Journal of Public Economics 168 (2018) 1–20

Contents lists available at ScienceDirect

Journal of Public Economics

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

Experimental evidence on scaling up education reforms in Kenya☆

Tessa Bold[a], Mwangi Kimenyi, Germano Mwabu[b], Alice Ng'ang'a[c], Justin Sandefur[d,*]

[a]Institute for International Economic Studies, Stockholm University, Sweden
[b]Department of Economics, University of Nairobi, Kenya
[c]Strathmore University, Nairobi, Kenya
[d]Center for Global Development, Washington D.C., United States of America

ARTICLE INFO

Article history:
Received 12 December 2016
Received in revised form 10 August 2018
Accepted 13 August 2018
Available online xxxx

JEL classification:
C93
I21
M51
O15

Keywords:
Education
Kenya
Contract teachers
Randomized evaluation
External validity
State capacity

ABSTRACT

What constraints arise when translating successful NGO programs to improve public services in developing countries into government policy? We report on a randomized trial embedded within a nationwide reform of teacher hiring in Kenyan government primary schools. New teachers offered a fixed-term contract by an international NGO significantly raised student test scores, while teachers offered identical contracts by the Kenyan government produced zero impact. Observable differences in teacher characteristics explain little of this gap. Instead, data suggests that bureaucratic and political opposition to the contract reform led to implementation delays and a differential interpretation of identical contract terms. Additionally, contract features that produced larger learning gains in both the NGO and government treatment arms were not adopted by the government outside of the experimental sample.

1. Introduction

A large share of children in low income countries learn little in primary school and complete their education lacking even basic reading, writing and arithmetic skills. There is growing descriptive as well as experimental evidence that low teacher effort can, at least partly, account for this policy failure and that by incentivizing teachers, for example by linking payments or tenure to performance, teacher effort could be raised and student learning outcomes could be significantly improved.[1]

Contract teacher programs – where new teachers are hired in an effort to reduce class sizes at wages below civil service levels, often under direct control of local schools, and without civil service tenure protections – are a case in point. A number of contract teacher trials, implemented across the developing world, have documented significant improvement in student learning outcomes at relatively low cost, thus making them a prime example of programs

☆ Professor Mwangi Kimenyi passed away in 2015. An earlier version of this paper circulated under the title "Scaling up What Works: Experimental Evidence on External Validity in Kenyan Education". We are indebted to the staff of the Ministry of Education, the National Examination Council, and World Vision Kenya, and to the leadership of Claudia Lagat, Mukhtar Ogle, and Salome Ong'ele. Paul Collier, Stefan Dercon, Geeta Kingdon, David Johnson, and Andrew Zeitlin helped conceive this project. The paper improved thanks to input from Michael Clemens, Michael Kremer, Karthik Muralidharan, Paul Niehaus, Lant Pritchett, David Roodman, Torsten Persson, Jakob Svensson, numerous seminar participants, the editor, and two anonymous referees. Maryam Akmal, Naureen Karachiwalla, Anne Karing, Joachim Munch, and Diogo Weihermann provided excellent research assistance. We acknowledge the financial support of the UK Department for International Development (DFID) as part of the "Improving Institutions for Pro-Poor Growth" (iiG) research consortium, the International Growth Centre (IGC), and the PEP-AUSAID Policy Impact Evaluation Research Initiative (PIERI). The views expressed here are the authors' alone.

* Corresponding author.
E-mail addresses: tessa.bold@iies.su.se (T. Bold), gmwabu@gmail.com (G. Mwabu), alicemnganga@yahoo.com (A. Ng'ang'a), jsandefur@cgdev.org (J. Sandefur).

[1] For primarily descriptive evidence on low teacher effort, see Chaudhury et al. (2006) and Bold et al. (2016). For experimental evidence on raising teacher effort, see Duflo et al. (2012), Duflo et al. (2015), Glewwe et al. (2010), and Muralidharan and Sundararaman (2011) and (2013). For recent reviews of the experimental evidence on ways to improve the quality of education, see Kremer et al. (2013), Krishnaratne et al. (2013), McEwan (2015), Conn (2014), and Glewwe and Muralidharan (2015).

https://doi.org/10.1016/j.jpubeco.2018.08.007
0047-2727/

to be recommended for countries with limited public resources to spend on education.[2]

Despite this evidence, however, questions remain whether similar positive effects can be realized when such incentive programs are operated at scale and transposed from NGO to government implementation. We hypothesize that both scale and government implementation will change the incentives provided by otherwise identical contract structures. Politics are a major channel for this effect. While employing a small number of contract teachers at wages far below civil service salaries may not provoke serious political opposition, large-scale implementation, absent other complementary policy changes, may face political resistance from vested interests (e.g. teacher unions) who will prevent government from exercising the de jure flexibility of teacher contracts. State capacity may be an equally important constraint. In settings with weak public monitoring systems, it is unclear whether interventions predicated on the operation of dynamic incentives can be efficiently implemented and enforced by government bureaucrats.[3,4]

In this paper we investigate these issues using experimental data from the pilot phase of a nationwide contract teacher program in Kenya that eventually employed 18,000 contract teachers. We estimate the effectiveness of contract teachers under the management of the Kenyan government and compare it to contract teachers managed by an NGO. From a sample of 192 government primary schools spanning all eight Kenyan provinces, 64 were randomly assigned to the control group, 64 to receive a contract teacher as part of the government program, and 64 to receive a contract teacher under the coordination of the local affiliate of an international NGO, World Vision Kenya.

To study the loss of fidelity when moving from an NGO pilot to a large-scale government program, we also experimentally varied other features of teacher contracts in addition to the implementing agency. In particular, we varied who controls hiring and firing decisions, what role local school management committees play, and how much teachers are paid. These contract variations emulate different versions of contract teacher programs in Sub-Saharan Africa and beyond and were designed to measure the trade-offs involved in moving from a program with strong local accountability to one with weaker control but a higher chance of political feasibility.[5]

Consistent with earlier findings, we find positive and significant effects of the program in schools where the contract teacher program was administered by an international NGO. Placing an additional contract teacher in a school where the program is managed by the NGO increased test scores by roughly 0.18 standard deviations. When moving from NGO to government implementation, however, these positive effects are virtually undone: in our preferred specification controlling for baseline characteristics, treatment effects were significantly smaller and indistinguishable from zero in schools receiving contract teachers from the Ministry of Education.

Beyond these average effects, the contract variations allow us to identify an 'optimal' design in each treatment arm and to estimate the cost in foregone learning of departing from it. We find the largest test score gains, 0.4 of a standard deviation, when teachers were paid a high salary and when parental school management committees were trained to oversee the teacher – regardless of the mode of hiring. The effect is both significantly different from zero and all other cells. Looking separately at each implementer, further illuminating differences emerge: The government achieved the highest test scores gains in the cells with central hiring, while the NGO achieved the highest test score gains in the cells where hiring was devolved to the school. Together, this suggests – sensibly and consistent with some other literature (Finan et al., 2015; Mbiti, 2016) – that more local accountability and better pay lead to better performance (with the caveat that there is an advantage apparent in our data for the government to rely on its central and established bureaucracies in the hiring and payment of teachers, despite the loss in local accountability that this may entail). Politically more palatable cells that dispensed with local control over teacher hiring and management altogether incurred significantly worse results, however. An important caveat to all our results looking at contract variations is that the number of schools in each cell becomes quite small for fine-grained comparisons, and statistical power is limited.

What explains the stark difference in treatment success between the government and the NGO as the program went to scale? We find evidence corroborating both the political resistance and state capacity mechanisms described above. Specifically, the prospect of a nationwide contract teacher program with 18,000 new contract teachers provoked organized resistance from the national teachers union, which demanded permanent civil service employment and union wages for all government hired teachers. The pattern of heterogeneous treatment effects as well as direct surveys of teachers suggest that the union's response, and the controversy surrounding the national scale-up that followed, adversely affected the credibility of dynamic incentives for teachers, in turn lowering their performance, even though teachers in the experiment were not formally covered by union collective bargaining. Importantly, these effects are only discernible in the treatment arm where government hired and managed contract teachers. We further show that monitoring and implementation of the program may have been compromised in several ways in the government treatment arm. For example, schools in the government treatment arm received fewer monitoring visits, and teachers experienced longer salary delays, though only the latter were significantly, negatively correlated with improvements in pupil test performance.

Overall, our results confirm the findings of previous contract teacher interventions regarding the ability of contract teachers to significantly improve learning in public primary schools across diverse baseline conditions in a low-income country — but not in the institutional context of government implementation. Our finding of a fairly large, significant treatment effect from the NGO arm of the contract teacher program implies that the null effect on the government side is not due to a failed intervention in the usual sense. Rather, our more tentative findings, about the link from teachers' expectations and union representation to the failure of the government treatment arm, point to specific mechanisms through which political general equilibrium effects can undermine the government scale-up of successful NGO programs. While some contract variations generated significant, positive learning gains under government implementation, those were not the ones adopted in the eventual national scale-up.

[2] See Duflo et al. (2015)), Muralidharan and Sundararaman (2013) and discussion in Kremer et al. (2013).

[3] Here, dynamic incentives refer to the fact that continued employment as a contract teacher is in principle dependent on satisfactory performance. Beyond this, contract teachers may also have more long-term career concerns if progression to a permanent contract is performance-dependent.

[4] There is a large literature pointing to a theoretical and empirical relationship between state capacity and the provision of public goods and services (see among others Acemoglu, 2005; Besley and Persson, 2011). Conversely, when state capacity is weak, private actors, such as NGOs may be more effective at delivering public services. Reinikka and Svensson (2010) find that religious not-for-profit health care providers in Uganda provided higher quality care than government facilities, while Banerjee et al. (2008) show that government supervisors in Indian public clinics sabotaged an NGO program to monitor absenteeism by nurses.

[5] A cell in which teachers were selected centrally by district education officials and paid the higher salary is similar to West African national contract teacher programs in which large swathes of teachers are employed as contract teachers at initially lower wages essentially as a cost-cutting exercise but with little additional accountability. On the other hand, a cell in which communities are trained in the day-to-day management of the contract teachers, are actively involved in choosing which teachers to employ and teachers are paid a low salary is similar to the Kenyan 'harambee' system in which local communities raise funds to employ contract teachers under the oversight of the community.

Our findings are not meant to imply that successful trials with NGO implementation cannot be scaled up by government – as recent large scale deworming campaigns inspired by the work of Miguel and Kremer (2004) aptly demonstrate. However, they do raise important questions about constraints to large-scale public implementation, especially for programs that may be politically sensitive or require complementary support from the public bureaucracy to work successfully. Importantly, our work also shows that randomized-controlled trials can be used to assess and identify constraints to scaling-up; i.e., shed light precisely on the issue of external validity that is often raised as a weakness of the RCT method itself. A natural next step, pursued in Banerjee et al. (2016), would be to also identify complementary policies to deal with these implementation constraints. Indeed, we attempt to contribute to this question by examining how complementary accountability training and variation in salaries and in the reliance on existing government bureaucracies can improve the effectiveness of contract teachers even within the constraints of government implementation.

The rest of the paper is organized as follows. Section 2 describes the public primary schooling system in Kenya. Section 3 outlines the experimental design and randomization procedures based on a multivariate matching algorithm and reports tests for balance using baseline data. Section 4 discusses compliance. Section 5 presents the main treatment effect estimates, comparing the relative effectiveness of NGO and government implementation based intention-to-treat (ITT) effects exploiting both across and within-school variation in treatment intensity. It also presents evidence on complementary experimental treatment variations. Section 6 explores possible mechanisms explaining the government-NGO performance gap. Section 7 concludes.

2. Context

Primary school enrollment is relatively high in Kenya, but learning levels in primary schools are low. According to the most recent national data prior to our study, from the 2006 Kenya Integrated Household Budget Survey, net primary enrollment was 81%, with government primary schools accounting for approximately 90% of this (Bold et al., 2011). Among children in third grade however, only 3 out of 10 can read a story in English or do simple division problems from the second grade syllabus (Mugo et al., 2011).

2.1. School finance and governance

In January 2003, the Kenyan government abolished all school fees in government primary schools. This "Free Primary Education" (FPE) policy established the current system of school finance in which government primary schools are prohibited from collecting revenue and instead receive a central government grant – commonly known as "FPE funds" – of approximately $13.50 per pupil per annum to cover non-salary costs.[6]

The FPE reform created a new governing body for each government primary school, equivalent to a local school board, known as a school management committee (SMC). The SMC is chaired by the head teacher and comprised representatives from the Ministry of Education, parents from each grade, teachers, and in some cases local community or religious organizations. The SMC manages a bank account where the government deposits FPE funds for each school.

[6] Except where otherwise noted, we convert Kenyan shillings to U.S. dollars using the prevailing exchange rate at the time of the baseline survey in July 2009, 74.32 shillings per dollar.

2.2. Civil service teachers and PTA teachers

Formally, all teachers in Kenyan public primary schools are civil servants employed by the Teacher Service Commission (TSC), a centralized bureaucracy under the direction of the Ministry of Education. Salaries are paid directly from Nairobi to individual teachers' bank accounts. At the beginning of 2011 the Ministry of Education reported a shortage of 61,000 civil service teachers (across roughly 20,000 primary schools) relative to its target of a 40:1 pupil-teacher ratio.

Civil-service teacher shortages reflect demand-side, rather than supply-side constraints. At the time of the experiment, the Ministry was operating under a net hiring freeze for civil service teachers. The relatively high salaries of civil service teachers create a long queue of qualified graduates seeking civil service jobs, which are allocated according to an algorithm that primarily rewards time in the queue rather than merit.

To address teacher shortages, many schools also informally contract local teachers known as Parent-Teacher Association (PTA) teachers, which are funded directly by parents. In the sample of schools surveyed for this study in 2009, 83% of teachers were employed by the civil service (TSC) and the remaining 17% by PTAs. Civil-service teachers earned an average of $261 per month, compared to just $56 per month for PTA teachers.

PTA teachers, as well as the contract teachers discussed below, are often drawn from the queue of graduates awaiting civil service jobs.

2.3. Contract teachers

A priori, there are multiple reasons to expect contract teachers to improve education outcomes. First, they provide additional teaching staff with similar educational qualifications at much lower cost. Second, because their contracts are, in theory, renewable conditional on performance, schools may retain only good teachers – a selection effect. Third, contract teachers lacking permanent job tenure should have stronger dynamic incentives to increase teaching effort – an incentive effect.

In 2009 the government of Kenya announced an initiative to provide funds to schools to employ teachers on contract outside of the civil service system. The current study was designed as an evaluation of a pilot phase of this initiative. The variations in teacher contracts described in Section 3.2 were chosen to inform the design of the eventual national scale-up.

However, scale-up of the national program occurred before the pilot was completed due to political pressure from outside the Ministry of Education. The randomized pilot program analyzed here was launched in June 2010, and in October 2010 the Ministry hired 18,000 contract teachers nationwide, nearly equivalent to one per school. These 18,000 teachers were initially hired on two-year, non-renewable contracts, at salary levels of roughly $135 per month, somewhat higher than the highest tier for the pilot phase.

The allocation of these teachers, coming after the launch of the randomized pilot, provides us with an opportunity to assess impact while the progam is going to scale. It also poses an obvious threat to the internal validity of our estimates. We show in Section 4.3, however, that these teachers were allocated without regard to the distribution of contract teachers in the experimental pilot.

2.4. Organizational structure of implementing agencies: Ministry of Education and NGO

The Ministry of Education is responsible for all government primary schools in Kenya, which account for 90.2% of gross primary enrollment. As of 2005 the Ministry's budget for primary education totalled $731 million (Otieno and Colclough, 2009), compared to

Table 1
Experimental design: number of schools in each program variation.

		Low salary		High salary	
		SMC training	No training	SMC training	No training
Gov't implementation	Local control	12	12	4	4
(64 schools total)	Central control	12	12	4	4
NGO implementation	Local control	12	12	4	4
(64 schools total)	Central control	12	12	4	4

In addition to the 128 schools shown in the table, the experiment included 64 pure control schools. Cells show the number of schools assigned to each treatment, prior to attrition. The decision to place more sample in the low-salary cells was motivated by cost considerations.

just $4 million per annum in international aid to Kenya for primary education channeled through NGOs (OECD, 2012).

To implement programs such as the contract teacher initiative studied here, the Ministry relies on local staff in the district education offices. In principle, district staff should make routine visits to all schools. In practice, the Ministry's ability to directly call on these district officials to carry out specific tasks is limited.

World Vision Kenya is the local affiliate of a large international NGO. Despite being one of the larger international NGOs with a presence in the country, World Vision is active in only a small fraction of Kenyan districts – highlighting again the constraints to scaling up with a non-governmental service provider. Within its areas of operation, World Vision employs permanent staff and paid "volunteers", who monitor and implement all World Vision program activities. World Vision is not traditionally active in school level education programs and instead focuses on community and household-level interventions.

3. Program and research design

The experiment was implemented from June 2010 to October 2011 in 14 districts spanning all 8 Kenyan provinces. 24 schools were sampled from each province, yielding 192 schools in total. One contract teacher per school was randomly assigned to 128 out of 192 sampled schools.

All schools in the study are public (i.e., government) primary schools. In a randomly chosen sub-sample of 64 out of the 128 treatment schools, an NGO was assigned responsibility solely for the contract teacher program. In the other 64 treatment schools, the government took responsibility for the contract teacher program. The timing and intervention protocols for the contract teacher program were identical in the NGO and government treatment arms. The baseline and follow-up data collection, including testing of pupils, was conducted by the same team of enumerators over the same dates.

3.1. Program details

Schools were given funds to hire a contract teacher. Half of the teachers in the experiment were assigned to second grade in 2010, and half to third grade in 2010. In 2011, all the contract teachers were placed in third grade. This created variation in treatment both between schools, as well as within schools because of different lengths and timing of exposure of a given cohort of students.

The head teacher was charged with allocating students to either the existing teacher or the contract teacher. Schools were told that if they were not satisfied with the performance of the contract teacher or if the contract teacher left for other reasons, they could hire a replacement. Head teachers were instructed to split the class to which the new contract teacher was assigned, maximizing the reduction in class sizes in the assigned grade rather than re-allocating teachers across grades.

3.2. Treatment variations

The random assignment of schools to NGO versus government implementation, which is at the center of this study, was overlaid by three additional treatment variations designed to identify the optimal design for the nationwide contract teacher program.

Out of the total 128 contract teacher positions created, 96 were offered KES 5000 ($67) per month, while 32 were offered KES 9000 ($121) per month. See Table 1 for a summary of the design. The high salary was equivalent to 50% of the average entry level civil service teacher salary. The low salary was roughly equivalent to the average PTA teacher salary.

We also tested two modalities for recruiting and paying teachers. In the local cell, responsibility for recruiting and paying contract teachers was assigned to the school management committee, in order to strengthen local control over the teacher's performance. The central-hiring cell in the experimental design was more similar to the civil service model. Teachers were paid directly by the Ministry or World Vision headquarters in Nairobi and district education officers and NGO officials, respectively, were responsible for selecting candidates. In all treatment arms, it was left to the school management committee to decide whether a teacher's performance was satisfactory.

Finally, we explored the importance of local accountability on teacher (and in turn, student) performance with a training intervention that placed particular emphasis on sensitizing school management committees about the contract teacher program in their school and encouraging them to take a more active role in monitoring teacher presence and performance.

Teachers in all treatment arms were required to have completed teacher training and hold a P1 certificate.

3.3. Sample

The experimental sample attempts to be representative of schools with high pupil-teacher ratios. Within each of the eight provinces, districts were chosen non-randomly by the implementing partners, based in part on the location of the offices of the partnering NGO.[7] Within each province, schools with a pupil-teacher ratio below the median were excluded from the sampling frame. Using this sampling frame of high pupil-teacher ratio schools, schools were chosen through simple random sampling within the selected districts.[8] In each school, the sampling frame consisted of all the students in first,

[7] The sample draws from 14 districts in total, using multiple districts from the same province where necessary to reach sufficient sample size. These 14 districts were: Nairobi province (North, West, East); Central province (Muranga South); Coast province (Malindi); Eastern province (Moyale and Laisamis); North Eastern (Lagdera, Wajir South, Wajir West); Nyanza province (Kuria East and Kuria West); Rift Valley province (Trans Mara); Western province (Teso).

[8] Consistent with the sampling frame, Table A.1 shows that the schools in our sample are larger, employed more teachers, and had lower test score performance (as measured by grades in the national primary leaving exam, KCPE) than the average Kenyan primary school.

Table 2
Timeline.

Activity	Month	Year
Sample selection and baseline data collection[1]	July	2009
EMIS data collection used for assignment of 18,000 teachers[2]	March	2010
Random assignment of experimental contract teachers	May	2010
RCT contract teachers placed in schools	June	2010
18,000 non-experimental contract teachers placed in schools	October	2010
18,000 contract teachers made "permanent and pensionable"	September	2011
Follow-up data collection in schools	October	2011
RCT intervention ends (i.e., teacher contracts end)	December	2011
Additional phone surveys of teachers	March	2012

second, and third grade who were present in school on the day of the baseline survey and of all students in third and fourth grade who were present on the day of the follow-up survey. The sample consists of a repeated cross-section of students.[9]

3.4. Data and timeline

The effect of the randomized intervention is measured by comparing differences in academic assessments in math and English across assignment groups.[10] The survey instruments were designed with the collaboration of the Kenya National Examination Council (KNEC) to conform to the national curriculum. The baseline survey – including pupil exams and questionnaires regarding pupil characteristics and school facilities – was conducted in July and early August of 2009 by the KNEC and the research team. The baseline survey was administered to 176 of the 192 schools in the experimental sample. 16 schools, due to transport and security constraints, could not be reached in time.[11]

Teachers were placed in treatment schools in June 2010; their contracts ended in October 2011. Follow-up data collection was conducted in the same sample of schools in October 2011 (see timeline in Table 2). Roughly 15,000 students were tested in the baseline and follow-up surveys. At baseline, we sampled 10 children in first grade and 10 children in second grade for one-on-one testing, and 20 children in third grade for written testing at the school.[12] At follow-up, we sampled 30 children from third grade; i.e., the cohort that was in first grade at the time of the baseline, and 20 children from fourth grade; i.e., the cohort that was in second grade at the time of the baseline, for testing at the school.[13] 10 of the sampled children from third grade were tested one on one, the remainder from Standard 3 and 4 sat written exams. If there were fewer children per grade than the specified sample – an issue that mainly occurred in the North Eastern and Eastern province – the entire grade was sampled, and preference was given to one-on-one testing over written exams.

The one-on-one tests were designed to assess the component skills of literacy as outlined and measured by UNESCO as part of their global literacy assessment and monitoring program. In particular, items covered letter identification, word reading, oral reading fluency, and reading comprehension. The written tests contained similar material but did not assess oral reading fluency. On the mathematics side, we assessed standard components of numeracy such as pre-number skills related to sorting and pattern recognition, number recognition and ordering, number operations, geometry and problem solving. The difficulty of the test was adapted to the grade level. The written tests were marked blindly by a separate set of examiners, whereas the one-on-one tests were marked directly by the (independent) enumerators who administered them. The average age of students at baseline was 7.6 years for the first-grade test, 8.8 years for the second-grade test, and 10.4 years for the third-grade test, with no significant difference in age or numbers across the three groups. The average age of students who sat the test at follow-up was 10.5 years in third grade and 11.4 years in fourth grade, again balanced across all three groups. Both at baseline and at follow-up, a school questionnaire asking for staffing and enrollment was administered to the head master and a pupil questionnaire asking for basic demographic and socio-economic information was administered to the students sitting the test. At follow-up we also collected data on the implementation of the contract teacher program including information on monitoring and presence. Throughout the program, data on the hiring of contract teachers, IDs, salary payments and turnover was collected.

In the term following the end of the contract teacher program, a questionnaire was administered to all contract teachers who had been employed through the program asking for demographic and socio-economic information, their experience through the program, their labor market experience since then and their political attitudes and involvement with the national controversy surrounding the employment of 18,000 contract teachers.

3.5. Randomization and balance

To guarantee that the sample is balanced between treatment and control schools, an optimal multivariate matching algorithm was used (see Greevy et al., 2004; Bruhn and McKenzie, 2009). Treatment and control schools were matched along the following dimensions: baseline scores on the first-grade test, pupil-teacher ratio, number of classrooms, number of civil service teachers, number of PTA teachers and average pay of teachers employed by the Parent-Teacher Associations at baseline and results in nationwide end-of-primary leaving exams from the end of 2005. Baseline data were incomplete and not fully processed at the time of randomization, and district average values were used where data was not available. The algorithm created groups of three schools, which were matched along the above dimensions, and then randomly assigned them to the three primary treatment arms: control, additional teacher with government implementation, and additional teacher with NGO implementation. Fig. 2 in the appendix shows the distribution of schools assigned to the control group and government or NGO implementation across the eight provinces.

[9] A condition of cooperating with government was to work both at baseline and at endline with the government's own Kenyan National Examination Council. This and the wide geographic spread of the study made it difficult to follow individual students.

[10] The official language of instruction is mother tongue in grades 1–3 and English in grades 4–8. In practice, however, the majority of lessons even in lower primary are held in English (Piper and Miksic, 2011). For this reason and because English is crucial for further progression through primary school, we assess children's literacy competence on the basis of English.

[11] All but one of the schools that could not be surveyed at baseline were located in the remote Eastern or North Eastern provinces.

[12] We included an oral one-on-one component because this is the preferred method for testing young children who may not be literate enough to sit written tests. Since this is a very time-consuming way to test students, however, we also gave students (from third grade onward) written tests, which allowed us to test a larger sample of children.

[13] If a child that was sampled, was not present, enumerators were told to sample a replacement. Both at baseline and at follow-up, the replacement rate was roughly one in five children across all treatment arms.

Table 3
Balance at baseline.

	Treatment	Control	Diff	Gov	NGO	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
A: Variables used in randomization						
Pupil-teacher ratio	61.13	63.95	−2.82	61.95	60.30	1.65
No. of classrooms	12.31	11.44	0.87	12.49	12.11	0.38
No. of civil service teachers	8.25	7.31	0.95	8.40	8.11	0.30
No. of PTA teachers	1.85	1.40	0.45	2.00	1.70	0.30
Avg. Pay PTA Teachers (Ksh)	3446.31	3237.38	208.93	3438.05	3454.57	−16.53
Score Primary Leaving Exam	233.71	235.93	−2.22	232.75	234.67	−1.92
Pupil test score, English, Std 1	0.11	0.00	0.11	0.11	0.12	−0.01
Pupil test score, Math, Std 1	0.09	0.00	0.09	0.11	0.08	0.02
B: Other variables						
No. students taking exam, Std 1	9.09	9.15	−0.06	8.86	9.32	−0.46
No. students taking exam, Std 2	9.20	8.92	0.28	9.23	9.18	0.05
No. of students taking exam, Std 3	17.36	17.61	−0.25	16.91	17.81	−0.89
Date of test	0.57	0.50	0.07	0.58	0.56	0.02
Age of children taking test in Std 1	7.60	7.70	−0.10	7.66	7.55	0.11
Age of children taking test in Std 2	8.86	8.70	0.16	8.84	8.88	−0.04
Age of children taking test in Std 3	10.35	10.46	−0.11	10.23	10.49	−0.25
% of boys taking test in Std 1	0.53	0.54	−0.006	0.56	0.50	0.07
% of boys taking test in Std 2	0.56	0.51	0.05	0.55	0.57	−0.02
% of boys taking test in Std 3	0.52	0.56	−0.03	0.51	0.54	−0.03
Pupil test score, Std 2	0.14	0.00	0.14	0.23	0.05	0.18
Pupil test score, Std 3	0.03	0.00	0.03	0.07	−0.003	0.08
Pupil test score, Std 1,2 & 3	0.09	0.00	0.09	0.14	0.04	0.11
Pupil test score – Engl., (1,2,& 3)	0.07	0.00	0.07	0.13	0.02	0.12
Pupil test score – Math, (1,2,& 3)	0.11	0.00	0.11	0.12	0.11	0.01
School avg. test score, Std 1,2 & 3	0.06	0.00	0.06	0.13	−0.01	0.15
Share replaced on day of testing	0.20	0.24	−0.04*	0.19	0.20	−0.02
Pupil wearing shoes	0.59	0.56	0.03	0.60	0.56	0.04
Pupil repeated	0.46	0.45	0.008	0.44	0.49	−0.04
Pupil had extra tuition	0.36	0.32	0.03	0.34	0.37	−0.03
Pupil lives with parents	0.92	0.92	−0.001	0.92	0.92	−0.002
Pupil has eaten breakfast	0.25	0.23	0.03	0.23	0.28	−0.06
Pupil disabled	0.17	0.19	−0.02	0.17	0.18	−0.01
Pupil absences	0.74	0.82	−0.08	0.73	0.76	−0.03
Wealth Index	0.01	0.01	0.0005	−0.01	0.05	−0.06
Parental education	0.29	0.36	−0.06	0.29	0.30	−0.02

School level statistics are based on 176 schools with baseline information, pupil level statistics are based on 6276 pupils from these schools. Standard errors for pupil level information are clustered at the school level. Asterisks denote significance at the 1% (***), 5% (**) and 10% (*) level. The wealth index is created by standardizing each asset indicator using the control mean and control standard deviation and then calculating a row mean across the standardized asset indicator values for each pupil. Parental education is the mean of the dummies for father and mother's education, where 1 indicates secondary, some secondary, or higher education, and zero indicates no schooling, primary or some primary education. †: Date of test is defined as 1 if the school was visited during the second half of the survey and zero otherwise.

Table 3, Panel A, shows balance tests, for the sample of 176 schools with baseline survey data, for the variables used in the block randomization.[14] Panel B reports whether randomization was also successful in achieving balance on baseline indicators that were not explicitly used in the matching algorithm, namely, average standardized test scores (for grades 2, 3 and overall), as well as several other student and test-specific variables.

The number of pupils tested and their age and gender are similar across the three assignment arms. Schools in the three groups are also, on average, tested at approximately the same dates. Although a slightly higher share of sampled students was replaced in control schools on the day of testing, there are no significant differences in the socio-economic composition of students across treatment arms. None of the baseline comparisons with respect to average standardized test scores yield any significant differences. However, we do observe economically meaningful differences in magnitude for baseline test scores, especially for second grade, which are higher in schools where the contract teacher is managed through the government.[15]

Though not significant, the size of the imbalance is such that the estimated treatment effect in the government treatment arm is sensitive to the inclusion of baseline test scores as a conditioning variable. If the imbalance can be treated as 'chance bias' that is due to systematic differences between treatment arms, conditioning would be indicated to obtain consistent estimates of the treatment effect (Bruhn and McKenzie, 2009). If, instead, it arose from data collection and processing flaws, conditioning would result in inconsistently estimated treatment effects. Most damaging, if the imbalance signalled that randomization was tampered with, estimated treatment effects would be inconsistent with or without conditioning.

[14] See Section 3.4. Of the 16 schools with no baseline survey data, 7 of the schools were assigned to the NGO arm, 7 were assigned to the government arm, and 2 were assigned to the comparison group. While a higher share of treatment schools compared to comparison schools could not be surveyed at baseline, there is no statistical difference in the share of schools surveyed between the three intervention arms.

[15] *A priori*, the first- and second-grade tests are anticipated to be more accurate because they were collected through one-on-one tests which have greater reliability for young children with limited literacy, while third grade students took paper-and-pencil tests administered in larger groups.

Table 4
Implementation and compliance.

	Treatment	Control	Diff.	Gov.	NGO	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Teacher recruitment						
Ever employed a teacher	0.87			0.88	0.86	0.02
						(0.06)
No. of months employed a teacher (out of 17)	12.30			11.59	13.00	−1.41
						(1.08)
Panel B: Effects within school						
Class size	60.67	69.45	−8.78	60.47	60.88	−0.41
			(6.14)			(6.42)
Teacher always in correct grade	0.70			0.71	0.69	0.02
						(0.09)
Teacher ever in correct grade	0.95			0.97	0.94	0.02
						(0.04)
Panel C: Reallocation across schools						
Size of treatment cohort	155.83	166.95	−11.12	146.29	166.07	−19.78
			(15.91)			(18.79)
% change in cohort enrollment	−0.11	−0.09	−0.01	−0.14	−0.08	−0.06
			(0.04)			(0.05)
% change in grade enrollment	−0.02	−0.02	0.0004	−0.04	0.008	−0.05
			(0.04)			(0.06)
No. of teachers from 18,000 program	0.65	0.48	0.17	0.68	0.62	0.06
			(0.17)			(0.21)
No. of TSC teachers	9.96	10.10	−0.14	10.15	9.75	0.41
			(1.11)			(1.32)
No. of PTA teachers	2.06	1.74	0.32	2.03	2.09	−0.06
			(0.35)			(0.43)
Panel D: Student composition						
Share replaced on day of testing	0.20	0.18	0.02	0.20	0.19	0.01
			(0.02)			(0.03)
Change in pupils who are male	−0.03	−0.007	−0.03	−0.05	−0.02	−0.03
			(0.02)			(0.03)
Change in pupils with shoes	−0.009	0.03	−0.04	0.02	−0.04	0.06
			(0.03)			(0.04)
Change in pupils who repeat	−0.02	0.01	−0.03	−0.01	−0.02	0.01
			(0.03)			(0.04)
Change in pupils with extra tuition	−0.02	0.08	−0.10	−0.05	0.02	−0.07
			(0.05)*			(0.06)
Change in pupils living with parents	−0.04	−0.04	−0.0004	−0.05	−0.03	−0.02
			(0.01)			(0.02)
Change in pupils who had breakfast	−0.05	−0.02	−0.03	−0.03	−0.07	0.04
			(0.05)			(0.06)
Change in pupils who are disabled	0.04	0.02	0.01	0.01	0.06	−0.05
			(0.04)			(0.05)
Change in pupils absent in last 5 days	−0.006	−0.02	0.02	−0.05	0.04	−0.09
			(0.09)			(0.11)
Change in wealth	−0.10	−0.004	−0.09	−0.12	−0.07	−0.05
			(0.08)			(0.10)
Change in parental education	−0.01	0.03	−0.04	0.04	−0.06	0.09
			(0.05)			(0.07)

The unit of observation is the school except for row 2 and 3 in Panel B. Column 1 shows averages for all treatment schools, and column 2 for all (pure) control schools. Column 3 measures the gap between columns 1 and 2, with standard errors in parentheses, and asterisks denoting differences that are significance at the 1% (***), 5% (**) and 10% (*) level. Columns 4–6 repeat this exercise comparing treatment schools assigned to government or NGO implementation.

To examine the likely source of the imbalance we follow standard practices in the experimental literature.[16] Regarding the integrity of the randomization procedure, we note that baseline data was collected before the randomization had taken place and was not available to anyone other than the research team at the time of the randomization. Second, the randomization took place in the presence of the research team with colored balls being drawn from bags. Third, none of the p-values in the balance table are significant and the results seem reasonable, especially given the large number of comparisons. We are therefore confident that the random assignment procedure was correctly implemented.

Regarding the quality of data collection and processing, we use standard IRT tests of differential item functioning (DIF), which provide a natural test for differential measurement error across treatment groups. Examining test statistics from a Mantel-Haenszel (MH) test, we find significant evidence of DIF (i.e., a p-value of 0.05 or less) for 2 out of 23 binary (i.e., non-partial credit) items at grade 1 when comparing the NGO and government samples, and similarly for just 1 out of 28 binary items at grade 2, and 5 out of 40 binary items at grade 3. This overall rate of DIF is slightly higher than anticipated

[16] For example, Standard Operating Procedure for Donald Green's lab states: "A p-value of 0.01 or lower should prompt a thorough review of the random assignment procedure and any possible data-handling mistakes. If the review finds no errors, we will report the imbalance test, proceed on the assumption that the imbalance is due to chance, and report estimates with and without covariate adjustment."

Table 5
Labor supply of contract teachers.

	LPM			Logit model		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Teacher months						
NGO implementation	.120	.120		.120	.121	
	(.066)*	(.065)*		(.066)*	(.065)*	
High salary		.118			.117	
		(.064)*			(.063)*	
Local recruitment		.140			.141	
		(.065)**			(.065)**	
Geographic density			−.0002			−.0002
			(.0002)			(.0002)
Lagged KCPE score			.002			.002
			(.001)			(.001)
Pupil-teacher ratio			.003			.003
			(.002)			(.003)
Obs.	2060	2060	2044	2060	2060	2044
Panel B: Ever hired a teacher						
NGO implementation	−.018	−.017		−.018	−.016	
	(.062)	(.061)		(.061)	(.051)	
High salary		.139			.130	
		(.070)**			(.047)***	
Local recruitment		.112			.100	
		(.061)*			(.055)*	
Geographic density			−.0002			−.0002
			(.0002)			(.0001)
Lagged KCPE score			.001			.0001
			(.0009)*			(.0008)*
Pupil-teacher ratio			.003			.004
			(.002)			(.002)
Obs.	125	125	124	125	125	124

The unit of observation is the school, with monthly observations from June 2010 to October 2011. In Panel A, the dependent variable is a binary indicator of whether a teacher was present and teaching in a given school in a given month. In Panel B, the dependent variable is a dummy for set to 1 if the schools filled the vacancy in any of the 17 months of the program and zero otherwise. For the logit model, the table reports marginal effects and their standard errors. In these and all subsequent tables, standard errors are reported in parentheses and asterisks denote significance at the 1% (***), 5% (**) and 10% (*) level. Standard errors in Panel A are clustered at the school level.

by chance (with 8 out of 91 total items reporting a p-value of less than 0.05), but does not point to cause for concern in our view. We also note that – despite being higher on average – all treatment arms have common support (see Fig. 3) so that the imbalance is not driven by outliers.

Given this, we proceed on the assumption that it is opportune to condition on baseline test scores both to correct chance bias and to improve precision (Bruhn and McKenzie, 2009; Roberts and Torgerson, 1999; Egbewale, 2015). A potential drawback of reliance on this specification is that baseline data is not available for 14 schools and missingness at baseline may be related to treatment effects. For this reason and to address any lingering concerns about differential measurement error, we also present regressions without controls as well as a number of econometric specifications that are less sensitive to the inclusion of baseline scores.

3.6. Sample attrition

A total of 192 schools were initially sampled for the experiment and assigned to either the intervention or the comparison groups. However, due to transport and security conditions, 4 schools (1 in the government treatment arm, 1 in the NGO treatment arm and 2 in the control group) could not be visited, thus reducing the effective sample to 188 – an attrition rate of 2%. Comparing sampled and attrited schools at endline across the ten variables used in the randomization, we find two significant differences: schools that attrited have lower pupil-teacher ratios and pay lower salaries to PTA teachers. To safeguard against non-random attrition, we estimate Lee bounds for the main results.

4. Compliance and implementation

Random assignment of a school to the treatment group created a job vacancy for a contract teacher. To simulate a scalable program, the onus then fell on district and school officials – under the guidance of their implementation partner; i.e., either the government or the NGO – to recruit a suitable teacher, place him or her in either second or third grade, and split that grade into two (or more) streams. Examining compliance is both of independent interest and can shed light on the mechanisms underlying the treatment effects on learning we document in the next section.

4.1. Teacher recruitment

The 128 schools assigned to receive a contract teacher as part of the experimental evaluation had varying success in recruiting and retaining contract teachers. Of the 64 schools assigned to the government (NGO) treatment arm, 88% (86%) were successful in hiring a contract teacher at some point during the program. The schools that were not able to hire at all are primarily located in hard to reach areas in Eastern province, but also in Nairobi West, where there was some unwillingness to participate in the program. However, even when the school was successful in hiring, teachers did not necessarily stay with the school for the entire duration of the program, and when a vacancy opened up, it was not always filled. As a consequence, out of the seventeen months of the program, schools in the government (NGO) arm actually employed a teacher for 11.6 (13.0) months on average (see Panel A of Table 4).

Table 5 examines the vacancy rate more closely, modeling success in filling a vacancy as a function of variations in contract and salaries

that were manipulated by the experiment. In the top panel, the dependent variable is a binary indicator of whether a teacher was employed in a given school in a given month, with monthly observations spanning the duration of the experiment from June 2010 to October 2011. In the bottom panel, it is a dummy set to 1 if the school ever hired a teacher and zero otherwise. We estimate both a linear probability model and a logit model.

We examine three experimental determinants of teacher labor supply. First, Table 5 shows that NGO implementation led to 12% more months with a filled vacancy, relative to the government treatment arm, and this effect is significant across all specifications. Second, local control over teacher hiring and payment had an effect of similar magnitude to the salary differential, raising the probability of a filled vacancy by a robustly significant 14% across specifications. Third, offering a "high" salary increases the probability of filling a teaching vacancy by just under 12%, mirroring results from Ferraz and Finan (2009) and Deserranno (2016). This effect is significant and consistent between the LPM and logit models. The first and second findings point to the challenges of government implementation, and a possible, partial solution (decentralized hiring). The third effect suggests that the failure to recruit a teacher was sensibly related to experimentally controlled wage offers, suggesting that limited supply of contract teachers, at least in certain areas, could constrain the nationwide implementation of a low-cost contract teacher program.

In addition, we also examine how teacher hiring was related to endogenous school characteristics such as the pupil-teacher ratio, test score performance and density of surrounding schools. Though imprecisely estimated, the coefficients suggest a negative association between hiring and a higher density of schools in a 5 km radius, a positive association with baseline student performance, and a positive association with the existing pupil-teacher ratio at the start of the program. Overall, the patterns are the same regardless of whether we look at the inframarginal (months of teachers) or extra-marginal (ever hired a teacher) success in hiring with the notable exception that we only find evidence that NGO implementation led to more hiring success when we look at months of employment rather than the binary indicator of ever employing a contract teacher.

4.2. Changes in school and classroom characteristics induced by the program

The contract teacher intervention was intended to operate via two channels: reducing class size by adding more teaching staff; and increasing the quality and motivation of this additional staff through the contract structure. Importantly, our ability to measure both effects using test-score data on the target cohort of pupils also hinges on schools' willingness to comply with the intervention by (a) placing the contract teacher in the correct grade, and not reallocating the existing teacher for that grade, such that the class-size reduction is concentrated on the treatment cohort.[17]

Overall, compliance was good, in the sense that schools placed teachers in second or third grade as instructed (Table 4, Panel B). Ninety-five percent of teachers were employed in the correct grade at least some of the time, 70% were exclusively employed in the treatment grades and only three teachers reported that they were never placed in the grades that were tested at endline and/or intended to be exposed. Non-compliance, such as it was, had two sources: (i) teachers in schools where the contract teacher was placed in third grade in 2010 progressed with their cohort to fourth grade; (ii) about 30% of teachers were asked to teach in higher grades in addition to the experimental grades. Since the scheduled teaching time is shorter for lower grades than for upper grades this does not necessarily reduce exposure for the experimental grades.

Compliance was less impressive on the issue of splitting classes. Instructions were to keep existing teachers in place, and reduce class sizes for the targeted grade. In practice, class sizes in the treatment cohort fell by only about 10%, and this reduction is not significant. This suggests that existing teachers may have been reassigned to another grade. Notably, Duflo et al. (2015) find that class size effects explained little or none of their positive results from contract teachers in an NGO program in Western Kenya.

Importantly, there are no significant differences in compliance between the government and the NGO. Neither teacher placement nor changes in class size were significantly different between the NGO and government sample. This suggests that any differential effects on test scores will not be driven by the inability (or unwillingness) of the implementing agency to follow the intervention protocol.

4.3. Reallocation across schools

A second question is the extent to which teachers and pupils endogenously reallocated in response to the program.

First, random assignment to the treatment group may affect a school's hiring of PTA teachers or the probability of being assigned a TSC teacher and/or one of the 18,000 teachers from the national contract teacher program.[18] If staff levels responded endogenously to the placement of a contract teacher through the research program, then the estimated treatment effect may be biased (most likely downwards). We explore this possibility in the last three rows of Table 4, Panel C. Across the board, there are no significant differences between treatment and control schools (or between NGO and government treatment arm) in terms of number of PTA teachers, number of civil service teachers, and number of teachers from the national contract teacher program. Of course, it is still possible that schools in the government and NGO treatment arm responded differently to the national-scale up and we examine this possibility formally in Section 6.

Second, we are concerned with possible shifts in school enrollment in response to the program. The survey consists of a panel of schools, not a panel of students. Thus, estimated treatment effects may be due to changes in performance for a given pupil, and/or changes in the composition of pupils. In either case, these are causal effects, but with very different interpretations. To shed light on which of these two channels drives our results, Table 4 reports enrollment levels at the end of the program and percentage changes in enrollment between 2009 and 2011 in the treatment cohort. There are no significant differences in enrollment in the treatment cohort between treatment and control schools and between the government and NGO treatment arm. Overall, there is a small reduction in enrollment in all schools (enrollment in the treatment cohort drops by roughly 10% between 2010 and 2011), but this trend is uniform across the various treatment arms. We cannot rule out that these net enrollment changes mask larger gross changes, leading to changes

[17] For comparison, in Muralidharan and Sundararaman (2013) a contract teacher was provided to a school with no restrictions on how they were to be assigned or used. The result is that the estimated treatment effect combines both class size and incentive effects. In contrast, in Duflo et al. (2015) contract teachers were assigned to a given grade and students randomly assigned to contract or existing teacher, thus allowing the authors to separate class size effects from the incentive effect.

[18] *A priori*, we would not expect the hiring of the eighteen thousand contract teachers in the national scale-up to respond to the employment of teachers in the experiment. Firstly, the allocation of contract teachers in the national program was based on administrative enrollment data collected before the beginning of the experiment described here. Secondly, the steering group, which included several high-ranking government officials, specifically agreed that allocation of teachers in the national and in the experimental program would be independent of each other.

in the unobserved ability of pupils. We argue that the observed net enrollment changes would have to mask implausibly large (and systematic) changes in gross enrollment for this to be a concern in the estimation.

Consistent with this, we find little or no difference in the evolution of the socio-economic composition of pupils across treatment arms at follow-up: across ten variables, we find one significant difference between treatment and control (fewer students with extra tuition in the treatment schools), and no significant differences across government and NGO implementation.

To summarize, we find that the contract teacher job vacancies created by the experimental program were filled in roughly 70% of months overall, with a quantitatively small but significant difference between NGO and government. Teachers were overwhelmingly placed in the correct grade, though often replacing rather than complementing the existing teacher, yielding small net changes in class size in our sample. None of these reallocations differed between the NGO and government treatment arm. Finally, there is little evidence of reallocation of teachers or pupils across schools in response to the program.

On the basis of these compliance patterns, we interpret the estimated parameters in the next section as causal treatment effects on a given cohort of pupils, with a more limited role for class size reductions.

5. Results

As noted in the introduction, scaling up successful education programs in many low-income countries typically implies a transition from working with non-governmental organizations to working within governments. The experiment here is designed to address this central question of whether the Kenyan government can implement a fairly standard contract teacher program. We present ITT effects of the contract teacher program as a whole on learning outcomes, then separately for the NGO and the government treatment arms. As a robustness check, we show that these effects are driven by exposure to contract teachers *per se* in Section 5.2, where we also exploit experimental variation in length of exposure to the program within schools. We further explore the robustness of the findings in the Online Appendix.

On average, the NGO treatment arm yields a significant increase in overall learning of around 0.2 standard deviations once controlling for baseline covariates, while the government treatment arm shows no effect. But a more nuanced picture emerges in Section 5.3 where we show that both overall and in each implementing agency there exist variants of the program design that lead to sizeable test score gains. Nevertheless, these treatment variants generally raise the overall effect size without reducing the gap between the government and NGO arms, thus they provide little guidance to explain the differential performance by implementing agency, a topic we turn to in Section 6.

5.1. Comparing the effectiveness of contract teachers under government and NGO management

We begin by estimating the average intention-to-treat (ITT) effect of school-level assignment to the contract teacher program on test scores, then proceed to compare the effects of the NGO and government treatment arms. The dependent variable Y_{ijt} comes from a test in English and Maths administered in 2009 and again in 2011, standardized relative to control schools in each year. The ITT effect is measured by the coefficient on the random assignment variable Z_{jt} in Eq. (1), where $Z_{j,t=0} = 0$ and $Z_{j,t=1} = 1$ if the school was assigned a teacher and zero otherwise.

$$Y_{ijt} = \alpha_1 + \beta_1 Z_{jt} + \gamma_1 \mathbf{X}_{jt} + \varepsilon_{1ijt} \quad (1)$$

The coefficient β_1 measures the causal effect of being assigned to treatment status, averaging over schools with varying degrees of success in recruiting contract teachers. We estimate Eq. (1) for three variants of the test score, the combined English and Maths test score, and the score in each subject separately, and report results both at individual student level and collapsed to the school level. For each test score and sample, we use three alternative sets of controls: first, we use a single cross-section of post-treatment data without controls, second we include initial test scores averaged at the school level $\left(\bar{Y}_{j,t-1} \in \mathbf{X}_{jt}\right)$.[19] Third, we pool both pre- and post-treatment data in a standard differences-in-differences specification including controls for school-level fixed effects and a time dummy.[20]

The first row in each Panel of Table 6 presents the estimates of the average ITT effect for the different specifications. The point estimate is fairly consistent across all specifications, at roughly 0.1 standard deviations for the pooled test score, with slightly larger and significant effects in English and lower effects in Mathematics. Collapsing results at the school, the treatment effect is estimated at 0.2 school level standard deviations, but equally imprecise.[21]

The remainder of Table 6 examines how the treatment effect of a contract teacher differs by implementing agency. In each case, we regress scores on the random assignment variable Z_{jt}, interacted with indicators for assignment to the NGO or government treatment arm:

$$Y_{ijt} = \alpha_2 + \beta_2^{ngo}(Z_{jt} \times \text{NGO}_{jt}) + \beta_2^{gov}(Z_{jt} \times \text{Gov}_{jt}) + \gamma_2 \mathbf{X}_{jt} + \varepsilon_{2ijt} \quad (2)$$

The β^{ngo} and β^{gov} coefficients are ITT measures in that they capture the causal effect of being assigned to the NGO or government treatment arms. As above, we present results for pooled scores and separately for each subject, in the student and the school sample, and for three different sets of controls $\mathbf{X}_{jt}$.

The results reveal large and significant differences in the performance of the contract teacher program between the two treatment arms when controlling for baseline achievement. The ITT effect of contract teachers in the NGO treatment arm is estimated to be between 0.15 and 0.18 of a standard deviation in the student sample, and between 0.3 and 0.34 in terms of school level standard deviations, an effect that is both economically meaningful and statistically significant in the regressions with baseline school average test scores and fixed effects. The effects are somewhat larger and significant in English and somewhat smaller and imprecise in Mathematics. The ITT effect of contract teachers in the government treatment arm is between a quarter to a third smaller in the simple cross-section (estimated at 0.12 standard deviations in the student sample and 0.21 standard deviations in the school sample). This, however, turns into an essentially zero effect in our preferred specifications which

[19] Initial test scores are averaged at the school level since the students sampled at baseline are in general not the same as the students sampled at follow-up. $\bar{Y}_{j,t-1}$ is the average score over first, second, and third grades.

[20] We do not include dummies to control for stratification in the analysis here, but present such a specification as a robustness check in the appendix. The main reason given by Bruhn and McKenzie for always including stratum dummies is as a safeguard against the small number of cases where failing to do so results in overly optimistic standard errors. This is not the case in our data, however: standard errors in the specification that does not control for stratum dummies are more conservative than in the specification that includes such dummies. Second, our preferred specification controls for baseline performance either through school averaged baseline scores or through school fixed effects. In the latter case, stratification block dummies are absorbed by the school fixed effects. In the former case, including stratification dummies would reduce the sample further since it effectively drops not only schools with missing baseline scores but also schools that are matched to them in the block randomization. Given the sample size in this study, we prefer to retain these observations rather than control for the method of stratification.

[21] Earlier versions of the paper explored heterogeneous treatment effects on baseline characteristics, showing that learning impacts are negatively associated with schools' baseline learning outcomes. Results are available on request. For the sake of brevity, the following sections only disaggregate effects along experimental dimensions.

Table 6
Treatment effects.

	Both subjects			English			Maths		
Sample of students	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pooling treatment arms:									
Z	.138	.103	.078	.148	.122	.101	.080	.047	−.0001
	(.090)	(.075)	(.080)	(.087)*	(.071)*	(.074)	(.073)	(.070)	(.088)
NGO vs. Govt.									
Z× NGO	.153	.184	.175	.167	.201	.195	.083	.091	.056
	(.109)	(.088)**	(.091)*	(.104)	(.080)**	(.083)**	(.091)	(.089)	(.105)
Z× Gov	.124	.021	−.022	.129	.042	.004	.078	.002	−.057
	(.106)	(.090)	(.095)	(.104)	(.084)	(.086)	(.087)	(.084)	(.107)
Coeff. 1 - Coeff. 2	.029	.163	.197	.039	.159	.190	.005	.088	.113
	(.119)	(.095)*	(.094)**	(.112)	(.082)*	(.081)**	(.101)	(.102)	(.118)
No. of students	8812	8220	14,432	8812	8220	14,432	8812	8220	14,432
Sample of schools									
Pooling treatment arms									
Z	.254	.182	.175	.269	.219	.216	.192	.093	.037
	(.177)	(.141)	(.145)	(.169)	(.130)*	(.138)	(.192)	(.182)	(.186)
NGO vs. Govt. :									
Z× NGO	.295	.328	.336	.311	.353	.365	.228	.212	.168
	(.205)	(.163)**	(.164)**	(.195)	(.150)**	(.152)**	(.222)	(.211)	(.226)
Z× Gov	.213	.037	.017	.228	.087	.070	.156	−.025	−.092
	(.204)	(.162)	(.169)	(.194)	(.149)	(.158)	(.221)	(.210)	(.223)
Coeff. 1 - Coeff. 2	.083	.291	.319	.083	.267	.295	.072	.237	.261
	(.205)	(.166)*	(.165)*	(.195)	(.153)*	(.145)**	(.222)	(.215)	(.251)
No. of schools	188	174	348	188	174	348	188	174	348
Baseline test scores (school average)		X			X			X	
School fixed effects			X			X			X

The dependent variable is a standardized score on a math and English test administered to pupils in grades 3 and 4 in 2011 and grades 1, 2 and 3 in 2009, either pooled or separately in each subject. Columns 1,2,4,5,7 and 8 use only the 2011 (follow-up) test data as the dependent variable. Columns 1,4 and 7 use no additional controls and and columns 2,5 and 8 control for baseline school average scores. Columns 3,6 and 9 use both follow-up and baseline test scores as the dependent variable and control for school fixed effects. At baseline, *Z* takes a value of zero for all schools. In the follow-up survey *Z* takes a value of 1 for schools randomly assigned to any treatment arm. *NGO* is an indicator variable for the NGO treatment arm and *Gov* is an indicator variable for the government treatment arm. Standard errors are clustered at the school level in the student sample regressions.

control for baseline scores or school fixed effects – both for the combined test score and separately for each subject.

Fig. 3 unpacks the result by displaying the kernel density of pupil-level test scores for each of the three treatment arms: government, NGO, and control, both at baseline in 2009 and the follow-up in 2011. The distributions are quite close in 2009, and move apart in 2011. Because test scores are standardized relative to control schools in each year, the overall shape of the distributions differs between years, but in each round all treatment arms are handled identically.

Fig. 4 shows the main ITT result graphically, comparing the kernel density of test score changes between control schools, the government treatment arm, and the NGO treatment arm. The ITT effect does not appear to be driven by outliers, as the NGO test-score distribution lies everywhere to the right of the government test-score distribution.

In Section A.2, we explore the robustness of the core results along the following dimensions: (a) the appropriateness of including baseline controls; (b) relatedly, the inclusion or exclusion of schools with missing baseline values; (c) Lee bounds and other controls to examine potential bias due to non-random attrition at follow-up; (d) a battery of robustness checks related to outliers, calculation of standard errors, stratification, and standardization of the test score variable.

We find no evidence that the large difference in treatment success between NGO and government implementation is driven by mean reversion of government test scores, missing baseline test scores or attrition of schools or students at follow-up. We also find the results robust with respect to trimming of outliers, stratification, randomization inference and method of standardization. We therefore conclude that the contract teacher program led, on average, to sizeable learning gains under NGO implementation, but had a negligible impact under government management.

How do the estimated effect sizes compare to earlier findings by Duflo et al. (2015) in Western Kenya and Muralidharan and Sundararaman (2013) in Andhra Pradesh, India? We find a 0.18 standard deviation ITT in the NGO treatment arm on combined math and English scores in a specification controlling for baseline school test performance.[22] Duflo et al. find a 0.3 standard deviation effect on math and literacy scores in a specification controlling for baseline pupil test scores, while Muralidharan and Sundararaman find an ITT effect of approximately 0.15 standard deviations, also on combined math and literacy also controlling for baseline pupil test scores. However, these results measure somewhat different things. Ideally one would standardize the effects to allow for differences in (a) the length of exposure, and (b) the proportion of test-takers directly exposed to treatment. A rough attempt to do so below suggests that effect sizes in the NGO treatment arm here are slightly larger than those found by Duflo et al. and slightly smaller than those found by Muralidharan and Sundararaman.

In the Duflo et al. study, researchers took a more active role in the management of the school during the experiment, controlling the allocation of pupils to classrooms to isolate the pupils directly exposed to contract teachers. The 0.3 standard deviation effect applies only to pupils in the contract teacher's classroom after 19 months exposure to the program. Adjusting this crudely to a 12-month duration implies a per pupil, per annum effect of 0.19 standard deviations.

[22] We focus on our NGO treatment arm to provide a sense of how reasonable the effect size is for the most successful component of our experimental evaluation. Arguably, implementation conditions in Muralidharan and Sundararaman are more comparable to our government treatment arm, in which case the relevant coefficient from our study is approximately zero.

In contrast, Muralidharan and Sundararaman adhere to a more "business-as-usual" model of school management: schools could allocate pupils to the contract teacher as they saw fit, and thus the ITT effect is measured for all pupils in the school, not just those (non-randomly) assigned to the contract teacher's classroom.[23] This implies an effect size per exposed pupil of about 0.6 standard deviations, or 0.3 standard deviations per annum.

In the current study, the one additional teacher provided by the program was spread across two grades over a duration of 17 months, and the ITT effect is computed over both the relevant grade cohorts. Thus the *prima facie* effect size of 0.18 standard deviations is equivalent to a per pupil, per annum effect of approximately 0.25 standard deviations under strong linearity assumptions. Furthermore, one could argue that teacher absenteeism was relatively high (27%) and roughly 30% of contract teacher positions were unfilled at any given point, so these effects should be seen as even larger relative to the actual days of teaching.

In sum, our estimates of learning gains for the NGO treatment arm are on the high end for comparable experimental pilots of contract teachers in the developing world: larger than evidence from a more controlled experimental setting elsewhere in Kenya, and roughly the same as effects found under more business-as-usual conditions in India. But we would caution that the strong linearity assumptions used to compare across studies here may exaggerate the differences in the effect sizes between Duflo et al. on the low end and the present study on the high end. For instance, if there are diminishing marginal returns to length of exposure to the program, the large per annum effect of the somewhat shorter program studied here would be reduced. Similarly, if contract teachers have positive spillovers on pupils not assigned to them as their primary instructor – particularly when assignment of pupils to classrooms is not randomized, but optimized by school management as in Muralidharan and Sundararaman and the current study – then our back-of-the-envelope calculations would overstate the degree to which these latter two studies find larger effects than Duflo et al. The final caveat would be that learning metrics differ across all these studies, as do the distributions of baseline learning levels, so the comparability of effect sizes across any two studies must be taken with some skepticism – while the comparison between treatment arms within our study is likely more reliable.

5.2. Intensity of treatment

We find some evidence that treatment effects on student learning are linked to experimental variation in the length and intensity of exposure to a contract teacher.

As noted in Section 3, contract teachers were randomly assigned to either second or third grade in the first year of the intervention, and all teachers were placed in third grade in the second year. As a result, our sample contains pupils who – under perfect compliance – would have experienced zero, seven, ten, or seventeen months of a contract teacher in their grade.[24] While the main specification in Eq. (1) defines all students in treatment schools as equally exposed, we now re-define the treatment as a continuous variable ranging from zero to seventeen.[25]

The pattern of effect sizes using this continuous treatment variable is similar to the binary school-level treatment specification. This indicates that the treatment effects we observe are indeed linked to the intended length of exposure of contract teacher deployment. The results are presented in Table 7 where the dependent variable is the pooled test score in English and Mathematics now standardized relative to the control group within each grade (since grades are no longer equally balanced across treatment and control). The upper panel shows the pooled results, the bottom panel distinguishes by implementing agency. For comparison, we repeat the original binary treatment specification in column (1) using the re-standardized test scores.

While it is reassuring that our results appear to be driven by students who were intended to have more exposure to a contract teacher, the fact that the effect sizes are similar – rather than larger – when comparing classes also within treatment schools could be a consequence of spillovers (through teachers being deployed in grades beyond the designated ones as seen in Section 4 and/or class size changes manifesting in other than the intended grade). We explore this by examining how test scores are related to (endogenous) variation in teacher deployment at both the school and grade level. In particular, we present OLS regressions of test scores on actual months the school hired a teacher, both in levels and interacted with a dummy for whether the school strictly followed the intervention protocol.[26]

Relative to the ITT estimates, we find bigger coefficients overall and in the NGO treatment arm, while the government treatment arm continues to show zero impact. Overall, we take these results, both exploiting exogenous and endogenous treatment variation, as evidence that it is exposure to contract teachers, rather than 'Hawthorne' effects, which are driving the treatment effects.

5.3. Contract variations and training

So far we've focused on the generic treatment of receiving a contract teacher, managed either by the government or the NGO. Within each treatment arm, we also randomly assigned three variations in the program design, related to training, devolution, and pay. First, half of the treatment schools received an overlapping training intervention for school management committees. Duflo et al. (2015) show that training school management committees (SMCs) in their governance responsibilities are an effective complement to the contract teacher intervention, which we sought to emulate here. Second, in half of treatment schools, the SMC was given direct, local control over teacher recruitment and payment. Third, in a quarter of the treatment schools, contract teachers received a considerably higher salary. Each of these variations is of independent interest, but may also shed light on the central discrepancy between NGO and government implementation.

The most flexible specification would be to estimate a fully saturated model, with separate treatment effects for each of these twenty-four cells. For each of the three program variations and the variation in government versus NGO implementation, there are two treatment conditions, yielding sixteen cells plus the pure control schools. Define four $1 - by - 2$ vectors for each school j, one for each treatment condition, e.g., $\mathbf{NGO} = \{I[NGO_{jt} = 0]I[NGO_{jt} = 1]\}$,

[23] The median number of teachers per school was 3, thus a given pupil would – assuming an equal division – have a 25% chance of being instructed by the contract teacher directly, plus any effect due to class size reduction.

[24] The four categories of duration of exposure to treatment in treatment schools emerge as follows: students in third grade at endline in schools where teachers were assigned to second grade in year 2010 were exposed to a full seventeen months of a contract teacher across 2010 and 2011, students in fourth grade at endline in schools where teachers were assigned to second grade in year 2010 received zero exposure; students who were in third grade at endline in schools where the contract teacher was placed in third grade in 2010 were exposed for ten months in 2011; and students in fourth grade at endline in schools where the contract teacher was placed in third grade in 2010 were exposed for seven months in 2010.

[25] Since there is overwhelming evidence that the impact of inputs applied at earlier dates fades out with time, we also estimate a specification (not shown here) where we take account of the timing of exposure. We find no significant difference between seven months of early exposure and ten months of late exposure.

[26] For the latter, we focus on the sample that was intended to be treated, omitting students in treatment schools where intended treatment was zero months.

Table 7
Intensity of treatment.

	Experimental variation		Endogenous variation	
	(1)	(2)	(3)	(4)
Z	.093			
	(.073)			
Length of intended exposure		.088		
		(.077)		
Mos. of Contract Teacher			.123	
			(.077)	
Months employed a teacher × Strict compliance				.130
				(.100)
Obs.	8220	8220	8220	7083
Z× NGO	.169			
	(.084)**			
Z× Gov	.015			
	(.087)			
Intended exposure × NGO		.168		
		(.098)*		
Intended exposure × Gov		.005		
		(.098)		
Months of teacher × NGO			.207	
			(.089)**	
Months of teacher × Gov			.022	
			(.103)	
Months of teacher × Strict compliance × NGO				.298
				(.136)**
Months of teacher × Strict compliance × Gov				−.023
				(.122)
Obs.	8220	8220	8220	7083

The dependent variable is a standardized score on a math and English test administered to pupils in grades 3 and 4 in 2011 and grades 1, 2 and 3 in 2009 (here standardized within each test and relative to the control group). The upper panel presents the results for the pooled scores and the bottom panel distinguishes by implementing agency. Column 1 repeats the main specification (column 2 in Table 6) with the re-standardized test score. Column 2 re-defines the school level treatment dummy to reflect the length of experimental exposure: It is set to 1 (17 out of 17 possible months) for students tested in third grade at endline in schools where the teacher was placed in second grade in 2010. In schools where the teacher was placed in third grade in 2010 the treatment variable is set to 10/17 for students tested in third grade at endline and to 7/17 for students tested in fourth grade at endline. It is set to zero for students tested in fourth grade at endline in schools where the teacher was placed in second grade in 2010 and third grade in 2011. Column 3 regresses on actual exposure in each treatment school (as share of 17 months a teacher was employed). Column 4 regresses on the interaction of months the school employed a teacher times strict compliance in terms of placement in the intended classroom comparing the sample that was intended to be treated to control schools. All regressions control for baseline test scores and standard errors are clustered at the school level.

and similarly for **SMC**, **Local**, and **High**. Our regression specification is then

$$Y_{ijt} = \alpha_3 + \boldsymbol{\beta}_3 \cdot (\mathrm{NGO}_{jt} \otimes \mathrm{SMC}_{jt} \otimes \mathrm{High}_{jt} \otimes \mathrm{Local}_{jt}) + \gamma_3 \mathbf{X}_{jt} + \varepsilon_{3ijt} \quad (3)$$

where $\otimes$ denotes the Kronecker product of vectors yielding all sixteen possible indicator variables, $\cdot$ denotes a dot product or inner product, and $\boldsymbol{\beta}_3$ is a 1-by-16 vector of coefficients. For reasons of both statistical power and ease of interpretation, we also present results that aggregate the coefficients in Eq. (3) into separate indicators for each of the three program variants and an overall treatment indicator, as well as several intermediate combinations that combine any two program variants.[27,28] Note again that none of the program variants applied to the control schools without a contract teacher. We present results with and without controlling for average baseline scores at the school level in Table 8.

For completeness, we present all sixteen coefficients from the fully saturated model in Table 8. In our preferred specification controlling for a school's baseline test scores (columns 4 and 6), four of the sixteen coefficients are statistically significant. To make more sense of these results, Fig. 1 groups the coefficients in linear combinations to test the effect of each contractual variation in isolation.[29] Note that the coefficients in Fig. 1 do not represent new estimates, but rather post-estimation combinations of the results from the fully saturated model in Table 8.[30]

We find some evidence that school management training had a positive effect on learning, particularly when combined with other program elements. Pooling the coefficients from all arms where the SMC was trained to supervise the contract teacher program, we find a significant positive effect on test scores of 0.2 standard deviations

[27] See Muralidharan et al. (2018) for a discussion of specification choice with cross-cutting designs. Reliance on a more parsimonious specification requires assumptions about equality of parameters across sub-groups that must be motivated by underlying hypotheses, or else run the risk of data-based model selection and incorrect inference.

[28] Random assignment to contract variation is balanced in the full sample and within each implementer for the accountability training and salary variation where differences across cells are both small and insignificant. For the hiring cross-cut, we find that schools in the NGO treatment arm had significantly higher test scores at baseline when hiring was done locally, while the opposite was true in the government treatment arm. Looking at the cells in the fully saturated model, we see sizeable imbalances as a consequence of the small number of schools in each cell. For this reason, we focus on results that control for baseline scores and those that aggregate over several cells. Results available on request.

[29] Note that the study is somewhat under-powered to study these cross-cutting interventions separately for government and NGO treatment arms. Given the small sample, it is worth noting the (ex post) power calculations for these tests. With an intraclass correlation of 0.33 in our endline test data and a correlation of 0.43 between baseline and endline scores, the MDE for a single treatment arm (i.e., government or NGO implementation) is approximately 0.26 standard deviations.

[30] For instance, the top left panel of Fig. 1 averages all coefficients from groups that received SMC training, and compares that to the average coefficient across all other groups that received a contract teacher, using Stata's lincom command.

Table 8
The effect of experimental contract variations on student learning.

	Pooled		NGO		Gov	
	(1)	(2)	(3)	(4)	(5)	(6)
No SMC - Central - Low	.149	.205	.123	.280	.175	.130
	(.112)	(.107)*	(.126)	(.128)**	(.159)	(.152)
SMC - Central - Low	.165	.160	.180	.155	.151	.165
	(.162)	(.133)	(.218)	(.181)	(.220)	(.177)
No SMC - Local- Low	.123	.103	.292	.276	−.045	−.070
	(.149)	(.130)	(.181)	(.183)	(.216)	(.165)
No SMC - Central- High	.062	−.138	−.239	−.027	.363	−.248
	(.216)	(.152)	(.140)*	(.085)	(.398)	(.283)
SMC - Local - Low	.022	−.043	−.020	.010	.064	−.096
	(.148)	(.118)	(.205)	(.142)	(.191)	(.171)
SMC - Central - High	.194	.345	−.128	.163	.516	.527
	(.221)	(.091)***	(.144)	(.078)**	(.407)	(.153)***
No SMC - Local - High	.003	−.044	.065	.032	−.060	−.119
	(.216)	(.123)	(.362)	(.193)	(.218)	(.127)
SMC - Local - High	.745	.459	1.276	.805	.214	.113
	(.323)**	(.192)**	(.611)**	(.364)**	(.185)	(.091)
Linear combinations						
SMC - Central - High	.013	.245			.393	.545
	(.222)	(.081)***			(.410)	(.149)***
SMC - Local - High	.642	.375	1.237	.678		
	(.322)**	(.186)**	(.612)**	(.364)*		
Baseline test scores (school average)		X		X		X
Obs.	8812	8220	5915	5580	5893	5549

See notes for Table 6. The dependent variable is the combined score on the English and Mathematics test, standardized relative to control schools in each year. Columns (3)–(6) report the coefficients from the fully saturated regression in Eq. (3), estimated with and without baseline controls. Column (1) reports the average of the coefficients in column (3) and (5) to give the treatment effects in the pooled sample. Column (2) reports the average of the coefficients in column (4) and (6).

in the sample as a whole, driven disproportionately by the government arm.[31] Disaggregating again to allow for contract variations to have complementary effects, we see that the combination of SMC training and a high salary appears to have a stronger positive effect, in the pooled sample and both the NGO and government-run arms. Combining results in an alternative sequence, we see that SMC training was particularly effective in the government arm when combined with central hiring.

In contrast, neither local control over hiring nor higher salaries had any effect on average when pooling the coefficients from various treatment arms (see top-middle and top-right panels of Fig. 1). These null results mask some heterogeneity between the NGO and government arms, however. There is some sign that local hiring had perverse effects in the government arm, and higher salaries had a small positive effect in the NGO-run program.[32]

Given the limited number of observations in each cell, we must be extremely cautious in interpreting these results. However, we see some evidence in Fig. 1 that coupling SMC training with high salary led to significantly better results (0.36 of a standard deviation in all treatment arms) than the average across all other cells. Results in the cross-section are qualitatively similar, but less precisely estimated. Compared to the aggregation that considers each contract variation in isolation, these results suggest that the anticipated positive effect of a higher salary for contract teachers is only 'switched on' when combined with local accountability training, and vice versa: treatment effects in the cell that combines SMC training with a low salary are much smaller.

Overall, the results suggest that there are important complementarities between each of the contract variations and between contract variations and implementing agency. These complementarities are seen most clearly by inspecting the coefficients in the fully saturated model in Table 8: The contract cell that one would expect *a priori* to function best, namely salary incentives coupled with strong local accountability, indeed leads to the highest test score gains, 0.81 of a standard deviation higher than in control schools, but only in the NGO treatment arm. In the government treatment arm, on the other hand, being able to rely on established bureaucracies in hiring is vital for test score gains to materialize. As a result, it is the cell that combines SMC training and high salary with central hiring that sees the largest test score gains relative to control schools, 0.53 of a standard deviation. When looking at pairwise comparisons of the cells, we cannot always reject the null that the 'optimal' design for each implementer performs identically to each of the other contract cells, and in particular to the one that is the polar opposite. However, we can conclude with confidence that the optimal cell leads to higher test scores than averaging across all other designs. Results are similar in the cross-section, but less precise.

Importantly, while the overall comparison points to an unsuccessful intervention in the government treatment arm, the analysis of the contract variations admits a more nuanced picture. In both treatment arms, there exist combinations that lead to large test score gains (though we caution against drawing overly strong conclusions given the small sample size and the highly non-linear nature of the results). Hence, concluding that the government cannot implement a contract teacher intervention would be too simplistic; it can if it gets the details right.

Successfully conducting a contract teacher program is however not quite the same as being able to scale it up. Although we do not

[31] In previous versions of the paper, we estimated the impact of each contract variation in isolation using a more parsimonious model with just one dummy variable for each contract variation. In such a specification, we found no result of any of the contract variations. The difference in results is due to the fact that the high salary cells are weighted differently when calculating average effects in the two specifications, namely according to number of observations in the parsimonious specification and equally in the fully saturated model, and that the fully saturated model allows for non-linearities while the parsimonious regression does not.

[32] While the overall null result on teacher pay stands in contrast to some of the results in the literature (Ferraz and Finan, 2009), it is consistent with recent research by de Ree et al. (2018), who found that a doubling of pay for Indonesian school teachers had no effect on test scores. Though it should be noted that in our context, the wage variation entails both a selection effect and an incentive effect, while de Ree et al. (2018) study a wage increase for ex-ante identical teachers.

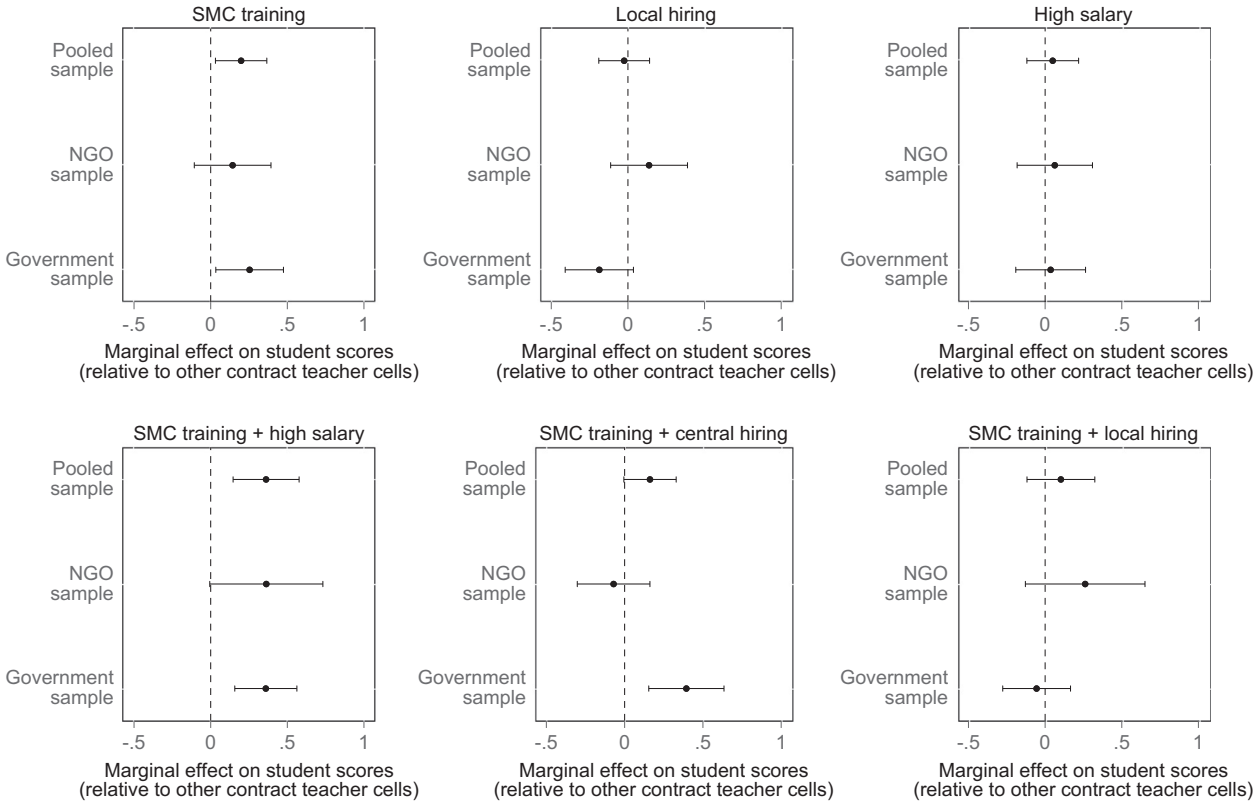

Fig. 1. The effect of experimental contract variations on student learning ctd. (Linear combinations of the coefficients in column (2), (4) and (6) of Table 8).

have data on the performance of teachers in the national scale-up, the contract variations do allow us to speculate a little. In particular, local accountability training, which was both an expensive and time-consuming part of the intervention seems to be crucial for benefits to materialize. This was the element that – alongside dynamic incentives – was dropped in the national scale-up, making it most similar to the 'no SMC training – Central – High salary' cell in our experiment, which incurred the worst results when controlling for baseline scores (negative and insignificant) and performed substantially worse than the optimal outcome in the cross-section in the government treatment arm. That is, based on our experimental design, there are contract variations with which government can successfully implement contract teacher programs, but they were not the variations endogenously chosen.

6. Mechanisms

We now turn to examining mechanisms which could explain the overall performance of the program and in particular the difference in performance between contract teachers in the NGO and government treatment arms. We explore three sets of explanations. We argue that the first two are likely a function of working with government, at least in Kenya and similar settings, independent of the scale of the program: differences in the selection and motivation of teachers hired, and weak monitoring and accountability within government systems. We also explore a third mechanism that may have undermined the government's performance that is potentially related to scaling up *per se*: the effect of the political response to the contract teacher program by the national teachers' union.

Methodologically, we proceed in three steps. First, we present treatment effects of random assignment to the government or NGO treatment arm on intermediate outcomes, such as the observable human capital of contract teachers recruited through the program, the number of monitoring visits made to treatment schools, and indicators of union activity and identification. Second, we report simple correlations between the final outcome variable (improvements in test score performance over the duration of the program) and these intermediate outcomes associated with various causal mechanisms. Third, we add interaction terms to the main treatment effects specification from equation (1) to examine the plausibility that the national controversy surrounding the hiring of 18,000 contract teachers disproportionately affected teachers in the government treatment arm, and thus helps to explain the differential effect on test scores. We use both observational (to measure absenteeism) and survey data collected as part of the follow-up survey, as well as data based on exit interviews with contract teachers conducted after the follow-up survey.[33]

6.1. Teacher selection

While the protocol for teacher recruitment was the same for the government and the NGO treatment arms, these institutions may have differed in the kind or quality of teachers they attracted. In practice, we see that the Ministry hired teachers with higher educational attainment, although there is no significant difference in terms of teaching qualifications (Table 9, Panel A). Teachers in the government arm are also more likely to be female. There is no significant difference in terms of age between government and NGO. Interestingly, none of these observable skills or demographic characteristics

[33] For the exit interviews, we were able to track 111 contract teachers drawn from 84 of the 108 schools that employed a teacher. There are more teachers than schools, because teachers who did not stay for the entirety of the program were replaced by new hires. Absenteeism data is available for 76 of the 108 treated schools. Attrition was not systematically related to treatment arm (government vs. NGO), treatment effects, initial pupil-teacher ratio and baseline scores.

are significantly correlated with changes in test scores (column 4, Table 9).[34]

The government and NGO arms may also differ in the extent of 'local capture' of the hiring process by existing public servant teachers, as Duflo et al. (2015) showed in an NGO program. The percentage of contract teachers who were friends of existing teachers or SMC members was two thirds in the government treatment arm, almost twice as high as in the NGO treatment arm (Panel A, Table 9). While this finding might suggest a corrupted hiring process in the government arm, it is also possible that teachers in the government arm were hired more locally and are therefore better connected. In any case, the indicator of local capture does not show the negative correlation with test score improvements that one might expect.

In sum, while we find minor differences in observable characteristics between teachers in the NGO and government treatment arms, observational data analysis provides little reason to suspect that these differences drove differences in treatment effects on learning. A possibility remains that the NGO (being a well-regarded and well-known organization in Kenya) may have been able to attract more motivated teachers. Certainly, from other contexts there is evidence that not-for-profit providers as well as more pro-social occupations may be more attractive to intrinsically motivated employees.[35] While we cannot rule this out, we deem it less likely to be the case here. First, the NGO partner, has never been active in school management or the employment of teachers, and second, regardless of the implementing agency, the employment contract was directly between the school and the teacher.

6.2. Monitoring and accountability

We find modest differences in teacher absenteeism and external supervision between the government and NGO treatment arms and some sign that these differences explain student outcomes; in particular, the significantly higher rate of salary delays in the government arm is strongly (negatively) correlated with learning gains.

Ex ante, there is strong reason to suspect that the Ministry's routine monitoring system of teachers operated by the Quality Assurance and Standards Directorate is quite weak and this could contribute to the different outcomes in the NGO and the government treatment arm. Our baseline survey shows roughly 25% absenteeism among civil service teachers, while the Kenyan Anti-Corruption Commission estimates that there are 32,000 ghost teachers on the government's payroll, representing 14% of all teachers (Siringi, 2007).

We compare government and NGO along three dimensions related to implementation and management of the program: teacher effort as measured by presence in the classroom during an unannounced visit, monitoring of schools, and successful management of the payroll (Table 9, Panel B).

Teacher presence in the classroom is indeed higher in schools managed by the NGO (73% versus 63%), and they were 11% more likely to have received a monitoring visit than schools in the government treatment arm. Only the latter difference is statistically significant, however, and while the correlations with test scores are large and have the anticipated signs, they are imprecisely estimated.

Similar differences are observed in the management of the payroll system and prompt payment of salaries. We find an average salary delay of roughly three months in the government arm, compared to two months in the NGO arm, and these delays are significantly negatively correlated with test score improvements. Taking the point estimates in Table 9 at face value, an increase in salary delays of one month accounts for one third of the difference in test scores between NGO and government. Related to this, we find significantly higher turnover in the government treatment arm, which is negatively (but not significantly) correlated with test scores.

6.3. Unionization, expectations and credibility of short-term contracts

We hypothesize that teachers' expectations and performance will differ when offered identical contracts by an international NGO or a national government. The effect of a fixed-term contract on teacher performance is likely mediated by teachers' beliefs about the credibility of that contract. Theoretically, short-term teacher contracts are predicated on the operation of dynamic incentives and career concerns (Holmstrom, 1982; Dewatripont et al., 1999a,b). While NGOs may be able to commit to employing teachers only if they perform well, the same contract may lack credibility within a weak public sector bureaucracy and highly unionized civil service system.[36]

Note that we focus on a difference in expectations, not actual union coverage.[37] In response to the government's ambitious plan to hire 18,000 contract teachers, the union filed a lawsuit and launched a series of labor actions which culminated in a national strike and the government conceding to make all these teachers permanent civil servants in September 2011. Formally, teachers employed in our research project were not covered by the negotiations between the government and the teachers' union, and there was no significant difference between treatment arms in the share of teachers employed as civil service teachers following the program. Nevertheless, we hypothesize that teachers in the government treatment arm were more likely to perceive the outcome of the union negotiation – which was ongoing through most of the intervention studied here – as affecting them personally, and further, that the prospect of a permanent unionized job undermined the dynamic incentives provided by a short-term teaching contract in the government treatment arm.

We present three pieces of evidence consistent with the idea that the political backlash had a negative impact on the effectiveness of the contract teacher program that was unique to the government treatment arm. First, we see large and significant differences in union identification between contract teachers in the NGO and government arms (see Panel A, Table 10).[38] Only 15% of teachers in the NGO treatment arm stated that the union represented their interests, while two and a half times as many (almost 40%) of teachers in the government treatment arm believed that the union represented them. Interestingly, this large difference in self-identification with the union is not reflected in any difference in active involvement, such as self-reported participation in the national strike.

[34] Comparing the contract teachers interviewed to a representative sample of contract teachers in Kenyan public primary schools from the World Bank Service Delivery Indicator (SDI) data set, we find teachers in our sample comparable in terms of age and gender ratio. In line with the stipulations of the program, they tend to have more education and training than the average contract teacher and are less likely to be born in the district in which they work. They are thus more representative of recent teacher trainees who are working as contract teachers while waiting for a civil service job.

[35] Reinikka and Svensson (2010) show evidence in the context of health providers in Uganda that is consistent with not-for-profit providers being able to attract more altruistic employees. Similarly, Deserranno (2016) shows that community health worker jobs that signal a more pro-social output attract more intrinsically motivated applicants.

[36] Note that the central role of expectations and dynamic incentives in the performance of contract teachers may make the challenge of scaling up within government institutions particularly challenging for programs of this type, in ways that would be less problematic for, say, providing additional textbooks or school infrastructure.

[37] An alternative hypothesis, suggested by a referee, is that fairness norms rather than expectations or dynamic incentives explain the political backlash we document in this section. Contract teachers working for less money alongside civil service teachers earning more money may view this as particularly unfair if the government is paying both parties.

[38] Note that in the text we use the phrase "self-identification with the union" or simply "union identification" to refer to the response to the question: "Do you believe the union represented your interests throughout the [experimental contract teacher] program?"

Table 9
Hiring, monitoring and implementation.

	Gov.	NGO	Difference	Corr. w/ test score gains
	(1)	(2)	(3)	(4)
Panel A: Socio-economic characteristics				
Age	29.983	29.760	.223	.0007
			(.938)	(.011)
Female	.550	.294	.256	.055
			(.097)***	(.099)
Post-secondary education	.200	.020	.180	−.098
			(.064)***	(.147)
Advanced professional qualification	.100	.137	−.037	.091
			(.061)	(.147)
Friend or relative of teacher or SMC member	.667	.373	.294	.056
			(.100)***	(.101)
Panel B: Monitoring and accountability				
Presence in school	.628	.727	−.099	.098
			(.110)	(.137)
Any monitoring visit to school	.850	.961	−.111	.210
			(.053)**	(.157)
Average salary delay (months)	3.000	2.094	.906	−.057
			(.292)***	(.034)*
Turnover	.714	.455	.260	−.100
			(.111)**	(.091)
Panel C: After the experiment				
Still working at program school	.379	.280	.099	.070
			(.098)	(.106)
Permanent and pensionable	.424	.469	−.046	.126
			(.092)	(.100)
Obs.	60	51	111	102

Summary statistics are based on exit interviews with 111 contract teachers (60 from the government and 51 from the NGO treatment arm, respectively) in 84 treatment schools. Absenteeism is based on 76 observations in treatment schools. Standard errors are clustered at the school level. "Presence in school" = 1 if the teacher was present in school during an announced visit; "Permanent and pensionable" = 1 if the teacher is employed as a civil-service teacher after the end of the RCT. Column 4 reports the coefficient in a regression of changes in test scores between 2009 and 2011 separately on each of the intermediate outcomes and a constant.

Second, we find a strong and significant relationship between union identification and changes in test scores. The difference in test scores between a teacher who felt represented by the union and a teacher who did not accounts almost exactly for the difference in test scores between NGO and government treatment arm. While these estimates are merely correlations, the results are consistent with the hypothesis that the national controversy surrounding the contract teacher scale-up spread to the contract teachers in the government treatment arm and negatively affected their performance, while teachers in the NGO treatment arm were largely immune to the political struggle between the government and the teachers union.

Third, we find that the contract teacher intervention generates reduced effects on learning and intermediate outcomes when contract teachers are exposed to either union representatives or the scaled-up national contract teacher program – but this heterogeneous effect only emerges in the government arm of the experiment. The underlying hypothesis here is that union representatives and contract teachers employed by the government in the national scale-up would signal to experimental teachers in the government treatment arm that the employment guarantee agreed upon by the government and the union would also extend to them.

Consistent with this, results show that contact with the union increases the likelihood of identifying with the union by a statistically significant 50% for teachers in the government treatment arm, but only by a mere 8% for teachers in the NGO treatment arm (column (1) and (2) of Panel B in Table 10). The difference between the two coefficients is significant at the 5% level. Similarly, the presence of one (or more) of the 18,000 contract teachers in a school where the experimental teacher is managed by the government is associated with a 12% higher probability of identifying with the union (though this coefficient is not significant), while the association is exactly zero in a school where the experimental teacher is managed by the NGO.[39] Furthermore, the treatment effect on student learning in the government treatment arm is 0.3 and 0.25 of a standard deviation lower, respectively, if the contract teacher had contact with the union or one of the 18,000 government contract teachers. In the NGO treatment arm, we see no such heterogeneity of effects.[40]

Taken at face value, the results in column (3) and (4) of Table 10 imply that our main result – the performance gap between NGO and government schools in the experiment – was roughly halved where the experimental subjects had only limited exposure to the national scale-up and surrounding controversy, i.e, where experimentally assigned contract teachers in the government treatment arm had no observed interaction with the teacher's union or the 18,000 non-experimental government contract teachers.

To summarize, we examined three hypotheses to explain the performance gap between the government and NGO treatment arms. We found limited evidence to support the idea that the government

[39] A month prior to the end line survey, a national teacher strike took place. One might therefore suppose that the results in our experiment may be driven by the fact that teachers were absent from school during the strike and that teachers in the government treatment arm were more likely to participate in the strike. However, as shown in Panel A of the table, there was no significant difference in reported strike participation between the two treatment arms and no significant effect of strike participation itself on test scores.

[40] The effect of labor strife on productivity has also been documented by Krueger and Mas (2004) in the case of American tire manufacturers. However, we do note that there was no difference in actual union activity (and in particular strike participation) between the two treatment arms, and therefore interpret the observed effect as being a consequence of different expectations as to the credibility of the short-term contract in the two treatment arms.

Table 10
Mechanisms: Political Economy and scaling up.

	Gov.	NGO	Difference	Corr. w/ test score gains
	(1)	(2)	(3)	(4)
Panel A: Summary statistics				
Desire a long-term job	0.632	0.706	−0.074	0.025
			(0.089)	(0.109)
Union represented my interests	0.377	0.149	0.228	−0.205
			(0.089)**	(0.111)*
Took any union action during program	0.428	0.444	−0.017	−0.032
			(0.041)	(0.220)
Union exposure	0.325	0.382	−0.057	−0.119
			(0.062)	(0.138)
Exposure to scale-up	0.333	0.379	−0.051	−0.109
			(0.088)	(0.103)
	Union identification		Test-score gains	
	(1)	(2)	(3)	(4)
Panel B: Regression results				
$Z \times$ Gov	0.084	0.157	−0.072	−0.075
	(0.101)	(0.116)	(0.152)	(0.120)
$Z \times$ NGO$\times$ Exposure to union	0.083		0.042	
	(0.120)		(0.186)	
$Z \times$ Gov$\times$ Exposure to union	0.548***		−0.292*	
	(0.168)		(0.158)	
$Z \times$ NGO$\times$ Exposure to gov't scale-up		−0.009		0.029
		(0.115)		(0.145)
$Z \times$ Gov$\times$ Exposure to gov't scale-up		0.121		−0.263*
		(0.154)		(0.143)
Observations	100	95	102	107

Panel A: "Union represented my interests" = 1 if the teacher said yes to, "Do you believe the union represented your interests throughout the [experimental contract teacher] program?"; "Desire for long-term employment" = 1 if the teacher mentioned long-term employment as their main expectation from the program; and "Took any union action during program" is the average of the following dummy variables: the teacher was a union member during the program; teacher could explain the purpose of union strike action against the contract teacher program; teacher reports participation in the national strike in 2011. "Union exposure" is the weighted average of the following dummy variables: "Was the school ever visited by a union representative?" and "Did the teacher ever attend a union meeting?". "Exposure to gov't scale-up" is a an indicator variable taking a value of 1 if one (or more) of the 18,000 (non-experimental) government contract teachers was also placed in the school. Column 4 reports the coefficient in a regression of changes in test scores between 2009-2011 separately on each of the intermediate outcomes and a constant. Panel B: The dependent variable in column (1) and (2) is union identification, which is a dummy variable set equal to 1 if the teacher said that the union represented his/her interests during the program, and zero otherwise. The dependent variable in column (3) and (4) is changes in test scores between 2009 and 2011. *Z* takes a value of 0 at baseline for all schools, and 1 in the follow-up survey only if the school was assigned to any treatment arm; *Gov* is an indicator variable for the government treatment arm. Standard errors are clustered at the school level.

program failed due to recruiting lower quality teachers, and somewhat stronger evidence that limited monitoring and accountability in the government program undermined results. Note that we characterize both of these mechanisms as features of working with the Kenyan government, regardless of scale. Finally, we presented a variety of evidence that the government program failed in part due to the political backlash it provoked. We consider this a function of going to scale *per se*, and argue that the measurable effects of the political backlash account for roughly half of the NGO-government performance gap.[41] The results of the implementation of the contract teacher program in Kenya, though only suggestive, are thus largely consistent with the "seesaw effect" stressed by Acemoglu (2010): large-scale policy interventions of this sort are likely to provoke political economy reactions from groups whose rents are threatened by reform, creating an endogenous policy response that counteracts the objectives of reform.

6.4. After the program

In panel C of Table 9, we examine the experience of contract teachers following the program. We find that about half the contract teachers have progressed to permanent and pensionable civil service status and about a third of teachers have been retained by program schools with no significant differences between intervention arms. We find a positive, but not significant relationship between teacher value added and subsequent civil service status (and a smaller positive relationship between performance and retention), which is of similar size across implementing agency. This is consistent with findings by Duflo et al. (2015) who find a strong relationship between teacher performance and subsequent progression to civil service status, but also our own understanding of the civil service employment procedure, which uses an algorithm for civil service employment that is heavily weighted towards time passed since graduation with little weight given to prior teaching experience or performance (Barton et al., 2017).

[41] This political dynamic is by no means unique to Kenya: teacher unions tend to be strong and vociferous opponents to accountability reforms in many countries (see Murillo (1999) and Bruns and Lucque (2014) for examples from Latin America, and cases in India (Compton and Weiner, 2012a,b) and the US (Barr, 2006). Theoretically, the issue of unions opposing such reforms is examined in Lindbeck and Snower (1989) and stated in general form in Lavy (2007). However, there are also examples where unions have collaborated in accountability reforms: during the 1990s and 2000s, the Chilean government implemented an ambitious reform of the education sector in cooperation with the teacher union (see Mizala and Schneider, 2014).

7. Conclusion

To the best of our knowledge, this paper is the first attempt to employ experimental methods to test organizational and political economy limitations to translating NGO tested programs to government implementation at national scale. We report on a randomized trial showing that contract teachers significantly raise pupil test scores when implemented by an international NGO. These effects disappear when the program is (a) implemented within the bureaucratic structures of the Kenyan government and (b) extended to a national scale. We show evidence that this is driven by the concomitant political response from vested interests opposed to the program.

Methodologically, notable caveats in our experimental design include the lack of pupil-level panel data, and statistically insignificant but non-trivial baseline imbalance in some outcomes despite randomization. We present evidence that student attrition and accretion do not drive our main effects, and that the NGO treatment led to learning gains after controlling for schools' baseline learning levels.

We also study several variants of the basic contract teacher intervention, randomizing devolution of hiring and firing authority to the school, training for school management committees, and teacher salary levels. Statistical power here is limited, results are mixed, and not consistently statistically significant, but there is some evidence that the combination of higher salaries and SMC training (and in the NGO treatment arm, devolution of hiring authority) produced much larger treatment effects. Notably, these elements were not adopted in the government's national scale-up of the program.

Our results are consistent with the hypothesis that the government, subject to union pressure, would struggle to credibly enforce teacher contracts. But our evidence stems from one particular Kenyan government institution, under pressure from a strong public sector union, compared to a well-established international NGO. We would be cautious in generalizing our results to the Kenyan government in its entirety, much less to developing-country governments and NGOs more broadly. In fact, a recent example from Kenya of a successful scale-up by government is the national deworming campaign inspired largely by the work of Miguel and Kremer (2004). So while our results are not intended to suggest a universal ranking of organizational effectiveness placing NGOs above public institutions, they do suggest a dimension of external validity of program evaluation to which future policy-oriented research should be attentive: namely an examination of government implementation when trying to understand the impact of accountability reforms and incentive programs, especially so in developing countries with weak public sector institutions. Our paper is an attempt to do just that using experimental methods.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpubeco.2018.08.007.

References

Acemoglu, D., 2005. Politics and economics in weak and strong states. J. Monet. Econ. 52, 1199–1226.

Acemoglu, D., 2010. Theory, general equilibrium, and political economy in development economics. J. Econ. Perspect. 24 (3), 17–32.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Mukherji, S., Shotland, M., Walton, M., 2016. Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India. Working Paper, J-PAL.

Banerjee, A., Duflo, E., Glennerster, R., 2008. Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system. J. Eur. Econ. Assoc. 6 (2-3), 487–500.

Barr, S., June 2006 Unions Oppose Senate's Pay-for-Performance Bill. Published online, June 30, 2006, at http://www.washingtonpost.com/wp-dyn/content/article/2006/06/29/AR2006062902029.html.

Barton, N., Bold, T., Sandefur, J., 2017. Measuring the Rents from Public Employment: Evidence from Kenya.

Besley, T., Persson, T., 2011. Pillars of Prosperity. Princeton University Press, Princeton.

Bold, T., Filmer, D., Martin, G., Molina, E., Stacy, B., Rockmore, C., Svensson, J., Wane, W., 2016. Enrollment without learning: teacher effort, knowledge, and skill in primary schools in Africa. J. Econ. Perspect. 31, 185–204. No. 4-Fall 2017.

Bold, T., Kimenyi, M., Mwabu, G., Sandefur, J., 2011. Why did abolishing fees not increase public school enrollment in Kenya? Center for Global Development Working Paper. 271.

Bruhn, M., McKenzie, D., 2009. In pursuit of balance: randomization in practice in development field experiments. Am. Econ. J. Appl. Econ. 1 (4), 200–232. October.

Bruns, B., Lucque, J., 2014. Great Teachers: How to Raise Student Learning in Latin America and the Carribean, Advance Edition. The International Bank for Reconstruction and Development / The World Bank, Washington, DC.

Chaudhury, N., Hammer, J., Muralidharan, K., Rogers, F.H., 2006. Missing in action: teacher and health worker absence in developing countries. J. Econ. Perspect. 20 (1), 91–116.

Compton, M., Weiner, L., October 2012a. More Police Attacks on Kashmiri Teachers. Published online, Oct. 08, 2012, at http://www.teachersolidarity.com/blog/more-police-attacks-on-kashmiri-teachers/.

Compton, M., Weiner, L., October 2012b, Striking Indian Contract Teachers Won't be Intimidated. Published online, Oct. 31, 2012, at http://www.teachersolidarity.com/blog/striking-indian-contract-teachers-wont-be-intimidated/.

Conn, K., 2014. Identifying effective education interventions in sub-saharan africa: a meta-analysis of rigorous impact evaluations. mimeo, Columbia University.

de Ree, J., Muralidharan, K., Pradhan, M., Rogers, H., 2018. Double for nothing? Experimental evidence on an unconditional teacher salary increase in Indonesia. Q. J. Econ. 133 (2), 993–1039. https://doi.org/10.1093/qje/qjx040. 1 May.

Deserranno, E., 2016. Financial incentives as signals: experimental evidence from the recuritument of health promoters in uganda. mimeo, Northwestern University.

Dewatripont, M., Jewitt, I., Tirole, J., 1999a. The economics of career concerns, part I: comparing information structures. Rev. Econ. Stud. (66), 183–198.

Dewatripont, M., Jewitt, I., Tirole, J., 1999b. The economics of career concerns, part II: application to missions and accountability of government agencies. Rev. Econ. Stud. (66), 199–217.

Duflo, E., Dupas, P., Kremer, M., 2015. School governance, teacher incentives, and pupil-teacher ratios: experimental evidence from kenyan primary schools. J. Public Econ. 123.

Duflo, E., Hanna, R., Ryan, S.P., 2012. Incentives work: getting teachers to come to school. Am. Econ. Rev. 102 (4), 1241–1278.

Egbewale, B.E., 2015. Statistical issues in randomized controlled trials: a narrative synthesis. Asian Pac. J. Trop. Biomed. 5 (5), 354–359.

Ferraz, C., Finan, F., 2009. Motivating politicians: the impacts of monetary incentives on quality and performance. NBER Working Paper. 14906.

Finan, F., Olken, B.A., Pande, R., 2015. The personnel economics of the state. Technical Report. National Bureau of Economic Research.

Glewwe, P., Ilias, N., Kremer, M., 2010. Teacher incentives. Am. Econ. J. Appl. Econ. 2, 205–227.

Glewwe, P., Muralidharan, K., 2015. Improving school education outcomes in developing countries: evidence, knowledge gaps, and policy implications, Rise-WP-15/001.

Greevy, R., Lu, B., Silber, J., Rosenbaum, P., 2004. Optimal multivariate matching before randomization. Biometrika 5 (2), 263–275.

Holmstrom, B., 1982. Managerial incentive problem: a dynamic perspective. Essays in Economics in Honor of Lars Wahlbeck. Swedish School of Economics, Helsinki.

Kremer, M., Brannen, C., Glennerster, R., 2013. The challenge of education and learning in the developing world. Science 340.6130, 297–300.

Krishnaratne, S., White, H., Carpenter, E., 2013. Quality education for all children? What works in education in developing countries? International Initiative for Impact Evaluation (3ie) Working Paper 20, New Delhi.

Krueger, A.B., Mas, A., 2004. Strikes, scabs, and tread separations: labor strife and the production of defective bridgestone/firestone tires. J. Polit. Econ. 112 (2), 253–289.

Lavy, V., 2007. Using performance-based pay to improve the quality of teachers. Futur. Child. 17 (1), 87–109.

Lindbeck, A., Snower, D.J., 1989. The Insider-Outsider Theory of Employment and Unemployment. MIT Press Books vol. 1. The MIT Press.

Mbiti, I.M., 2016. The need for accountability in education in developing countries. J. Econ. Perspect. 30 (3), 109–132.

McEwan, P., 2015. Improving learning in primary schools in developing countries: a meta-analysis of randomized experiments. Rev. Educ. Res. 85 (3), 353–394. https://doi.org/10.3102/0034654314553127.

Miguel, E., Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. Econometrica 72 (1), 159–217.

Mizala, A., Schneider, B.R., 2014. Negotiating education reform: teacher evaluations and incentives in Chile (1990–2010). Gov. Int. J. Policy Adm. Inst. 27 (1), 87–109.

Mugo, J., Kaburu, A., Limboro, C., Kimutai, A., 2011. Are our children learning: annual learning assessment report. Technical Report. Uwezo, Kenya.

Muralidharan, K., Romero, M., Wuthrich, K., 2018. Cross-cutting Treatments and (Incorrect) Inference in Experiments, mimeo, UCSD.

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

Muralidharan, K., Sundararaman, V., 2013. Contract teachers: experimental evidence from India. NBER Working Paper Series. 19440.

Murillo, M.V., 1999. Recovering political dynamics: teachers' unions and the decentralization of education in Argentina and Mexico. J. Interamerican Stud. World Aff. 41 (1), 31–57.

OECD, 2012. Credit Reporting System (CRS) Database. http://stats.oecd.org/Index.aspx?datasetcode=CRS1 Accessed March.

Otieno, W., Colclough, C., 2009. Financing education in Kenya: expenditure, outcomes and the role of international aid by. Research Consortium on Educational Outcomes & Poverty Working Paper. 25.

Piper, B., Miksic, E., 2011. Mother Tongue and Reading: Using Early Grade Reading Assessments to Investigate Language-of Instruction Policy in East Africa. In: Gove, A., Wetterberg, A. (Eds.), The Early Grade Reading Assessment: Applications and Interventions to Improve Basic Literacy. RTI Press, https://files.eric.ed.gov/fulltext/ED531301.pdf.

Reinikka, R., Svensson, J., 2010. Working for God? Evidence from a change in financing of nonprofit health care providers in Uganda. J. Eur. Econ. Assoc. 8 (December), 1159–1178.

Roberts, C., Torgerson, D.J., 1999. Baseline imbalance in randomised controlled trials. Br. Med. J. 319.

Siringi, S., August 2007. Kenya: Exposed - Country's 32,000 Ghost Teachers. Published online, Aug. 11, 2007, at http://allafrica.com/stories/200708110007.html.