
Teacher
Community
Student
ACTIONS
Whole School Visit
Update beliefs
Improved management
Improved teaching practice
More resources
Student learning
ASSUMPTIONS
Accurate diagnosis of problems
Recommendations are feasible and address binding constraints.
Information is new /salient
Stakeholders believe information
Head teacher is motivated and able to improvement management.
Head teacher has authority to influence teacher
Teacher is motivated and able to change behavior.
Community can overcome collective actions problems in increasing investments.
Changed behavior improves learning

Figure A.3: Theory of Change—Results Framework

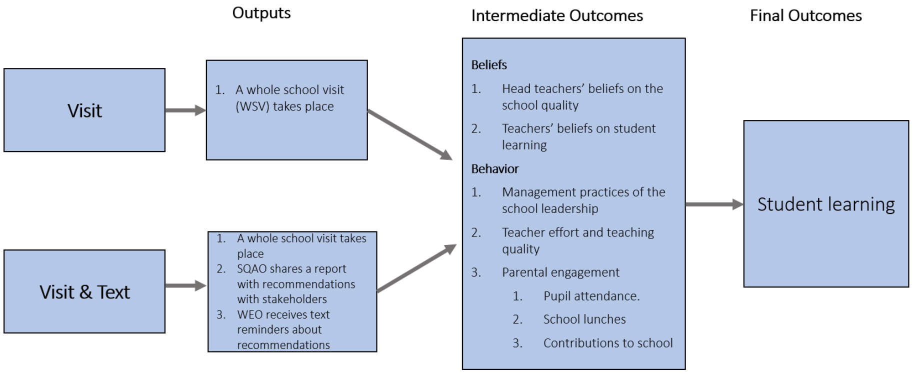

Outputs
Intermediate Outcomes
Final Outcomes
Visit
1. A whole school visit (WSV) takes place
Visit & Text
1. A whole school visit takes place
2. SQAO shares a report with recommendations with stakeholders
3. WEO receives text reminders about recommendations
Beliefs
1. Head teachers' beliefs on the school quality
2. Teachers' beliefs on student learning
Behavior
1. Management practices of the school leadership
2. Teacher effort and teaching quality
3. Parental engagement
1. Pupil attendance.
2. School lunches
3. Contributions to school
Student learning

Figure A.4: Theory of Change—follow-up by WEOs

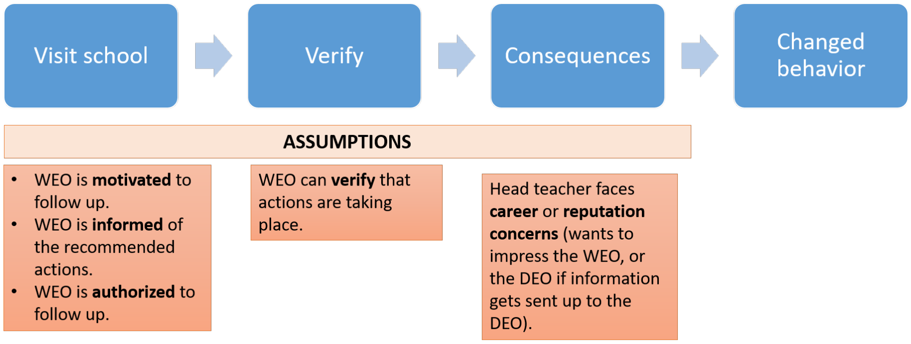
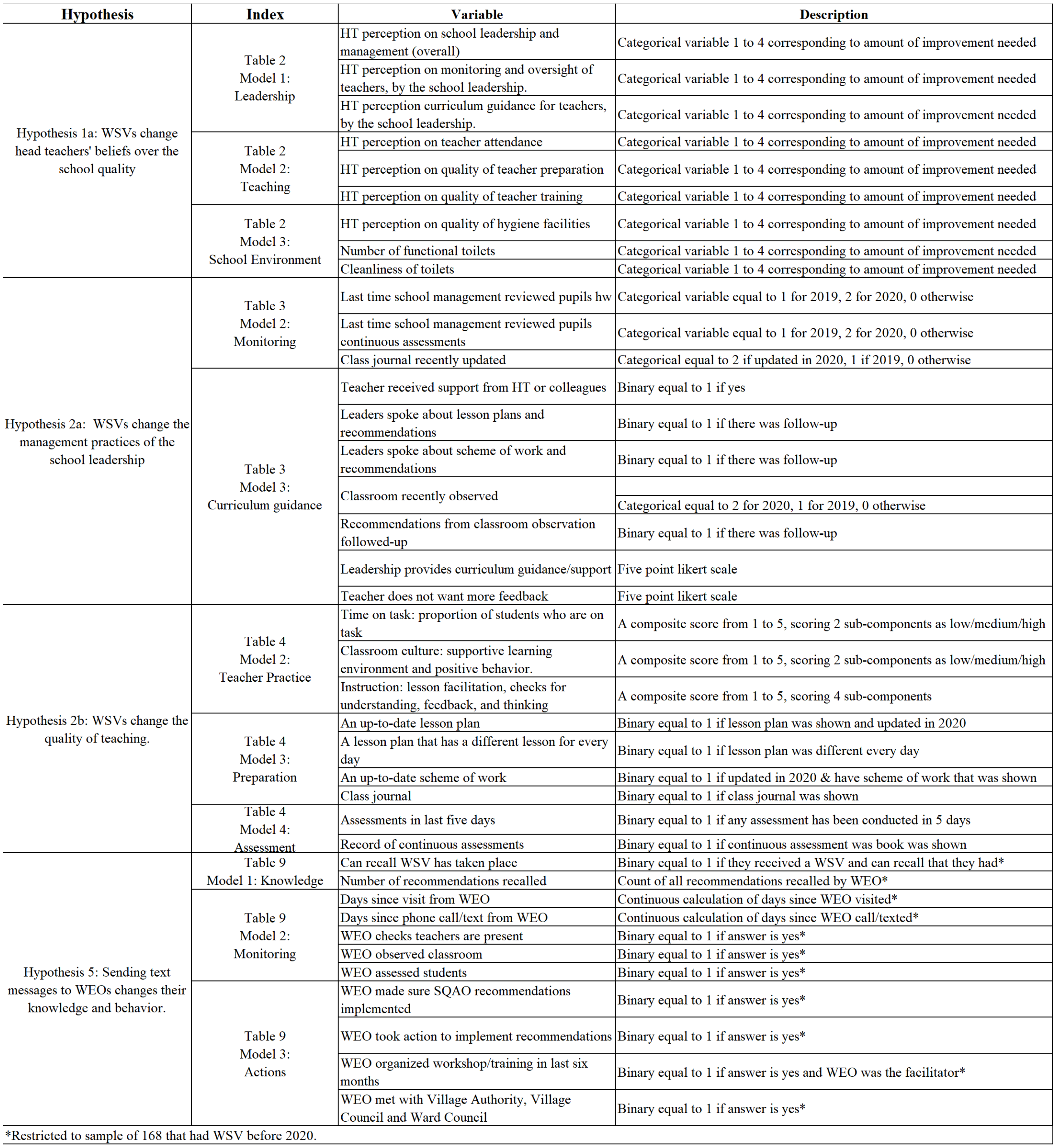

Visit school
Verify
Consequences
Changed behavior
ASSUMPTIONS
• WEO is motivated to follow up.
• WEO is informed of the recommended actions.
• WEO is authorized to follow up.
WEO can verify that actions are taking place.
Head teacher faces career or reputation concerns (wants to impress the WEO, or the DEO if information gets sent up to the DEO).

Figure A.5: Definitions of indices and variables



Hypothesis	Index	Variable	Description
Hypothesis 1a: WSVs change head teachers' beliefs over the school quality	Table 2 Model 1: Leadership	HT perception on school leadership and management (overall)	Categorical variable 1 to 4 corresponding to amount of improvement needed
		HT perception on monitoring and oversight of teachers, by the school leadership.	Categorical variable 1 to 4 corresponding to amount of improvement needed
		HT perception curriculum guidance for teachers, by the school leadership.	Categorical variable 1 to 4 corresponding to amount of improvement needed
	Table 2 Model 2: Teaching	HT perception on teacher attendance	Categorical variable 1 to 4 corresponding to amount of improvement needed
		HT perception on quality of teacher preparation	Categorical variable 1 to 4 corresponding to amount of improvement needed
		HT perception on quality of teacher training	Categorical variable 1 to 4 corresponding to amount of improvement needed
	Table 2 Model 3: School Environment	HT perception on quality of hygiene facilities	Categorical variable 1 to 4 corresponding to amount of improvement needed
		Number of functional toilets	Categorical variable 1 to 4 corresponding to amount of improvement needed
		Cleanliness of toilets	Categorical variable 1 to 4 corresponding to amount of improvement needed
Hypothesis 2a: WSVs change the management practices of the school leadership	Table 3 Model 2: Monitoring	Last time school management reviewed pupils hw	Categorical variable equal to 1 for 2019, 2 for 2020, 0 otherwise
		Last time school management reviewed pupils continuous assessments	Categorical variable equal to 1 for 2019, 2 for 2020, 0 otherwise
		Class journal recently updated	Categorical equal to 2 if updated in 2020, 1 if 2019, 0 otherwise
	Table 3 Model 3: Curriculum guidance	Teacher received support from HT or colleagues	Binary equal to 1 if yes
		Leaders spoke about lesson plans and recommendations	Binary equal to 1 if there was follow-up
		Leaders spoke about scheme of work and recommendations	Binary equal to 1 if there was follow-up
		Classroom recently observed	Categorical equal to 2 for 2020, 1 for 2019, 0 otherwise
		Recommendations from classroom observation followed-up	Binary equal to 1 if there was follow-up
		Leadership provides curriculum guidance/support	Five point likert scale
		Teacher does not want more feedback	Five point likert scale
Hypothesis 2b: WSVs change the quality of teaching.	Table 4 Model 2: Teacher Practice	Time on task: proportion of students who are on task	A composite score from 1 to 5, scoring 2 sub-components as low/medium/high
		Classroom culture: supportive learning environment and positive behavior.	A composite score from 1 to 5, scoring 2 sub-components as low/medium/high
		Instruction: lesson facilitation, checks for understanding, feedback, and thinking	A composite score from 1 to 5, scoring 4 sub-components
	Table 4 Model 3: Preparation	An up-to-date lesson plan	Binary equal to 1 if lesson plan was shown and updated in 2020
		A lesson plan that has a different lesson for every day	Binary equal to 1 if lesson plan was different every day
		An up-to-date scheme of work	Binary equal to 1 if updated in 2020 & have scheme of work that was shown
		Class journal	Binary equal to 1 if class journal was shown
	Table 4 Model 4: Assessment	Assessments in last five days	Binary equal to 1 if any assessment has been conducted in 5 days
		Record of continuous assessments	Binary equal to 1 if continuous assessment was book was shown
Hypothesis 5: Sending text messages to WEOs changes their knowledge and behavior.	Table 9 Model 1: Knowledge	Can recall WSV has taken place	Binary equal to 1 if they received a WSV and can recall that they had*
		Number of recommendations recalled	Count of all recommendations recalled by WEO*
	Table 9 Model 2: Monitoring	Days since visit from WEO	Continuous calculation of days since WEO visited*
		Days since phone call/text from WEO	Continuous calculation of days since WEO call/texted*
		WEO checks teachers are present	Binary equal to 1 if answer is yes*
		WEO observed classroom	Binary equal to 1 if answer is yes*
		WEO assessed students	Binary equal to 1 if answer is yes*
	Table 9 Model 3: Actions	WEO made sure SQAO recommendations implemented	Binary equal to 1 if answer is yes*
		WEO took action to implement recommendations	Binary equal to 1 if answer is yes*
		WEO organized workshop/training in last six months	Binary equal to 1 if answer is yes and WEO was the facilitator*
		WEO met with Village Authority, Village Council and Ward Council	Binary equal to 1 if answer is yes*

*Restricted to sample of 168 that had WSV before 2020.

Table B.1: Balance Tests. Head teacher and school characteristics

	(1) Control	(2) Visit	(3) Visit&Text	(1)-(2)	Difference (1)-(3)	(2)-(3)
Rural	0.729 (0.032)	0.768 (0.043)	0.798 (0.041)	-0.039	-0.069	-0.030
Pubic School	0.970 (0.012)	0.980 (0.014)	0.970 (0.017)	-0.010	0.000	0.010
Years at school	6.698 (0.411)	6.131 (0.535)	6.354 (0.462)	0.567	0.345	-0.222
Teaching experience (years)	17.749 (0.618)	17.657 (0.815)	17.636 (0.767)	0.092	0.112	0.020
Years in position	3.477 (0.271)	3.222 (0.391)	3.222 (0.366)	0.255	0.255	0.000
N	199	99	99			
Reduced sample						
Rural	0.723 (0.032)	0.765 (0.043)	0.796 (0.041)	-0.042	-0.073	-0.031
Pubic School	0.969 (0.012)	0.980 (0.014)	0.969 (0.017)	-0.010	-0.000	0.010
Years at school	6.733 (0.418)	6.163 (0.540)	6.306 (0.464)	0.570	0.427	-0.143
Teaching experience (years)	17.621 (0.623)	17.663 (0.823)	17.612 (0.774)	-0.043	0.008	0.051
Years in position	3.467 (0.275)	3.235 (0.395)	3.163 (0.365)	0.232	0.303	0.071
Conducted Whole School Visit	0.359 (0.034)	0.878 (0.033)	0.908 (0.029)	-0.519***	-0.549***	-0.031
N	195	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. District fixed effects are included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table B.2: Balance Tests: WEO and teacher characteristics

	(1) Control	(2) Visit	(3) Visit&Text	Difference (1)-(2)	 (1)-(3)	 (2)-(3)
Classroom obs						
Teacher Quality Index	0.077	0.067	-0.075	0.010	0.152	0.142
	(0.104)	(0.130)	(0.115)			
N	165	87	91			
WEO						
Male	0.841	0.856	0.845	-0.015	-0.004	0.011
	(0.031)	(0.045)	(0.044)			
Age (in 2020)	42.918	43.273	43.192	-0.355	-0.274	0.081
	(0.391)	(0.536)	(0.571)			
University Degree	1.851	2.163	1.935	-0.312	-0.084	0.229
	(0.142)	(0.229)	(0.212)			
N	199	99	99			
Teacher						
Age (in 2020)	37.525	38.026	38.420	-0.501	-0.895*	-0.394
	(0.312)	(0.483)	(0.450)			
Male	0.556	0.567	0.542	-0.011	0.014	0.025
	(0.020)	(0.026)	(0.030)			
N	1499	769	757			
Clusters	198	99	98			
Teacher (incl. replacements)						
Age (in 2020)	37.267	37.852	38.090	-0.584	-0.823	-0.238
	(0.333)	(0.509)	(0.467)			
Male	0.565	0.573	0.542	-0.009	0.022	0.031
	(0.019)	(0.027)	(0.032)			
N	1286	661	677			
Clusters	198	99	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. District fixed effects are included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table B.3: Balance: Student characteristics

	(1)		(2)		(3)			T-test	
	Control		Visit		Visit&Text			Difference	
Variable	N	Mean	N	Mean	N	Mean	(1)-(2)	(1)-(3)	(2)-(3)
Male	3525	0.498	1715	0.506	1751	0.493	-0.008	0.005	0.013
	[197]	(0.009)	[98]	(0.013)	[98]	(0.012)			
Age (in 2020)	3525	8.833	1715	8.995	1751	9.005	-0.162	-0.172	-0.010
	[197]	(0.096)	[98]	(0.060)	[98]	(0.084)			
Math	3522	0.010	1713	0.007	1746	-0.028	0.003	0.039	0.036
	[197]	(0.032)	[98]	(0.050)	[98]	(0.051)			
Kiswahili	3490	0.013	1692	-0.017	1734	-0.010	0.030	0.024	-0.007
	[197]	(0.035)	[98]	(0.047)	[98]	(0.049)			
Excl. students not assessed at midline									
Male	3318	0.499	1605	0.502	1658	0.486	-0.003	0.013	0.015
	[197]	(0.009)	[98]	(0.013)	[98]	(0.012)			
Age (in 2020)	3318	8.854	1605	8.973	1658	8.990	-0.119	-0.135	-0.017
	[197]	(0.094)	[98]	(0.062)	[98]	(0.086)			
Math	3316	0.023	1604	0.029	1653	-0.016	-0.006	0.039	0.045
	[197]	(0.032)	[98]	(0.052)	[98]	(0.053)			
Kiswahili	3292	0.026	1585	0.007	1644	-0.003	0.019	0.029	0.010
	[197]	(0.036)	[98]	(0.046)	[98]	(0.050)			

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

Table B.4: Student-level attrition analysis

	(1) Attrite	(2) Math	(3) Kiswahili	(4) Age
In-Kind	0.008	0.017	-0.010	0.122
	(0.009)	(0.055)	(0.053)	(0.108)
Recognition	-0.002	-0.031	-0.023	0.131
	(0.008)	(0.053)	(0.055)	(0.125)
Attrite		-0.268***	-0.278***	-0.371
		(0.087)	(0.091)	(0.531)
Attrite x In-Kind		-0.094	-0.120	0.738
		(0.127)	(0.144)	(0.550)
Attrite x Recognition		0.024	0.073	0.635
		(0.145)	(0.142)	(0.554)
F-Test	0.312	0.470	0.840	0.922
Control Mean	0.051	0.010	0.013	8.833
Observations	6991	6981	6916	6991
R-Squared	0.017	0.061	0.045	0.012

Table B.5: Balance Whole School Visits

		(1)		(2)		(3)	T-test	
		Control		Visit		Visit&Text	Difference	
Variable	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)-(2)	(1)-(3)
Whole School Visit								
Conducted	195	0.359	98	0.878	98	0.908	-0.519***	-0.549***
		(0.034)		(0.033)		(0.029)		
Days since visit date and 3/16/2020	70	236.914	85	225.329	88	231.875	11.585	5.039
		(15.983)		(9.673)		(11.466)		
At least two days	70	0.557	86	0.872	89	0.888	-0.315***	-0.330***
		(0.060)		(0.036)		(0.034)		
Ward Education Officer								
Present	70	0.529	86	0.744	89	0.708	-0.216**	-0.179**
		(0.060)		(0.047)		(0.048)		
Days present	60	0.683	78	1.231	83	1.157	-0.547***	-0.473**
		(0.110)		(0.108)		(0.102)		
Activities performed								
Observed teaching	70	0.743	86	0.977	89	0.921	-0.234***	-0.178***
		(0.053)		(0.016)		(0.029)		
Talked to teachers	70	0.914	86	0.860	89	0.944	0.054	-0.030
		(0.034)		(0.038)		(0.025)		
Talked to students	70	0.714	86	0.791	89	0.854	-0.076	-0.140**
		(0.054)		(0.044)		(0.038)		
Talked to parents	70	0.429	86	0.686	89	0.787	-0.257***	-0.358***
		(0.060)		(0.050)		(0.044)		
Assessed students	70	0.771	86	0.849	89	0.888	-0.077	-0.116**
		(0.051)		(0.039)		(0.034)		
Reviewed documents	70	0.914	86	0.942	89	0.944	-0.028	-0.030
		(0.034)		(0.025)		(0.025)		
School Self-Evaluation Form								
Filled SSEF	70	0.414	86	0.500	89	0.483	-0.086	-0.069
		(0.059)		(0.054)		(0.053)		
Very helpful	29	0.931	43	0.953	43	0.930	-0.022	0.001
		(0.048)		(0.032)		(0.039)		
Easy to understand	29	0.897	43	0.651	43	0.837	0.245**	0.059
		(0.058)		(0.074)		(0.057)		

Notes: The value displayed for t-tests are the differences in the means across the groups. Strata fixed effects included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table B.6: Balance Whole School Visits (cont.)

		(1)		(2)		(3)	T-test	
		Control		Visit		Visit&Text	Difference	
Variable	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)-(2)	(1)-(3)
Exit meeting								
Took place?	70	0.943	86	0.977	89	0.944	-0.034	-0.001
		(0.028)		(0.016)		(0.025)		
Teachers attended	70	0.929	86	0.953	89	0.910	-0.025	0.018
		(0.031)		(0.023)		(0.030)		
SMC member attended	70	0.343	86	0.547	89	0.596	-0.204***	-0.253***
		(0.057)		(0.054)		(0.052)		
Parents attended	70	0.286	86	0.535	89	0.461	-0.249***	-0.175**
		(0.054)		(0.054)		(0.053)		
Community leader attended	70	0.229	86	0.360	89	0.326	-0.132*	-0.097
		(0.051)		(0.052)		(0.050)		
Students attended	70	0.243	86	0.256	89	0.281	-0.013	-0.038
		(0.052)		(0.047)		(0.048)		
WEO attended	70	0.343	86	0.570	89	0.506	-0.227**	-0.163
		(0.057)		(0.054)		(0.053)		
DEO attended	70	0.057	86	0.000	89	0.011	0.057***	0.046**
		(0.028)		(0.000)		(0.011)		
No. parents attended	19	31.158	46	26.739	41	19.049	4.419	12.109
		(7.780)		(3.846)		(3.314)		
Report card								
Received	70	0.429	86	0.767	89	0.708	-0.339***	-0.279***
		(0.060)		(0.046)		(0.048)		
Publicly displayed	30	0.467	66	0.576	63	0.492	-0.109	-0.025
		(0.093)		(0.061)		(0.063)		
Did you learn something you did not know?								
Student learning	70	0.843	86	0.872	89	0.933	-0.029	-0.090
		(0.044)		(0.036)		(0.027)		
Teaching quality	70	0.829	86	0.895	89	0.910	-0.067	-0.082
		(0.045)		(0.033)		(0.030)		
Curriculum	70	0.757	86	0.849	89	0.865	-0.092	-0.108
		(0.052)		(0.039)		(0.036)		
Management quality	70	0.843	86	0.872	89	0.888	-0.029	-0.045
		(0.044)		(0.036)		(0.034)		
School environment quality	70	0.800	86	0.884	89	0.865	-0.084	-0.065
		(0.048)		(0.035)		(0.036)		
Community engagement	70	0.771	86	0.872	89	0.843	-0.101	-0.071
		(0.051)		(0.036)		(0.039)		

Notes: The value displayed for t-tests are the differences in the means across the groups. Strata fixed effects included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table B.7: 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.8: Head teacher beliefs, unpacked

	Leadership			Teaching			School environment			Community	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Overall	Monitoring	Curriculum guidance	Attendance	Preparation	Quality teaching	Hygeine	Functional toilets	Clean toilets	Overall	Pupil attendance
Visit	0.046	-0.047	-0.075	-0.160	-0.074	-0.021	0.055	0.069	0.018	0.163	-0.045
	(0.116)	(0.115)	(0.110)	(0.135)	(0.106)	(0.095)	(0.108)	(0.114)	(0.094)	(0.116)	(0.091)
Visit&Text	-0.098	-0.238**	-0.136	0.225*	-0.046	-0.016	-0.023	-0.126	-0.105	-0.064	-0.054
	(0.107)	(0.109)	(0.111)	(0.137)	(0.103)	(0.098)	(0.098)	(0.102)	(0.100)	(0.107)	(0.095)
F-Test	0.262	0.127	0.612	0.013	0.812	0.961	0.505	0.120	0.266	0.078	0.929
Control Mean	2.115	2.192	2.188	2.461	2.176	2.005	1.933	1.845	2.041	1.870	1.917
Observations	384	386	385	387	387	386	387	387	387	387	387
R-Squared	0.061	0.114	0.110	0.200	0.213	0.135	0.118	0.110	0.174	0.086	0.158

Notes: Each column represents a separate regression, estimated using equation 1. The outcome variables are head teacher' responses to the question: “Think back to the beginning of this school year (January/February 2019). How much room for improvement was there in the following areas”. Each variable is a categorical ranging from 1 ”A lot of room for improvement” to 4 ”No improvement was necessary”. Standard errors are clustered at the school level. * for p<.1; ** for p<.05; *** for p<.01; Estimates include strata fixed effects.

Table B.9: Teacher beliefs about student ability

	Prop. of students who can...		Proficiency at grade...			
	(1) Add	(2) Read	(3) 1	(4) 2	(5) 3	(6) 4
Visit	0.016	0.084*	-0.019	0.025	-0.006	0.008
	(0.036)	(0.048)	(0.020)	(0.029)	(0.023)	(0.013)
Visit&Test	-0.015	0.036	-0.008	-0.011	0.049**	-0.014
	(0.040)	(0.045)	(0.018)	(0.024)	(0.023)	(0.011)
F-Test	0.474	0.326	0.630	0.240	0.047	0.101
Control Mean	3.657	3.517	0.170	0.407	0.267	0.045
Observations	1524	1427	2626	2626	2626	2626
R-Squared	0.018	0.025	0.013	0.040	0.045	0.031

Notes: Each column represents a separate regression, estimated using equation 1. The dependent variables in columns (1) and (2) are the share of grade 2 students that the head teacher believes can do addition and read at a grade 2 level. Responses are coded as a categorical variable ranging from 1 “0-25 percent” to 4 ”75 to 100 percent”. Columns (3) to (6) are binary variables, each indicating the grade at which the teacher believes an avererage student in this school would be able to read a short story (one paragraph) for comprehension.

Table B.10: Monitoring

	(1) Overall	(2) See homework	(3) See assessment	(4) Class journal observed	(5) Class journal updated
Visit	0.021	-0.027	-0.051	0.066*	0.014
	(0.078)	(0.059)	(0.054)	(0.037)	(0.076)
Visit&Text	0.112	0.047	0.016	0.066*	0.099
	(0.076)	(0.061)	(0.055)	(0.036)	(0.075)
F-Test	0.327	0.307	0.287	1.000	0.337
Control Mean	-0.000	0.945	1.106	0.460	0.868
Observations	2369	2369	2369	2369	2369
R-Squared	0.117	0.096	0.064	0.153	0.126

Notes: Each column represents a separate regression, estimated using equation 1. The dependent variables in columns (2) to (5) are binary variables

Table B.11: Curriculum guidance

	(1) Overall	(2) Special support	(3) Follow-up lesson plan	(4) Follow-up scheme of work	(5) Recently observed	(6) Follow-up observation	(7) High level support	(8) Feedba
Visit	0.052	0.024	0.047	0.039	-0.026	-0.009	0.049	-0.038
	(0.059)	(0.026)	(0.031)	(0.034)	(0.054)	(0.014)	(0.042)	(0.026
Visit&Text	0.032	0.036	0.034	-0.006	-0.014	0.002	0.051	-0.049
	(0.060)	(0.025)	(0.031)	(0.032)	(0.052)	(0.015)	(0.044)	(0.029
F-Test	0.779	0.688	0.725	0.239	0.842	0.502	0.968	0.713
Control Mean	0.000	0.758	0.360	0.358	0.729	0.092	3.822	1.827
Observations	2369	2369	2369	2369	2369	2369	2369	2357
R-Squared	0.030	0.028	0.043	0.064	0.028	0.029	0.035	0.047

Notes: Each column is a separate regression estimated using equation 1. The dependent variable in column (1) is Kling index for curriculum guidance. Columns (2) to (6) are binary variables. The dependent variable in column is teachers' extent of agreement to the statement "The school leadership provides a high level of curriculum guidan feedback and professional support", ranging from 1 "Strongly disagree", to 6 "Strongly agree". The dependent v able in column (8) is teachers' response to the statement "I would like to receive more feedback about my teach from my Head Teacher", ranging from 1 "Strongly agree" to 6 "Strongly disagree"

Table B.12: WEO Knowledge

	(1) Overall	(2) No. rec. remembered	(3) Recall visit
Visit&Text	0.048	0.510	-0.015
	(0.139)	(0.586)	(0.050)
Control Mean	-0.000	4.107	0.893
Observations	168	168	168
R-Squared	0.202	0.272	0.205

Notes: See Table 8.

Table B.13: WEO Monitoring

	(1) Overall	(2) Days since last visit	(3) Days since last call/text	(4) Checked: teacher present	(5) Observed teaching	(6) Assessed student
Visit&Text	0.004	-2.470	2.246	-0.007	-0.006	0.003
	(0.165)	(6.732)	(9.246)	(0.045)	(0.076)	(0.068)
Control Mean	0.904	24.060	12.988	0.904	0.590	0.699
Observations	166	166	166	166	166	166
R-Squared	0.119	0.115	0.119	0.172	0.181	0.193

Notes: See Table 8.

Table B.14: WEO Action

	(1) Overall	(2) Followed up	(3) Action— improve learning	(4) Organized workshop	(5) Meet stakeholders this year
Visit&Text	0.272**	0.054	0.079	0.064	0.002
	(0.135)	(0.033)	(0.058)	(0.067)	(0.039)
Control Mean	-0.029	0.925	0.787	0.268	0.892
Observations	166	160	160	165	166
R-Squared	0.257	0.235	0.168	0.271	0.143

Notes: See Table 8.

Table B.15: School Environment

	(1) Toilet:Student Ratio	(2) Clean Toilets	(3) Goodstate Classrooms
Visit	0.000	-0.016	-0.001
	(0.003)	(0.094)	(0.001)
Visit&Text	0.001	-0.033	-0.002
	(0.003)	(0.097)	(0.001)
F-Test	0.647	0.878	0.510
Control Mean	0.023	2.734	0.017
Observations	393	397	393
R-Squared	0.247	0.215	0.256

Notes: Each column represents a separate regression, estimated using equation 1.