Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education

Tessa Bold, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur

Abstract

The recent wave of randomized trials in development economics has provoked criticisms regarding external validity. We investigate two concerns—heterogeneity across beneficiaries and implementers—in a randomized trial of contract teachers in Kenyan schools. The intervention, previously shown to raise test scores in NGO-led trials in Western Kenya and parts of India, was replicated across all Kenyan provinces by an NGO and the government. Strong effects of short-term contracts produced in controlled experimental settings are lost in weak public institutions: NGO implementation produces a positive effect on test scores across diverse contexts, while government implementation yields zero effect. The data suggests that the stark contrast in success between the government and NGO arm can be traced back to implementation constraints and political economy forces put in motion as the program went to scale.

JEL Codes: I20, I25, I28

Keywords: education, randomized control trials, Kenya.

Center for Global Development
www.cgdev.org

Working Paper 321
March 2013

Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education

Tessa Bold
Mwangi Kimenyi
Germano Mwabu
Alice Ng'ang'a
Justin Sandefur

We are indebted to the staff of the Ministry of Education, the National Examination Council, and World Vision Kenya, and in particular to Mukhtar Ogle and Salome Ong'ele. Julia Bless, Anne Karing, Naureen Karachiwalla, Phyllis Machio, Rachael Musitia, Joachim Münch and Diogo Weihermann provided excellent research assistance. Paul Collier, Stefan Dercon, Geeta Kingdon, David Johnson, and Andrew Zeitlin helped conceive this project. Michael Clemens, Michael Kremer, Karthik Muralidharan, Paul Niehaus, Lant Pritchett, David Roodman, Torsten Persson, Jakob Svensson and numerous seminar participants provided invaluable feedback. 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.

Bold: Institute for International Economic Studies, Stockholm University and Goethe University Frankfurt, tessa.bold@iies.su.se. Kimenyi: Brookings Institution, Washington D.C., kimenyi@brookings.edu. Mwabu: Department of Economics, University of Nairobi, gmwabu@gmail.com. Ng'ang'a: Strathmore University, Nairobi, alicemnganga@yahoo.com. Sandefur: Center for Global Development, Washington D.C., jsandefur@cgdev.org.

CGD is grateful to its funders and board of directors for support of this work.

Tessa Bold, Mwangi Kimenyi, and Germano Mwabu. 2013. "Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education." CGD Working Paper 321. Washington, DC: Center for Global Development.
http://www.cgdev.org/publication/scaling-what-works

Center for Global Development
1800 Massachusetts Ave., NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

1 Introduction

The recent wave of randomized trials in development economics has catalogued a number of cost-effective, small-scale interventions found to improve learning, health, and other welfare outcomes. Surveying this growing literature, Banerjee and He (2008) offer a menu of proven interventions which, at current spending levels on international development aid, could be scaled-up across the developing world.

Critics argue that these studies produce internally valid measures of the causal effect of an intervention ("it worked here"), but question their external validity ("it will work there") (Cartwright and Hardie 2012, Deaton 2010). In this paper we report on a randomized trial designed to assess two obstacles to external validity which arise in the translation of experimental evidence into policy advice. First, heterogeneous treatment response implies that estimates for one beneficiary population may not apply to other populations. Second, even with a homogenous population, treatment effects may be a function of the institution that implements the intervention. This may reflect efficiency differences between implementing organizations (Heckman 1991, Allcott and Mullainathan 2012), or the fact that government implementation at national or regional scale may induce general equilibrium effects or political economy responses (Acemoglu 2010). These obstacles are particularly relevant when national or global policy prescriptions are drawn from randomized trials of pilot projects, often conducted in non-randomly chosen locations and implemented by well-organized non-governmental organizations.

We report on a randomized trial embedded within the national scale-up of a contract teacher program by the Kenyan government. We compare the effectiveness of NGO and government implementation, and test for the presence of heterogeneous treatment effects in a nationwide sample. The question of NGO versus government implementation is paramount to the formulation of public policy, as governments are frequently the only institutional actors capable of taking education policies to scale.[1] Yet, within the growing literature of randomized impact evaluations on public service delivery in the developing world, trials involving governments rarely focus on accountability reforms or changes to provider incentives.[2] Instead, the effects of such policies are commonly forecast on the

[1] In Kenya, for instance, government schools account for 90.2% of gross primary enrollment. Furthermore, as of 2005 the Ministry's budget for primary education totalled $731 million (Otieno and Colclough 2009), compared to $4 million in international aid for education routed through NGOs in 2009 (OECD 2012).

[2] We examined 31 studies which measure the impact of education interventions on learning

basis of NGO-run pilots testing a theoretical mechanism that operates within, e.g., a clinic or school.[3] The external validity of the results from these NGO pilots hinges on the assumption that incentives for front-line service providers do not interact with the broader organizational and political context, allowing for extrapolation to the public sector.

At the school level, the experiment presented here replicates one of the most extensively tested, successful interventions to raise student learning in primary schools: the provision of contract teachers. Banerjee et al. (2007) present results from a randomized evaluation showing that an NGO program in urban India hiring young women to tutor lagging students in grades 3 and 4 led to a 0.28 standard deviation increase in tests scores. Muralidharan and Sundararaman (2010) evaluate a state-wide program in Andhra Pradesh, finding that hiring an extra contract teacher leads to an increase in test scores of 0.15 and 0.13 standard deviations on math and language tests. In both cases, the additional teachers lead to significant learning gains despite salary costs that are a small fraction of

outcomes in developing countries, roughly half of which cite significant government involvement in project implementation. Trials involving governments tend to focus on increasing inputs: governments have been directly involved in evaluations of the learning impacts of conditional cash transfer programs in Ecuador (Paxson and Schady 2007), Malawi (Baird, McIntosh and Özler 2011), and Nicaragua (Macours, Schady and Vakis 2011). Other studies have evaluated government programs involving school meals (Kazianga, de Walque and Alderman 2009), use of ICT in the classroom in Chile (Rosas, Nussbaum, Cumsille, Marianov, Correa, Flores, Grau, Lagos, Lopez, Lopez, Rodriguez and Salinas 2003) and Colombia (Barrera-Osorio and Linden 2009), provision of eye-glasses in China (Glewwe, Park and Zhao 2012), school construction in Afghanistan (Burde and Linden 2012) and reforms to local school management in Madagascar (Glewwe and Maïga 2011).

[3]In India, RCTs have examined NGO programs to encourage parental involvement in schools (Pandey, Goyal and Sundararaman 2008, Banerjee, Banerji, Duflo, Glennerster, and Khemani 2010), changes to the English and reading curriculum (He, Linden and MacLeod 2008, He, Linden and MacLeod 2009), use of information technology in the classroom (Linden, Banerjee and Duflo 2003, Inamdar 2004, Linden 2008), teacher performance pay (Muralidharan and Sundararaman 2011), student and parent incentives (Berry 2011), cameras in schools to discourage teacher absenteeism (Duflo, Hanna and Ryan 2012b), and as discussed below, contract teachers or tutors (Banerjee, Cole, Duflo and Linden 2007, Muralidharan and Sundararaman 2010). Similarly in Kenya, NGO pilot programs have examined the impact of contract teachers and tracking students (Duflo, Dupas and Kremer 2011), teacher incentives (Glewwe, Ilias and Kremer 2010), student incentives (Kremer, Miguel and Thornton 2009), physical school inputs (Kremer, Moulin and Namunyu 2003, Glewwe, Kremer, Moulin and Zitzewitz 2004, Glewwe, Kremer and Moulin 2009), and school meals (Vermeersch and Kremer 2005), while in Uganda Barr, Mugisha, Serneels and Zeitlin (2011) report on an RCT of an NGO program to facilitate community monitoring of schools. Notable exceptions to this pattern of government evaluations focusing on increased inputs that we are aware of include the evaluation of a World Bank-financed school management reform program in Madagascar cited above (Glewwe and Maïga 2011).

civil service wages. Finally, of particular relevance for the present study given its geographic focus, Duflo, Dupas and Kremer (2012a) show that exposure to an NGO-managed contract teacher program in government schools in Western Kenya raises test scores by 0.21 standard deviations relative to being taught by civil service teachers. Furthermore, their experimental design allows them to attribute this effect to contract teachers per se, rather than the accompanying reduction in class size from hiring an extra teacher.[4]

In 2009, the Kenyan government announced a nationwide contract teacher program that would eventually employ 18,000 teachers. We report on a randomized experiment embedded within this program, designed to test the government's ability to implement a fairly close variant of the NGO project described by Duflo et al. (2012a) and to replicate the results across diverse conditions. As part of the experimental evaluation, 192 schools were chosen from across 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. The random assignment of schools to NGO versus government implementation, which is at the center of this study, was overlayed by additional treatment variations in salary levels and recruitment procedures.

We find positive and significant effects of the program only 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, comparable in magnitude to the results in Muralidharan and Sundararaman (2010) in India and Duflo et al. (2012a) in Western Kenya. Treatment effects were significantly smaller and indistinguishable from zero in schools receiving contract teachers from the Ministry of Education.

What explains the difference in treatment success between the government and the NGO? We investigate a number of potential mechanisms and find (suggestive) evidence showing that the stark contrast in success between the government and NGO arm can be traced back to implementation constraints in the public sector and political economy forces put in motion as the program went to scale.

[4]See Bruns, Filmer and Patrinos (2011) for a summary of additional, non-experimental results on the impact of contract teachers, including Bourdon, Frölich and Michaelowa (2007) who find positive (negative) test-score effects on low (high) ability pupils in Mali, Niger, and Togo, and Goyal and Pandley (2009) who find contract teachers are equally or more likely to be present and teaching relative to civil service teachers in Madhya Pradesh and Uttar Pradesh, India, and that this higher effort is correlated with pupil performance. Finally, Atherton and Kingdon (2010) find that contract teachers in Indian schools perform better than regular teachers.

In contrast to earlier small-scale pilots, the prospect of a nationwide contract teacher program with 18,000 new contract teachers provoked organized resistance from the national teachers union.[5] As Acemoglu (2010) notes, 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 - - the "seesaw effect". More specifically, while a small number of contract teachers can be employed at wages far below civil service levels, a large cohort of contract teachers becomes politically potent and able to demand civil service protections. We find evidence that union actions to demand permanent civil service employment and union wages had a differential effect on teachers employed by the government and the NGO during the experimental evaluation, although neither were formally covered by union collective bargaining. Teachers in the government treatment arm were more likely to report that the union represented their interest, and this self-identification was significantly, negatively correlated with improvements in pupil test scores. Moreover, we find that test scores are lower where teachers in the government treatment arm (but not the NGO treatment arm) are in closer contact with the union or one of the 18,000 (non-experimentally allocated) teachers hired by the government in the national scale-up. We interpret this as evidence that the controversy surrounding the national scale-up adversely affected the credibility of dynamic incentives for teachers already employed by the government in the experiment, in turn lowering their performance.

We further show that monitoring and implementation of the program may have been compromised in several ways in the government treatment arm. Duflo et al. (2012a) find that nepotism and local capture undermine contract teacher performance. We show that nepotism and local capture are significantly higher in schools in the government treatment arm, but fail to replicate any correlation between these variables and test score gains. We also find that schools in the government treatment arm received fewer monitoring visits, and teachers experienced longer salary delays, though of these intermediate indicators, only salary delays were significantly, negatively correlated with improvements in pupil test performance.

Finally, we also exploit our nationwide sampling frame to measure the extent of heterogeneity of treatment response across schools, a more traditional concern regarding external validity in the experimental literature. Comparing treatment

[5] This political dynamic is not unique to Kenya. Unions in both the US and various states of India have opposed short-term contracts in recent years (Compton and Weiner 2012a, Compton and Weiner 2012b, Barr 2006). Where unions are more amenable to reforms, e.g. the case of on-going research by J-PAL in Ghana, more favorable results may be anticipated.

effects across different localities and baseline characteristics, we find no reason to question the external validity of earlier studies on the basis of their geographic scope.

Overall, our results confirm the findings of Duflo et al. (2012a) and Muralidharan and Sundararaman (2010), among others, regarding the ability of contract teachers to significantly improve learning in public primary schools across diverse baseline conditions in a low-income country. But the effects of this intervention appear highly fragile to the involvement of carefully-selected non-governmental organizations.

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 on both intention-to-treat (ITT) effects and average treatment effects for the treated (ATT), where actual treatment is defined as successfully recruiting a contract teacher. Section 6 tests for heterogeneous treatment effects across space and initial conditions. Section 7 explores possible mechanisms explaining the government-NGO performance gap. Section 8 concludes.

2 Context

Primary school enrollment is relatively high in Kenya, but learning levels in primary schools are poor. According to the most recent available national household survey from 2006, net primary enrollment was 81%, with government primary schools accounting for 72% (Bold, Kimenyi, Mwabu and Sandefur 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, Kaburu, Limboro and Kimutai 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 of 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. In some cells of the experimental design described below, funds to hire a contract teacher were transferred to this school bank account managed by the SMC, and SMC members participated in a training program to increase local accountability.

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. The combination of permanent contracts, direct payment from Nairobi and teacher shortages leads to limited local accountability for civil service teachers.

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

[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.

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 this 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 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 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.

3 Experimental 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.

3.1 Program details

Contract teachers were randomly assigned to teach either grade 2 or 3.[7] 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. As discussed below, compliance with these instructions was high but imperfect.

The experimental sample focuses on 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.[8] Within each province, schools with a pupil-teacher ratio below

[7] Half of the teachers in the experiment were assigned to grade 2 in 2010, and half to grade 3 in 2010. In 2011, all the contract teachers were placed in grade 3.

[8] 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

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.

The effects of the randomized interventions are measured by comparing baseline and follow-up academic assessments in math and English in all 192 schools. The survey instruments were designed with the collaboration of 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 August of 2009 by the KNEC and the research team. 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. Roughly 15,000 students were tested in the baseline and follow up survey.

3.2 Treatment variations

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

High versus low salary

Out of the total 128 contract teacher positions created, 96 were offered KES 5,000 ($67) per month, while 32 were offered KES 9,000 ($121) per month. 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. The salary variation was designed to explore to what extent salary was linked to performance and the Ministry's ability to reduce payroll costs without sacrificing teacher performance.

Central versus local hiring and payment

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

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).

Ministry or World Vision headquarters in Nairobi and district education officers and NGO officials, respectively, were responsible for selecting candidates.

School management committee training

Following Duflo et al. (2012a), we explored the importance of local accountability on teacher (and in turn, student) performance with a training intervention. We invited members of the school management committee in half of the treatment schools to a two-day training workshop. While school management committees have formal responsibility to monitor teachers and school finances, many parental representatives are unaware or ill-equipped to perform these duties. The training program drew on manuals developed by World Vision and the Ministry of Education, with a 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 performance.

3.3 Randomization

To guarantee that the sample is balanced between treatment and control schools, an optimal multivariate matching algorithm was used (see Greevy, Lu, Silber and Rosenbaum (2004) and Bruhn and McKenzie (2009)). Treatment and control schools were matched along the following dimensions: results in nationwide end-of-primary leaving exams, baseline scores on the grade 1 test, pupil-teacher ratio, number of classrooms, number of civil service teachers, number PTA teachers and average pay of teachers employed by Parent-Teacher Associations at baseline. The algorithm created groups of 3 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. Figure 1 in the appendix shows the distribution of schools assigned to the control group and government or NGO implementation across the eight provinces.

We also check whether randomization was successful in achieving balance on baseline indicators that were not explicitly used in the matching algorithm, namely, the outcome variable of interest, average standardized test scores at baseline. The successful outcome of the randomization is reported in Table 1, where we compare treatment and control schools, as well as schools assigned to the government and the NGO treatment arm.

3.4 Sample attrition

192 schools were part of the intervention, and all participated in the program. However, both in the baseline sample and in the follow up sample, we were not able to survey all the schools. In the baseline survey, 16 schools (7 in the government treatment arm, 7 in the NGO treatment arm and 2 in the control group) were not visited. In the follow up survey, 4 schools (1 in the government treatment arm, 1 in the NGO treatment arm and 2 in the control group) were not visited.

The main reason for not being able to include these schools is location. With one exception, they are all located in the remote Eastern or North Eastern province, places which are difficult to reach because of transport and in particular security conditions.[9] Comparing excluded and included schools on the basis of administrative data, we find no significant differences in terms of national end-of-primary-leaving exam scores and pupil teacher ratios.

Since the baseline survey was conducted 9 months before the randomization, missing schools in the baseline present no threat to the internal validity of our study, even though the size of the school panel is of course reduced. In total, there are 174 schools for which data is available both in the baseline and the follow-up. The fact that 4 schools, or 2 percent, attrited in the follow up survey is potentially more of a concern, since there could be some correlation between the treatment effect and the likelihood of not being surveyed. We explore this issue by estimating the treatment effect using Lee bounds in Section 5.4.

4 Compliance and Contamination

Random assignment of a school to the treatment group created a job vacancy for a contract teacher. The onus then fell on district and school officials to recruit a suitable teacher, place him or her in either grade 2 or 3, and split that grade into two (or more) streams. Compliance was imperfect at each of these steps. While compliance is of independent interest, it may also help explain differences in treatment effects between the NGO and government documented in the following section.

[9] An alternative would be to exclude schools with missing baseline information from the study. Results on the reduced sample are identical and available from the authors upon request.

4.1 Teacher recruitment

The 128 schools assigned to receive a contract teacher as part of the experimental evaluation had mixed success in recruiting and retaining contract teachers. The proportion of vacancies filled varied by salary level and recruitment method, and between government and NGO implementation.

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. However, 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 17 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 3). The difference between the government and the NGO arms, 1.4 months, is insignificant.

Table 2 examines the vacancy rate more closely, modeling success in filling a vacancy as a function of various demand-side policies that were manipulated by the experiment, as well as other exogenous and/or predetermined school characteristics. The dependent variable is a binary indicator of whether a teacher was present and teaching in a given school in a given month, with monthly observations spanning the duration of the experiment from June 2010 to October 2011. We estimate both a linear probability model and a logit model, with and without controls for school characteristics.

We examine three experimental determinants of teacher labor supply. First, Table 2 shows that NGO implementation led to between 12 and 14% more months with a filled vacancy, relative to the government treatment arm, and this effect is significant across all specifications.[10] 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 to 15% across specifications. Third, offering a "high" salary increases the probability of filling a teaching vacancy by just under 12%. This effect is significant and consistent between the LPM and logit models, but not robust to the inclusion of school-level controls in columns 3 and 6. In addition, the correlation between the probability of filling the teacher vacancy in our intervention and the general thickness of the labor market – measured as the ratio of applicants to vacancies

[10] Note that while this difference is significant in both the LPM and logit model using disaggregated monthly data, the difference between the average months a teacher was employed in the NGO and government treatment arms is not significant when the data is collapsed to the school level in Table 3.

for the 18,000 teachers hired in 2010 – is positive and significant.[11] This provides further evidence that failure to recruit a teacher was sensibly related to local labor market conditions.

4.2 Reallocation within treatment schools

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. Our ability to measure both effects using test-score data on the target cohort of pupils, however, hinges on schools' willingness to comply with the intervention by (a) placing the contract teacher in the correct grade, and (b) not reallocating the existing teacher for that grade, such that the class-size reduction is concentrated on the treatment cohort.[12]

Table 3, Panel B, reports whether schools complied with the intervention protocol in terms of placing the teacher in grade 2 and 3 and splitting the grade. Schools largely followed the instructions on the former point, but less so on the latter. 95% of teachers were employed in the correct grade at least some of the time and 72% were employed in the correct grade all of the time. In contrast, class sizes in the treatment grades were not reduced significantly.

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 gives us some confidence that results will not be driven by the inability (or unwillingness) of the implementing agency to conduct randomized controlled trials, or by a class-size mechanism.

[11] This is the coefficient in a regression of presence of a teacher on labor market thickness and a constant. It is significant at the 1% level with standard errors clustered at the school level.

[12] For comparison, note that the contract teacher evaluation described in Muralidharan and Sundararaman (2010) adhered to a fairly strict 'business as usual' protocol, whereby 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, Duflo et al. (2012a) deviate from business as usual, intervening within schools to ensure that contract teachers are assigned to a given grade, existing teachers are not reassigned so that class size reductions are maximized for the target pupils, and dictating the allocation of pupils between civil service and contract teachers. This allows the latter authors to separate class size effects from the effect of a contract teacher on a given class size.

4.3 Reallocation across schools

A common concern in many experimental evaluations is that the government or other implementing partner will reallocate resources (in this case, teachers) to compensate the control group. Here we consider the reallocation of teachers across schools, as well as endogenous movement of pupils in response to the contract teacher 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. 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 3, 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.

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 3 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 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. In addition, there is no *a priori* reason to suspect this phenomenon would differ systematically between the NGO and government treatment arm, which is the comparison of primary focus in what follows.

To summarize, we find that the contract teacher job vacancies created by this experimental program were filled in roughly 70% of months overall, with about a 12% higher success rate in filling vacancies on the NGO side. Teachers were

overwhelmingly placed in the correct grade, but they were often asked to cover additional grades as well. Existing teachers were often reallocated within schools to spread the teaching load evenly, yielding very small net changes in class size in our sample. None of these reallocations differed between the NGO and government treatment arm. Finally, there is no 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, and conclude these effects are unlikely to reflect reductions in class size. The possibility remains that differences between the NGO and government arm may be attributable to differences in recruitment success, which we explore further below.

5 Comparing the effectiveness of government and NGO programs

As noted in the introduction, scaling up successful education programs in many low-income countries requires 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 replicate successful NGO pilots. Section 5.1 presents the reduced-form treatment effects for the program as a whole, and the direct comparison of the NGO and government treatment arms. Given the performance gap between the government and NGO treatment arms in the ITT estimates, an obvious question arises as to whether this disparity can be explained by poor compliance, i.e., a failure to fully implement the program in the government treatment arm. Section 5.2 presents instrumental variables estimates of the impact of actual presence of a contract teacher in a given school in a given month, (as opposed to mere random assignment) on student performance in both the NGO and government treatment arms. We find that differences in compliance between the government and NGO program do nothing to explain differences in treatment effects.

5.1 ITT effects

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} is the score on a math and English test 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 equation (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 equation (1) with three alternative sets of controls ($\mathbf{X}_{jt}$): first, a simple cross-sectional OLS regression with no controls; second, controlling for initial test scores averaged at the school level, $\bar{Y}_{j,t-1}$; and third, a school-level fixed effects regression. While the cross-sectional regression without controls provides a consistent estimate of β_1 due to randomization, controlling for variations in initial conditions and focusing on relative changes over time using the lagged-dependent variable and fixed effects models may improve power and precision.

Columns 1 to 3 of the top panel of Table 4 present the results for each of these three estimates of the average ITT effect. The point estimate is fairly consistent across all three specifications, at approximately 0.1 standard deviations and just insignificant in all three specifications.

The bottom panel of Table 4 repeats estimation from the top panel, allowing for the effect to differ by implementing agency. In each case, we regress scores on a treatment variable and the treatment variable interacted with a dummy for government implementation. Thus for the ITT we estimate

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

As above, we estimate three variations of each of these equations with varying sets of controls ($\mathbf{X}_{jt}$)

Results from specifications including either school fixed effects or a lagged dependent variable both show that the overall effect of a contract teacher is driven by the NGO program, with essentially zero effect in the government treatment arm. Columns 1 to 3 in the bottom panel of Table 4 compare the effect of assignment to NGO versus government implementation of the project. The coefficient on Z_{jt} shows that NGO implementation raises scores by 0.15 to 0.18 standard deviations. This coefficient is statistically significant at the 5% level in the lagged dependent variable and fixed effects models, and at the 10% level in the cross-section. The coefficient on $Z_{jt} \times \text{Gov}_{jt}$ shows the relative effect of moving from NGO to government implementation. This effect is consistently

negative, and statistically significant at the 10% level in the lagged dependent variable model and at the 5% level in the fixed effects model. Adding the coefficients on Z and $Z \times$ Gov gives the simple ITT effect within the government sample, which is statistically insignificant across all three specifications, and has a combined point estimate of almost precisely zero in the lagged dependent variable and fixed effects model.

Figure 2 shows this result graphically, comparing the kernel density of test score changes between control schools and all treatment schools (top panel) and between government and NGO treatment schools (bottom panel). 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.

5.2 IV estimates

Can differences in the probability of filling contract teacher vacancies described in Section 4.1 explain the difference in government and NGO performance? We address this question using instrumental variables to estimate the local average treatment effect of employing a contract teacher. IV estimates allow us to test whether NGO-government differences are attributable to differences in recruitment success, under the maintained assumption that random assignment to treatment status influences test scores only through teacher presence.

Define T as the proportion of months during the experiment that a school employed a contract teacher. T is clearly endogenous to unobserved factors such as the quality of school management which may also directly affect pupil performance. Random assignment satisfies the exclusion restriction for a valid instrument for contract teacher presence under the assumption stated above, allowing us to estimate local average treatment effects for schools hiring a teacher in both the government and NGO program.

As a benchmark, we present a naïve OLS regression of test scores on treatment status, where T_{jt} measures the proportion of months out of a possible 17 months total duration of the program that a contract teacher was in place in a given school.

$$Y_{ijt} = \alpha_3 + \beta_3 T_{jt} + \gamma_3 \mathbf{X}_{jt} + \varepsilon_{3ijt} \tag{3}$$

Columns 4 to 6 in the top panel of Table 4 report the estimates of equation (3). As seen, the effect is slightly larger than the ITT effect at roughly 0.13 standard deviations, but is insignificant across all three specifications. The treatment variable ranges from zero to one, where one implies a school employed a teacher for all 17 months of the program. Thus the point estimates can be interpreted

as the comparison of a school with no teacher to one with a full 17-months' exposure to treatment. Columns 4 to 6 in the bottom panel of Table 4 report the results from the naïve OLS estimates comparing the effect of NGO and government treatment on test scores.

$$Y_{ijt} = \alpha_4 + \beta_4 T_{jt} + \beta_4' T_{jt} \times \text{Gov}_{jt} + \gamma_4 \mathbf{X}_{jt} + \varepsilon_{4ijt} \tag{4}$$

The point estimates on T are statistically significant for both the lagged dependent variable and fixed effects models, with point estimates of 0.22 and 0.24, respectively. As in the ITT regressions, however, the coefficients on the interaction of treatment and government implementation ($T\times$ Gov) are statistically significant and almost perfectly negate the overall treatment effect, implying zero effect in schools where the program was administered by the government.

Because of the obvious potential bias affecting OLS estimates of β_3, we use the random assignment, Z, to instrument actual treatment, T. Thus we estimate the local average treatment effect as

$$Y_{ijt} = \alpha_5 + \beta_5 \hat{T}_{jt} + \gamma_5 \mathbf{X}_{jt} + \varepsilon_{5ijt} \tag{5}$$

where $\hat{T}_{jt}$ are the predicted values from the first-stage regression

$$T_{jt} = \alpha_6 + \delta_6 Z_{jt} + \gamma_6 \mathbf{X}_{jt} + \varepsilon_{6ijt}. \tag{6}$$

To distinguish the impact in the government and NGO treatment arm, we estimate

$$Y_{ijt} = \alpha_7 + \beta_7 \hat{T}_{jt} + \beta_7' \widehat{T_{jt} \times \text{Gov}}_{jt} + \gamma_7 \mathbf{X}_{jt} + \varepsilon_{7ijt} \tag{7}$$

where $\hat{T}_{jt}$ and $\widehat{T_{jt} \times \text{Gov}}_{jt}$ are the predicted values from the following first-stage regressions

$$T_{jt} = \alpha_8 + \delta_8 Z_{jt} + \delta_8' Z_{jt} \times \text{Gov}_{jt} + \gamma_8 \mathbf{X}_{jt} + \varepsilon_{8ijt}. \tag{8}$$

$$T_{jt} \times \text{Gov}_{jt} = \alpha_9 + \delta_9 Z_{jt} + \delta_9' Z_{jt} \times \text{Gov}_{jt} + \gamma_9 \mathbf{X}_{jt} + \varepsilon_{9ijt}. \tag{9}$$

Results from estimating equations (5) and (7) are presented in Columns 7-9 of Table 4, with the results of interest found in the bottom panel. Instrumentation has a small and statistically insignificant effect on the treatment coefficients in Columns 7-9 vis-a-vis the OLS estimates in Columns 4-6. The overall LATE estimate ranges from 0.20 in the cross-section to 0.24 in the lagged dependent variable model. Once again, in both the lagged dependent variable and fixed effects models, the interaction of treatment and government implementation has a negative effect (and is significant at the 5% level in the fixed effects specification), with the combined point estimate insignificantly different from zero in the government treatment arm.

5.3 Contract variations and training

Duflo et al. (2012a) show that training school management committees in their governance responsibilities is an effective complement to the contract teacher intervention. We replicate a similar SMC training intervention in half of the schools in both the NGO and government treatment arm. (Note that SMC training was not conducted if the school was not assigned a contract teacher, as the training curriculum focused heavily on the SMC's responsibilities in recruiting, paying, and monitoring the performance of the teachers.) Table 5 shows the ITT effect of the SMC training on test scores, extending the specification in equation (1).

As seen in the top panel, columns 1 to 3, the coefficient on the interaction of Z and the indicator variable for SMC training is positive but statistically insignificant in all three specifications. The bottom panel shows the results separately for the NGO and government treatment arms. Again, the SMC training has no significant effect in any specification.[13]

In addition to the SMC training intervention, Table 5 also tests whether varying two dimensions of the teachers' contract had any effect on test score performance:(i) receiving a higher salary offer; or (ii) being recruited and paid locally by the SMC rather than by district or national officials. In section 4.1 we showed that these variations had significant positive effects on schools' ability to recruit and retain contract teachers. However, Table 5 shows no such effect on test score performance. There is no significant difference in test performance between either contract variant, in the overall sample or in either the NGO or government sub-sample.

Overall, results indicate that the institution providing the contract teacher is the key determinant of the program's impact, whereas variants in contract details or complementary training have no marginal effect.

5.4 Robustness checks

We explore a number of alternatives for the ITT specification in equation (2) to examine the robustness of the core results. First, we collapse all test scores at

[13]Note that the experimental design has insufficient power to detect effects of interesting magnitudes when examining these cross-cutting interventions separately in the NGO and government treatment arm. We report them only for the sake of transparency, given the disparity between the overall ITT effect of the contract teacher program in the NGO and government samples.

the school (and year) level and estimate

$$Y_{jt} = \alpha_{10} + \beta_{10} Z_{jt} + \beta'_{10} Z_{jt} \times \text{Gov}_{jt} + \gamma_{10} \mathbf{X}_{jt} + \varepsilon_{10jt} \tag{10}$$

with school fixed effects. This specification differs from estimates based on equation (2) in two respects. It is more conservative in terms of the standard errors than the pupil-level regression using clustered standard errors (Angrist and Pischke 2009). Point estimates from equations (2) and (10) also differ due to an implicit re-weighting of the observations. Pupil sample sizes vary across schools in relationship with school enrollment up to a maximum of twenty pupils per class. Below this ceiling of twenty pupils, schools with more pupils receive more weight, and the estimates using pupil-level data can be interpreted roughly as the effect on an average pupil. Estimates using the collapsed data represent, instead, the the treatment effect in the average school in the sample.

The results are presented in the first column of Table 6. The conclusions are unchanged, being assigned a contract teacher increases test scores by 0.18 standard deviations in the NGO treatment arm, which is significant at the 5% level, but has no measurable effect in the government treatment arm.

Second, we explore whether school attrition at follow up biases our results. To do so, we estimate bounds on the coefficients in (10) using the procedure proposed by Lee (2009).[14] The results are presented in column (2) of Table 6, and again, the conclusions remain unchanged.

Finally, we examine whether pupil attrition (or increases) drive the results by including the percentage change in enrollment in the treatment cohort and treatment grade as explanatory variables. The results are reported in column (4) and (5). Including the enrollment variables, which are themselves significant and positively related to test scores, does not change the results. The effect of an additional contract teacher in the government treatment arm is still zero and insignificant, while the effect in the NGO treatment arm is 0.17, though the latter is just shy of significant when controlling for percentage changes in cohort size.

In sum, the core results are therefore robust to a number of different (and more conservative specifications).

[14] The upper Lee bound is estimated by equating the proportion of treated and non-treated schools by trimming the test score distribution in treated schools from the top – equivalent to a worst case scenario.

6 Heterogeneous treatment response

In addition to the institutional considerations raised above, a more traditional concern about the generalizability of RCT results stems from possible heterogeneous response to treatment associated with differences in school or pupil characteristics. The broad geographic dispersion of our sample is helpful in both addressing and testing the basis for this concern.

The estimates in Table 4 provide an unbiased estimate of the intention-to-treat effect for schools within the sampling frame – i.e., schools with high pupil-teacher ratios in the 14 study districts. In general, if the treatment effect varies with school or pupil characteristics, and the sampling frame differs from the population of interest for policymaking, results from any evaluation (experimental or otherwise) will not be broadly applicable. Estimation of heterogeneous treatment effects, combined with knowledge of the distribution of exogenous characteristics in the sample and population, may provide a bridge from internal to external validity.

Two issues to be addressed in estimating heterogeneous effects are (i) selecting the dimensions of heterogeneity, and (ii) hypothesis testing with multiple comparisons (Green and Kern 2012). On the former question, the literature on medical trials commonly takes a data-driven approach based on boosting algorithms (Friedman, Hastie and Tibshirani 2000). An alternative approach to studying heterogeneity, more common in the social sciences and which we use here, is hypothesis driven. Specific interaction terms, $\mathbf{X}_{jt}$, are proposed based on *ex ante* hypotheses and tested in an extension of equation (1) including school fixed effects.

$$Y_{ijt} = \alpha_{11} + \beta_{11} Z_{jt} + \beta_{11}^{'} \left(Z_{jt} \times \frac{\mathbf{X}_{jt} - \mu_x}{\sigma_x} \right) + \gamma_{11j} + \varepsilon_{11ijt} \qquad (11)$$

We explore three hypotheses. The first is that the intervention's effect will be stronger where the supply of teachers is higher, reducing the risk of unfilled vacancies and potentially increasing contract teachers' motivation to maintain employment. As a rough proxy for the supply of teachers in a given area, we use the count of other primary schools within a 5-mile radius of the school. We assume that a higher density of primary schools implies a higher population density, particularly for skilled labor, and a thicker labor market for teachers.

Our second hypothesis about heterogeneity is that the addition of a contract teacher will have a larger effect in schools with a higher initial pupil-teacher ratio, as these schools will experience a larger reduction in class size due to

treatment. Finally, our third hypothesis is that the treatment will be more effective in schools with lower initial test scores. This hypothesis is more speculative, but is motivated by the attention paid to tracking and remedial education in the contract teacher literature (Banerjee et al. 2007, Duflo et al. 2012a).

Our sample overlaps with the study area of Duflo et al. (2012a) in one district of Western Kenya, Teso.[15] Figure 3 shows kernel densities of the three baseline characteristics associated with our three hypotheses about heterogeneous treatment effects – pupil-teacher ratios, geographic density of schools, and baseline test scores – for our entire sample (blue lines), and exclusively for the Western Kenya schools in our sample (red lines). As seen, our Western Kenya sub-sample has somewhat higher pupil-teacher ratios at baseline (mean of 69.2 compared to 60.9 for the rest of the sample), and somewhat lower baseline test scores (mean of -0.25 on the standardized test compared to 0.11 for the remaining provinces). Mean geographic density is similar in Western Kenya (3.64) as compared to the full sample (3.27), but the variance is much lower (standard deviation of 0.25 compared to 1.69).

Table 7 shows the results from estimating the heterogeneous ITT effects in equation (11). Because the variables measuring exogenous heterogeneity have been standardized, all coefficients can be interpreted as the change in the treatment effect implied by a one standard-deviation change in the independent variable. For instance, column 1 shows that the ITT effect is roughly 0.7 percentage points larger in locations with a higher density of schools, in line with the sign of our hypothesis but close to zero and entirely insignificant. Column 2 shows no consistent relationship between initial pupil-teacher ratios and the treatment effect. Turning to our third hypothesis, we explore two measures of schools' initial level of academic achievement: scores on an independent national standardized test administered to grade 8 pupils in 2005, and scores on the baseline test used in the primary analysis here. Column 3 shows a significantly negative relationship between initial test scores in the baseline and subsequent treatment effects (coefficient of -.115), implying that the intervention is somewhat progressive.

There is evidence that this last result is less than robust. First, examining columns 7 and 11, the heterogeneity in effects with respect to baseline scores is unique to the government treatment arm. No such effect is observed in the NGO sample. Second, columns 4, 8, and 12 interact the treatment assignment with an alternative measure of baseline academic performance: lagged scores on a national standardized test, the KCPE exam. The interaction terms are

[15] Teso comprises one of two districts in Duflo et al.'s (2012a) sample, and is the only Western Kenya district in our sample.

insignificant and of inconsistent signs.

Nevertheless, taking these coefficients at face value, the results imply slightly larger treatment effects in Western Kenya where baseline test scores were roughly one-quarter standard deviation below the national mean. Specifically, the estimates in column (3) of Table 7 imply that the overall ITT effect would be about 0.028 (-0.115×-0.25) standard deviations higher relative to the overall effect of 0.078. The estimates in column (7) imply that the gap between the Ministry and NGO treatment effects would be approximately 0.058 (-0.161×-0.25) standard deviations smaller, relative to an overall gap of nearly one-fifth of a standard deviation.

What do these findings imply for the external validity of evidence from a single province in Kenya, in particular Western Kenya? There is no evidence that NGO treatment effects in Western Kenya should be expected to differ whatsoever from NGO treatment effects estimated in the other districts in our sample. There is limited, and non-robust evidence that the intervention is progressive in the government treatment arm, with a larger effect for schools with lower baseline performance. Overall, we believe the limited heterogeneity of the treatment effects estimated here should lend some confidence to policymakers wishing to forecast the effect of small-scale interventions across Kenyan schools based on results from Western Kenya alone. There is little reason to question the external validity of Duflo et al. (2012a), at least within Kenya, on the basis of heterogeneous response across schools. But there are good reasons to believe that results based on NGO-led interventions do not extend to government implementation.

7 Mechanisms

We now turn to examining mechanisms which could explain the difference in performance between contract teachers in the NGO and government treatment arms. We explore two dimensions that we argue are a function of working with government, independent of the scale of the program: characteristics and effort of teachers hired, and weak monitoring and accountability systems. We also explore a third channel 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 (see Table 8, columns 1-3).

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 (Table 8, column 4).[16] Third, we add interaction terms to the main treatment effects specification from equation (2) to examine the plausibility that the national controversy surrounding the hiring of 18,000 contract teachers disproportionately affected teachers in the government treatment arm (Table 9), and thus helps to explain the differential effect on test scores.[17]

7.1 Teacher selection

In theory, the protocol for teacher recruitment was the same for the government and the NGO schools in our sample. In practice, the NGO may have put more effort into recruiting high quality candidates. We test this hypothesis by comparing the observable characteristics of contract teachers hired in each treatment arm.

Inasmuch as there are any significant differences, the Ministry hired teachers with more observable skills. As can be seen from Table 8, Panel A, teachers in the NGO treatment arm have less tertiary education. There is no significant difference in terms of teaching qualifications and age between government and NGO. Teachers in the government arm are more likely to be male. Interestingly, none of these observable skills or demographic characteristics are significantly correlated with changes in test scores.

Another way in which recruitment could be compromised is through local capture of the hiring process. Instead of hiring the best qualified candidate, existing teachers lobby for employing their friends and relatives. Duflo et al. (2012a) note that local capture of this kind significantly reduced the positive impact of contract teachers on learning, but that schools which received a complementary

[16] Column 4 of Table 8 reports the coefficient in a regression of changes in test scores between 2009-2011 on each of the intermediate outcomes and a constant. As discussed by Imai, Keele, Tingley and Yamamoto (2011) and Green, Ha and Bullock (2010), there is no widely accepted empirical technique for establishing the role of intermediate outcome variables as part of a causal chain. Correlations between final and intermediate outcomes are at best suggestive of causal channels.

[17] The data in Table 8 and 9 is based on exit interviews with contract teachers conducted after the follow-up survey. We were able to track 111 contract teachers drawn from 84 of the 108 schools that employed a teacher. Comparison of the full sample to the sample of contract teachers shows that attrition was not systematically related to treatment effects, results in end-of-primary leaving exams or pupil-teacher ratios. Results available upon request.

training intervention for the School Management Committee were less prone to local capture.

Consistent with Duflo et al. (2012a), we find that the hiring process was compromised by local capture under government implementation. 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. While this difference is suggestive of a corrupted hiring process in the government program, it is worth noting that our indicator of local capture does not show the significant, negative correlation with test score improvements that one might expect.

7.2 Monitoring and accountability

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 7, Panel B).

Teacher presence in the classroom is indeed higher in schools managed by the NGO (73% versus 63%), but the difference is not significant between treatment arms. Presence in the class room is positively, but not significantly, correlated with test scores.

There is a significant difference between the monitoring activities of the NGO and the government. Schools in the NGO treatment arm were 15% more likely to have received a monitoring visit than schools in the government treatment arm. However, the likelihood of receiving a monitoring visit is not a significant correlate of changes in test scores.

Similar differences are observed in the management of the payroll system and prompt payment of salaries. Both in the government treatment arm and in the NGO treatment arm, salary delays occurred, but they were significantly more severe under government implementation – with an average delay of roughly three months in the government arm, compared to 2 months in the NGO arm.

The salary delays display a significant negative correlation with test score improvements. Taking the point estimates in Table 8 at face value, an increase in salary delays of 1 months (roughly the average difference between NGO and government) accounts for one third of the difference in test scores between NGO and government.

We interpret these findings on teacher presence, monitoring and salary delays as different dimensions of a common problem: low top-down accountability in the government bureaucracy, especially in the link from Nairobi to the district offices. Salary delays were often related to the inability of government officials in Nairobi to confirm the identity or payment details of teachers contracted locally, preventing timely completion of bank transfers. In either case, district-level employees of the Ministry failed to carry out their duties under the program: conducting monitoring visits and reporting back information to Nairobi. Although the SMC training was designed to compensate for this low top-down accountability, the results in Section 5.3 show that it failed to have the intended effect. In contrast, district-level employees of the NGO appear to be more accountable and responsive to their superiors in Nairobi.

7.3 Unionization, expectations and credibility of short-term contracts

The effect of a fixed-term contract on teacher performance is likely mediated by teachers' beliefs about the credibility of that contract. Will the contract be terminated if their performance is poor, or is this an empty threat? We hypothesize that teachers' expectations will differ when offered identical contracts by an international NGO or a national government.

This hypothesis is grounded in the highly unionized and politicized nature of public sector teaching in Kenya, as in many developing countries. In this case, the government's ambitious plan to employ 18,000 contract teachers nationwide posed a significant threat to the Kenyan National Union of Teachers. The teachers' union waged an intense political and legal battle against the contract teacher program, including a lawsuit which delayed implementation by over a year, street protests in central Nairobi, and a two-day national strike, demanding permanent civil service employment and union wage levels for all contract teachers. By June 2011, 4 months before the impact evaluation ended, the government acquiesced to union demands to absorb the contract teachers into civil service employment at the end of their contracts.

Formally, teachers employed in our research project were not covered by the

negotiations between the government and the teacher union.[18] Nevertheless, we hypothesize that teachers in the government treatment arm were more likely to perceive the outcome of the union negotiation 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 explore this hypothesis in Panel C, Table 8. Two thirds of teachers overall expressed the hope that the experimental contract teacher program would be a stepping stone to permanent employment, with no significant difference between government and NGO. We do, however, see large and significant differences when we ask whether teachers felt that the union was supporting them in this desire. Only 14% 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.[19] Interestingly, this large difference in self-identification with the union is not reflected in any difference in active involvement, such as participating in the national strike.

When relating these variables to changes in test scores, 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 the estimates in column 4 of Table 8 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 employed by the government in the experiment and negatively affected their performance, while teachers employed by the NGO were largely immune to the political struggle between the government and the teachers union.

Table 9 presents further evidence consistent with this interpretation. In particular, we hypothesize that union representatives and contract teachers employed by the government in the national scale-up would signal to experimental teachers

[18] As shown in Table 8, Panel D, roughly 45% of the program teachers are now employed on permanent and pensionable contracts, with no significant difference between teachers previously in the government and NGO treatment arms. In contrast, all of the 18,000 contract teachers hired by the government outside of the experimental evaluation received permanent tenure.

[19] 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?"

in the government treatment arm that the employment guarantee agreed upon by the government and the union would also extend to them. This in turn would lead experimental teachers in the government arm to believe that the union was representing their interests throughout the program. In contrast, experimental teachers in the NGO arm – just like existing PTA teachers – would be made to understand that they would not be covered by the employment guarantee. If this hypothesis is correct, then we would expect contact with the unions or one of the 18,000 contract teachers to strengthen identification with the union for teachers in the government treatment arm, but not for teachers in the NGO treatment arm.

We examine this hypothesis in column (1) and (2) of Table 9. Contact with the union increases the likelihood of identifying with the union (that is, stating that the union represented one's interests) by 50% for teachers in the government treatment arm (a significant effect), but only by a mere 8% for teachers in the NGO treatment arm (an insignificant effect). The difference between the two coefficients is significant at the 5% level.[20] Similarly, placing one (or more) of the 18,000 contract teachers in a school where the experimental teacher is managed by the government increases his or her probability of identifying with the union by 12% (though this coefficient is not significant), while the effect is exactly zero in a school where the experimental teacher is managed by the NGO.

Second, we hypothesize that for experimental teachers in the government treatment arm, greater exposure to the controversy surrounding the 18,000 government contract teachers (and the union's demands that they be permanently employed) undermines the credibility of the dynamic incentives provided by the short-term contracts in the experiment. Where teachers find the threat of contract termination less credible, we would expect them to exert less effort and hence have lower test score gains. Taken together, this implies a negative association between exposure to the 18,000 government contract teachers and union lobbying and changes in test scores for teachers in the government treatment arm, but not for teachers in the NGO treatment arm.

We examine this hypothesis in column (3) and (4) of Table 9. In the government treatment arm, either having contact with the union or placing one of the 18,000 government contract teachers in the school significantly reduces test-score gains by 0.3 and 0.25 of a standard deviation respectively. In the NGO treatment arm, exposure to the national controversy had no effect on test score gains.

Taken at face value, the results in column (3) and (4) of Table 9 imply that

[20] Contact with the union is defined as the average of having been visited by a union representative and having attended a union meeting.

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 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. In this final sub-section (7.3), 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. However, this evidence is only suggestive. Formally, the limitations of the Kenyan government (at any scale) and the weaknesses exposed by scaling up are co-linear in our experimental design, and both are contained within the coefficient on the interaction of treatment status and government implementation.

8 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 the external validity of experimentally-estimated treatment effects of social programs.

We report on a randomized trial replicating earlier experimental results showing that contract teachers significantly raise pupil test scores when implemented by an international NGO. These effects disappear entirely when the program is (a) implemented within the bureaucratic structures of the Kenyan government and (b) extended to a national scale. We show that this latter point matters less in terms of the heterogeneity of the beneficiary population, and more in terms of the concomitant political response from vested interests opposed to the program.

Our results suggest that scaling-up an intervention (typically defined at the school, clinic, or village level) found to work in a randomized trial run by a specific organization (often an NGO chosen for its organizational efficiency) requires

an understanding of the whole delivery chain. If this delivery chain involves a government Ministry with limited implementation capacity or which is subject to considerable political pressures, agents may respond differently than they would to an NGO-led experiment. Lack of attention to interaction between the intervention being tested and the broader institutional context, and adjustment of policy accordingly, may imply very different effects than those implied by a simple extrapolation of the estimates of the controlled experiment (cf Reinikka and Svensson 2005).

How externally valid are our findings on the limits of external validity? We have focused on particular mechanisms undermining external validity – government capacity and political economy responses – whose relevance depends on the context and intervention in question. For instance, even within the field of experimental economics research in developing-country settings, we would argue that our results have limited relevance to studies estimating effects driven by biological processes (e.g. Miguel and Kremer 2004) or a production function relationship (e.g. de Mel, McKenzie and Woodruff 2008). But we argue our results are highly relevant to studies, randomized or other, estimating the impact of reforms to the delivery of health, education, sanitation, policing, or other government services, and especially so in developing countries with weak public sector institutions.

In the terminology of Shadish, Campbell and Cook's (2002) classic text on generalizing experimental results, this is a question of 'construct validity' rather than external validity per se, i.e., of identifying the higher order construct represented by the experimental treatment. In most of the experimental evaluation literature in development economics, the treatment construct is defined to include only the school- or clinic-level intervention, abstracting from the institutional context of these interventions. Our findings suggest that the treatment in this case was not a "contract teacher", but rather a multi-layered organizational structure including monitoring systems, payroll departments, long-run career incentives and political pressures.

Our results are also potentially relevant to debates on the generalizability of RCT results beyond economics. While the education literature has focused on measuring and controlling for the "fidelity" of implementation to explain replication failures (Borman, Hewes, Overman and Brown 2003), our results point to the underlying institutional obstacles to fidelity that must be considered in any attempt to translate experimental findings into government policy. In the literature on clinical trials in health, a distinction is frequently made between efficacy studies conducted in a more controlled setting, and effectiveness studies that more closely mimic "real world" conditions. Both treatment arms in the

present study meet standard criteria for an effectiveness study: i.e., representative sampling, use of intention-to-treat analysis, clinically relevant treatment modalities (Gartlehner, Hansen, Nissman, Lohr and Carey 2006). Yet the gap in performance between the NGO and government program suggests an important difference between many effectiveness studies, as commonly defined, and the real-world institutional setting of public service delivery in the developing world.

References

Acemoglu, Daron, "Theory, general equilibrium, and political economy in development economics," *Journal of Economic Perspectives*, 2010, *24* (3), 17–32.

Allcott, Hunt and Sendhil Mullainathan, "External Validity and Partner Selection Bias," *NBER Working Paper*, 2012, (18373).

Angrist, Joshua and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Atherton, Paul and Geeta Kingdon, "The relative effectiveness and costs of contract teachers in India," *CSAE Working Paper Series*, 2010, *2010-16.*

Baird, Sarah, Craig McIntosh, and Berk Özler, "Cash or Condition? Evidence from a Cash Transfer Experiment," *Quarterly Journal of Economics*, 2011, *126* (4), 1709–1753.

Banerjee, Abhijit and Ruimin He, *Reinventing Foreign Aid*, MIT Press, Cambridge, Massachusetts,

___, Rukmini Banerji, Esther Duflo, Rachel Glennerster, , and Stuti Khemani, "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," *American Economic Journal: Economic Policy*, 2010, *2* (1), 1–30.

Banerjee, Abhijit V, Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying Education: Evidence From Two Randomized Experiments in India," *Quarterly Journal of Economics*, 2007, *122* (3).

Barr, Abigail, Frederick Mugisha, Pieter Serneels, and Andrew Zeitlin, "Information and collective action in the community monitoring of schools:

Field and lab experimental evidence from Uganda," mimeo, Centre for the Study of African Economies, Oxford 2011.

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

Barrera-Osorio, Felipe and Leigh Linden, "The Use and Misuse of Computers in Education: Evidence from a Randomized Controlled Trial of a Language Arts Program," 2009.

Berry, James, "Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India," 2011.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, and Justin Sandefur, "Why Did Abolishing Fees Not Increase Public School Enrollment in Kenya?," *Center for Global Development Working Paper Series*, 2011, *271.*

Borman, G.D., G.M. Hewes, L.T. Overman, and S. Brown, "Comprehensive School Reform and Achievement: A Meta-Analysis," *Review of Educational Research*, 2003, *73*, 125–230.

Bourdon, J., M. Frölich, and K Michaelowa, "Teacher Shortages, Teacher Contracts and Their Impact on Education in Africa," *IZA Discussion Paper, Institute for the Study of Labor, Bonn, Germany.*, 2007, (2844).

Bruhn, Miriam and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, October 2009, *1* (4), 200–232.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos, *Making Schools Work: New Evidence on Accountability Reforms*, Washington, DC: The International Bank for Reconstruction and Development / The World Bank, 2011.

Burde, Dana and Leigh Linden, "The Effect of Village-Based Schools: Evidence from a Randomized Controlled Trial in Afghanistan," 2012.

Cartwright, Nancy and Jeremy Hardie, *Evidence-Based Policy: A Practical Guide to Doing it Better*, Oxford University Press, 2012.

Compton, Mary and Lois Weiner, "More police attacks on Kashmiri teachers," October 2012a. Published online, Oct. 08, 2012, at http://www.teachersolidarity.com/blog/more-police-attacks-on-kashmiri-teachers/.

___ **and** ___ , "Striking Indian Contract Teachers won't be intimidated," October 2012b. Published online, Oct. 31, 2012, at http://www.teachersolidarity.com/blog/striking-indian-contract-teachers-wont-be-intimidated/.

de Mel, Suresh, David McKenzie, and Christopher Woodruff, "Returns to Capital in Microenterprises: Evidence from a Field Experiment," *Quarterly Journal of Economics*, 2008, *123.*

Deaton, Angus, "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 2010, *48* (2), 424–455.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 2011, *101* (5).

___ **,** ___ **, and** ___ , "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools," *NBER Working Paper*, 2012, (17939).

___ **, Rema Hanna, and Stephen Ryan**, "Incentives Work: Getting Teachers to Come to School," *American Economic Review*, 2012, *102* (4), 1241–1278.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, 2000, *28* (2), 337–407.

Gartlehner, Gerald, Richard A. Hansen, D. Nissman, K. Lohr, and T. Carey, "Criteria for Distinguishing Effectiveness From Efficacy Trials in Systematic Reviews," Technical Report 06-0046, AHRQ Technical Review 12 (Prepared by the RTI International – University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016.) 2006.

Glewwe, Paul, Albert Park, and Meng Zhao, "A Better Vision for Development: Eyeglasses and Academic Performance in Rural Primary Schools in China," *University of Minnesota, Center for International Food and Agricultural Policy Working Papers*, 2012, *120032.*

___ **and Eugenie Maïga**, "The Impacts of School Management Reforms in Madagascar: Do the Impacts Vary by Teacher Type?," *The Journal of Development Effectiveness*, 2011, *3* (4), 4353–469.

___ , **Michael Kremer, and Sylvie Moulin**, "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, 2009, *1* (1), 112–135.

___ , ___ , ___ , **and Eric Zitzewitz**, "Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya," *Journal of Development Economics*, 2004, *74*, 251–268.

___ , **Nauman Ilias, and Michael Kremer**, "Teacher Incentives," *American Economic Journal: Applied Economics*, 2010, *2*, 205–227.

Goyal, S. and P. Pandley, "Contract Teachers," Technical Report 28, World Bank South Asia Human Development Sector Report, Washington, DC 2009.

Green, Donald P. and Holger L. Kern, "Modeling Heterogenous Treatment Effects in Survey Experiments using Bayesian Additive Regression Trees," *Public Opinion Quarterly*, 2012, *76* (3), 491–511.

___ , **Shang E. Ha, and John G. Bullock**, "Enough Already about "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose," *Annals of the American Academic of Political and Social Science*, 2010, *628*, 200–208.

Greevy, Robert, Bo Lu, Jeffrey Silber, and Paul Rosenbaum, "Optimal multivariate matching before randomization," *Biometrika*, 2004, *5* (2), 263–275.

He, Fang, Leigh Linden, and Margaret MacLeod, "How to Teach English in India: Testing the Relative Productivity of Instruction Methods within the Pratham English Language Education Program," 2008.

___ , ___ , **and** ___ , "A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trial," 2009.

Heckman, James J., "Randomization and Social Policy Evaluation," *NBER Working Paper*, July 1991, (107).

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto, "Unpacking the Black Box: Learning about Causal Mechanisms from Experimental and Observational Studies," *American Political Science Review*, 2011, *105* (4), 765–789.

Inamdar, Parimala, "Computer skills development by children using 'hole in the wall' facilities in rural India," *Australasian Journal of Educational Technology*, 2004, *20* (3), 337–350.

Kazianga, Harounan, Damien de Walque, and Harold Alderman, "Educational and Health Impacts of Two School Feeding Schemes: Evidence from a Randomized Trial in Rural Burkina Faso," *World Bank Policy Research Working Paper*, 2009, *4976*.

Kremer, Michael, Edward Miguel, and Rebecca Thornton, "Incentives to Learn," *The Review of Economics and Statistics*, 2009, *92* (3), 437–456.

___, Sylvie Moulin, and Robert Namunyu, "Decentralization: A Cautionary Tale," *Poverty Action Lab Paper No. 10*, 2003.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 2009, *76(2)*, 1071–1102.

Linden, Leigh, "Complement or Substitute? The Effect of Technology on Student Achievement in India," 2008.

___, Abhijit V Banerjee, and Esther Duflo, "Computer-Assisted Learning: Evidence from a Randomized Experiment," *Poverty Action Lab Paper No. 5*, 2003.

Macours, Karen, Norbert Schady, and Renos Vakis, "Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood: Evidence from a Randomized Experiment," *Human Capital and Economic Opportunity: A Global Working Group*, 2011, *2011-007*.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 2004, *72*.

Mugo, John, Amos Kaburu, Charity Limboro, and Albert Kimutai, "Are Our Children Learning: Annual Learning Assessment Report," Technical Report, Uwezo Kenya 2011.

Muralidharan, Karthik and Venkatesh Sundararaman, "Contract Teachers: Experimental Evidence from India," 2010.

___ and ___, "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, 2011, *119* (1), 39–77.

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

Otieno, Wycliffe and Christopher Colclough, "Financing Education in Kenya : Expenditure , Outcomes and the Role of International Aid by," *Research Consortium on Educational Outcomes & Poverty Working Paper*, 2009, (25).

Pandey, Priyanka, Sangeeta Goyal, and Venkatesh Sundararaman, "Community Participation in Public Schools: The Impact of Information Campaigns in Three Indian States," *World Bank Policy Research Working Paper*, 2008, *4776.*

Paxson, Christina and Norbert Schady, "Does Money Matter? The Effects of Cash Transfers on Child Health and Development in Rural Ecuador," *World Bank Policy Research Working Paper*, 2007, *4226.*

Reinikka, Ritva and Jakob Svensson, "Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda," *Journal of the European Economics Association*, 2005, *3* (2/3), 259–267.

Rosas, Ricardo, Miguel Nussbaum, Patricio Cumsille, Vladimir Marianov, Monica Correa, Patricia Flores, Valeska Grau, Francisca Lagos, Ximena Lopez, Veronica Lopez, Patricio Rodriguez, and Marcela Salinas, "Beyond Nintendo: design and assessment of educational video games for first and second grade students," *Computers and Education*, 2003, *40*, 71–94.

Shadish, William R., Thomas D. Campbell, and Donald T. Cook, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, 2002.

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

Vermeersch, Christel and Michael Kremer, "School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation," *World Bank Policy Research Working Paper*, 2005, *3523.*

Table 1: Balance at baseline

	Control	Treatment	Gov.	NGO	Diff. (2)−(1)	Diff. (3)−(1)	Diff. (4)−(1)	Diff. (3)−(4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pupil-Teacher Ratio	43.33	54.11	55.74	52.15	10.77	12.41	8.82	3.59
					(7.39)	(8.82)	(9.76)	(11.09)
No. of classrooms	11.44	12.31	12.49	12.11	.87	1.05	.67	.38
					(.99)	(1.20)	(1.15)	(1.26)
No. Civil Service Teachers	7.31	8.25	8.40	8.11	.95	1.10	.80	.30
					(.99)	(1.15)	(1.17)	(1.22)
No. PTA Teachers	1.40	1.85	2.00	1.70	.45	.60	.30	.30
					(.28)	(.36)*	(.32)	(.37)
Avg. Pay PTA Teachers	3237.38	3446.31	3438.05	3454.57	208.93	200.67	217.19	-16.53
					(437.29)	(573.37)	(477.63)	(587.28)
Primary leaving exam score	235.93	233.71	232.75	234.67	-2.22	-3.18	-1.26	-1.92
					(6.06)	(6.79)	(7.09)	(6.74)
Pupil standardized test score - English, grade 1	-.07	.04	.03	.04	.11	.10	.11	-.01
					(.13)	(.15)	(.15)	(.16)
Pupil standardized test score - Math, grade 1	-.06	.03	.04	.02	.09	.10	.08	.02
					(.10)	(.12)	(.12)	(.13)
Pupil standardized test score - English & Math	2.11e-09	.09	.14	.04	.09	.14	.04	.11
					(.09)	(.10)	(.10)	(.10)

School level statistics are based on 176 schools with usable baseline information, pupil level statistics are based on 6,276 pupils from these schools. Standard errors for pupil level information are clustered at the school level. Here and in all subsequent tables, standard errors are reported in brackets and asterisks denote significance at the 1% (***), 5% (**) and 10% (*) level.

Table 2: Labor supply of contract teachers

	Linear Probability Model			Logit Model		
	(1)	(2)	(3)	(4)	(5)	(6)
NGO implementation	.120	.120	.144	.120	.121	.145
	(.066)*	(.065)*	(.062)**	(.066)*	(.065)*	(.064)**
High salary		.118	.092		.117	.090
		(.064)*	(.066)		(.063)*	(.065)
Local recruitment		.140	.154		.141	.149
		(.065)**	(.063)**		(.065)**	(.065)**
Geographic density						-.0003
						(.0002)
Lagged KCPE exam score			.001			.002
			(.001)			(.001)
Pupil-teacher ratio			.003			.003
			(.002)			(.003)
Obs.	2,060	2,060	2,044	2,060	2,060	2,044

The unit of observation is the school, with monthly observations from June 2010 to October 2011. The dependent variable is a binary indicator of whether a teacher was present and teaching in a given school in a given month. Columns 1, 2, 3 and 4 restrict the determinants of teacher presence to factors controlled by the experiment, while columns 2 and 4 include other exogenous and/or predetermined school characteristics. For the logit model, the table reports marginal effects and their standard errors. All standard errors are clustered at the school level.

Table 3: Compliance and contamination

	Treatment	Control	Diff.	Gov.	NGO	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Teacher Recruitment						
Ever employed a teacher	.87			.88	.86	.02
						(.06)
No. of months employed a teacher (out of 17)	12.30			11.59	13.00	-1.41
						(1.08)
Panel B: Reallocation within school						
Class size	60.67	69.45	-8.78	60.47	60.88	-.41
			(6.14)			(6.42)
Teacher always in correct grade	.72			.76	.69	.07
						(.09)
Teacher ever in correct grade	.95			.97	.94	.02
						(.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	-.11	-.09	-.01	-.14	-.08	-.06
			(.04)			(.05)
% change in grade enrollment	-.02	-.02	.0004	-.04	.008	-.05
			(.04)			(.06)
No. of teachers from 18,000 program	.65	.48	.17	.68	.62	.06
			(.17)			(.21)
No. of TSC teachers	9.96	10.10	-.14	10.15	9.75	.41
			(1.11)			(1.32)
No. of PTA teachers	2.06	1.74	.32	2.03	2.09	-.06
			(.35)			(.43)

Table 4: Treatment effects

	ITT			OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pooling treatment arms:									
Z	.138	.103	.078						
	(.090)	(.076)	(.080)						
Mos. of Contract Teacher				.135	.126	.125	.190	.141	.105
				(.097)	(.080)	(.082)	(.124)	(.103)	(.108)
NGO vs gov't implementation:									
Z	.153	.184	.175						
	(.109)	(.088)**	(.091)*						
$Z\times$ Gov	-.029	-.163	-.197						
	(.119)	(.095)*	(.094)**						
Mos. of Contract Teacher				.154	.217	.234	.199	.238	.225
				(.122)	(.093)**	(.091)**	(.142)	(.113)**	(.116)*
Mos. of Contract Teacher x Gov				-.043	-.200	-.239	-.019	-.208	-.256
				(.146)	(.121)*	(.118)**	(.164)	(.130)	(.128)**
Coeff. 1+ Coeff.2	.124	.021	-.022	.111	.017	-.005	.180	.030	-.032
	(.106)	(.090)	(.095)	(.120)	(.105)	(.110)	(.155)	(.130)	(.134)
Lag dependent var.		X			X			X	
School fixed effects			X			X			X
Obs.	8,812	8,220	15,088	8,812	8,220	15,088	8,812	8,220	15,088

The dependent variable in all columns is a standardized score on a math and English test administered to pupils in grades 1, 2 and 3 in 2009 and grades 3 and 4 in 2011. Columns 1, 4 and 7 use only the 2011 (follow-up) test data. 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. 'Mos. of Contract Teacher' is a continuous, and potentially endogenous, treatment variable measuring months of exposure to a contract teacher; Gov is an indicator variable for the government treatment arm. Standard errors are clustered at the school level.

Table 5: Intent-to-treat effects of cross-cutting interventions

	SMC Training			Local Hiring			High Salary		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pooling treatment arms:									
Z	.108	.086	.073	.144	.152	.136	.113	.102	.080
	(.100)	(.085)	(.093)	(.102)	(.088)*	(.095)	(.093)	(.081)	(.086)
$Z\times$ Cross-cut	.063	.036	.009	-.012	-.094	-.114	.104	.008	-.010
	(.120)	(.096)	(.096)	(.119)	(.095)	(.096)	(.153)	(.108)	(.095)
Coeff. 1 + Coeff. 2	.170	.122	.082	.133	.057	.022	.217	.109	.070
	(.116)	(.094)	(.094)	(.114)	(.090)	(.092)	(.154)	(.110)	(.099)
NGO vs gov't implementation:									
Z	.131	.202	.222	.065	.182	.204	.142	.174	.171
	(.115)	(.099)**	(.108)**	(.114)	(.101)*	(.114)*	(.111)	(.095)*	(.100)*
$Z\times$ Gov	-.046	-.227	-.300	.160	-.063	-.144	-.057	-.143	-.182
	(.147)	(.122)*	(.128)**	(.151)	(.131)	(.137)	(.129)	(.109)	(.113)
$Z\times$ Cross-cut	.045	-.036	-.095	.176	.003	-.059	.045	.042	.018
	(.174)	(.131)	(.127)	(.171)	(.130)	(.127)	(.240)	(.153)	(.125)
$Z\times$ Cross-cut $\times$ Gov	.035	.135	.208	-.376	-.191	-.098	.119	-.093	-.072
	(.239)	(.190)	(.188)	(.236)	(.186)	(.188)	(.304)	(.210)	(.182)
Cross-cut and NGO (Coeff. $1+3$)	.176	.166	.127	.241	.186	.145	.187	.216	.189
	(.162)	(.120)	(.113)	(.159)	(.117)	(.107)	(.233)	(.146)	(.118)
Cross-cut and Gov (Coeff. $1+2+3+4$)	.165	.073	.035	.025	-.068	-.096	.248	-.020	-.065
	(.137)	(.118)	(.120)	(.128)	(.106)	(.115)	(.175)	(.131)	(.122)
Lag dependent variable		X			X			X	
School fixed effects			X			X			X
Obs.	8,812	8,220	15,088	8,812	8,220	15,088	8,812	8,220	15,088

See notes for Table 4. Columns 1, 4 and 7 use only the 2011 (follow-up) test data. 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. In each column, the 'cross-cut' variable – denoting a cross-cutting experimental treatment or variation of the contract-teacher treatment – is defined according to the column heading. Standard errors are clustered at the school level.

Table 6: Robustness checks

	(1)	(2)	(3)	(4)	(5)
Z	.179	.189	.148	.176	.170
	(.089)**	(.096)**	(.103)	(.103)*	(.103)
$Z\times$ Gov	-.159	-.195	-.123	-.163	-.166
	(.093)*	(.093)**	(.105)	(.105)	(.105)
$Z \times$ Gov (upper Lee bound)		-.174			
		(.096)*			
% change in cohort enrollment			.002		
			(.0007)**		
% change in grade enrollment				.001	
				(.0009)	
Obs.	174	174	163	163	163
Lag dependent variable	X				
School fixed effects		X	X	X	X

The dependent variable in all columns is a standardized score on a math and English test administered to pupils in grades 1, 2 and 3 in 2009 and grades 3 and 4 in 2011 collapsed at the school level. Z represents an indicator variable for random assignment to any treatment arm; Gov is an indicator variable for the Ministry of Education treatment arm.

Table 7: Heterogeneous treatment effects

	Full Sample				Government & Control Sample				NGO & Control Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Z	.086	.076	.082	.078	-.013	-.022	.001	-.022	.184	.182	.172	.176
	(.082)	(.081)	(.080)	(.080)	(.096)	(.095)	(.092)	(.095)	(.095)*	(.091)**	(.091)*	(.090)*
$Z\times$ Density	.007				-.016				.017			
	(.053)				(.078)				(.068)			
$Z\times$ PTR		-.022				-.075				.051		
		(.047)				(.060)				(.066)		
$Z\times$ $Y_{t=0}$			-.115				-.161				-.044	
			(.050)**				(.070)**				(.064)	
$Z\times$ KCPE				.012				-.039				.054
				(.046)				(.065)				(.061)
Obs.	14,956	15,088	14,496	15,088	10,023	10,100	9,756	10,100	10,064	10,196	9,861	10,196

See notes for Table 4. All equations include school fixed effects and standard errors are clustered at the school level. Columns 1-4 include the full sample of schools. In columns 5-8 (9-12) the sample is restricted to the control schools and those assigned to treatment by the government (NGO). 'Density' measures the number of primary schools within a 5-mile radius of the school; PTR measures pupil-teacher-ratios; $Y_{t=0}$ is the average test score at baseline; and KCPE is the lagged, average national exam score in the school.

Table 8: Mechanisms

	Gov.	NGO	Difference	Corr. w/ test score gains
	(1)	(2)	(3)	(4)
Panel A: Socio-economic characteristics				
Age	29.983	29.760	.223	.002
			(.938)	(.010)
Female	.550	.294	.256	.057
			(.097)***	(.097)
Post-secondary education	.200	.020	.180	-.091
			(.064)***	(.145)
Advanced professional qualification	.100	.137	-.037	.097
			(.061)	(.145)
Friend/relative of teacher/SMC member	.667	.373	.294	.051
			(.100)***	(.100)
Panel B: Monitoring and accountability				
Presence in school	.628	.727	-.099	.101
			(.110)	(.134)
Any monitoring visit to school	.850	.961	-.111	.184
			(.053)**	(.155)
Average salary delay (months)	3.000	2.117	.883	-.056
			(.291)***	(.034)*
Panel C: Unionization and expectations				
Desire a long-term job	.632	.706	-.074	.027
			(.089)	(.107)
Union represented my interests	.377	.149	.228	-.197
			(.089)**	(.110)*
Took any union action during program	.428	.444	-.017	-.028
			(.041)	(.217)
Panel D: After the experiment				
Still working at program school	.379	.280	.099	.072
			(.098)	(.104)
Permanent and pensionable	.424	.469	-.046	.126
			(.092)	(.098)
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 72 observations in treatment schools. Standard errors are clustered at the school level. Dummy variables are defined as: "Presence in school" $= 1$ if the teacher was present in school during an announced visit; "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 "Permanent and pensionable" $= 1$ if the teacher is employed as a civil-service teacher after the end of the RCT. "Took any union action during program" is the average of the following dummy variables: the teacher joined the union after the program; teacher could explain the purpose of union strike action against the contract teacher program; teacher participated in the national strike in 2011. 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.

Table 9: Mechanisms: Political Economy and scaling up

	Union identification		Test-score gains	
	(1)	(2)	(3)	(4)
$Z \times$ Gov	0.084	0.157	-0.065	-0.075
	(0.101)	(0.116)	(0.149)	(0.119)
$Z \times$ NGO$\times$ Union exposure	0.083		0.040	
	(0.120)		(0.183)	
$Z \times$ Gov$\times$ Union exposure	0.548***		-0.304*	
	(0.168)		(0.154)	
$Z \times$ NGO$\times$ Exposure to gov't scale-up		-0.009		0.016
		(0.115)		(0.143)
$Z \times$ Gov$\times$ Exposure to gov't scale-up		0.121		-0.258*
		(0.154)		(0.141)
Observations	100	95	102	107

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 2009-2011. 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. "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. Standard errors are clustered at the school level.

9 Appendix

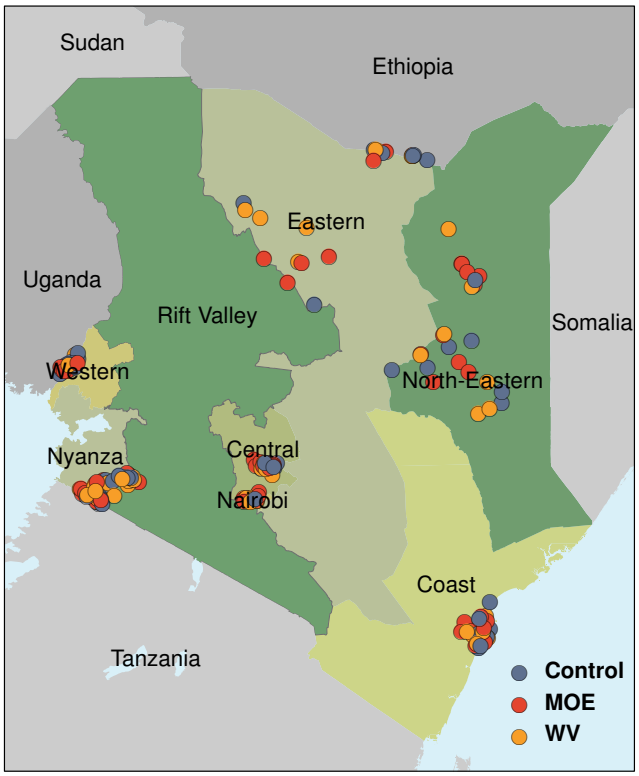

Figure 1: Treatment & control sites across Kenya's 8 provinces. (MOE and WV denote implementation by the Ministry of Education and World Vision, respectively.)

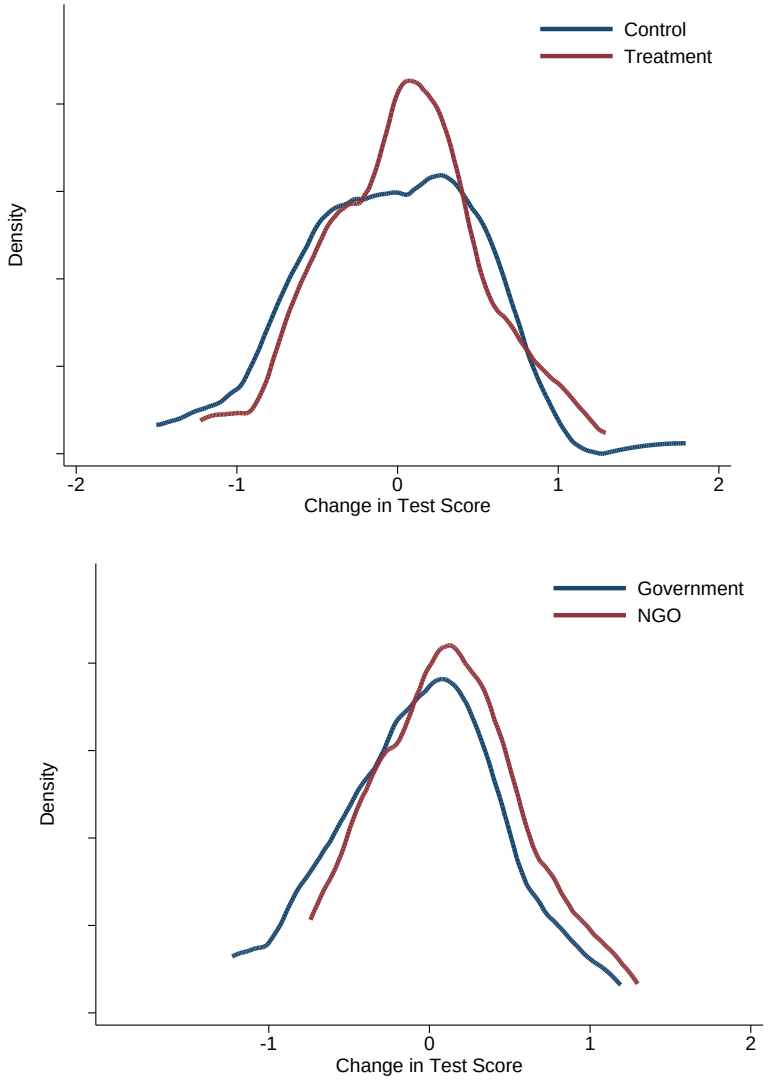

Figure 2: Kernel densities of change in school-level average test scores by treatment status.

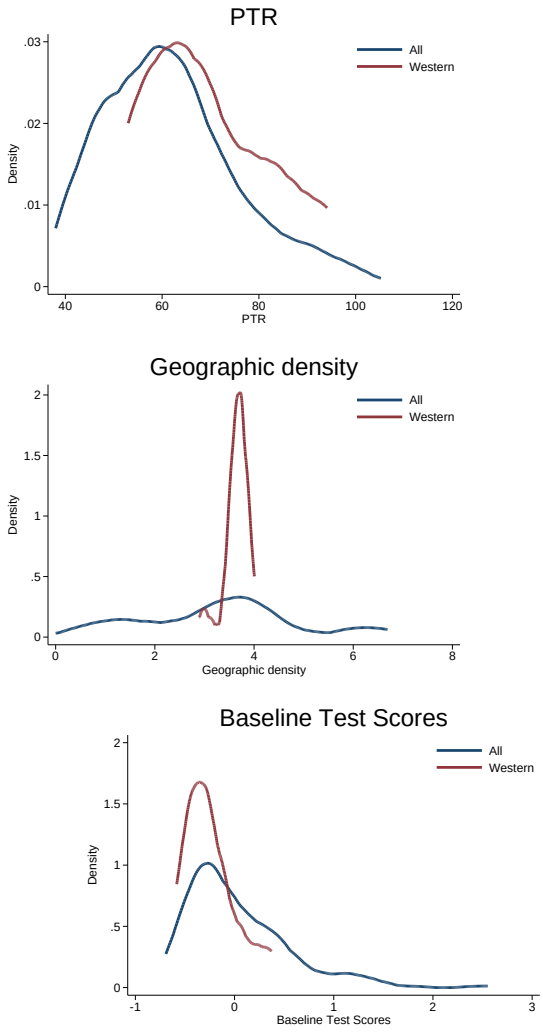

Figure 3: Kernel densities of baseline school characteristics.