Interventions & Institutions Experimental Evidence on Scaling up Education Reforms in Kenya*

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

Preliminary Draft - Comments Welcome

April 11, 2012

Abstract

The recent wave of randomized trials in development economics has provoked criticisms regarding external validity and the neglect of political economy. We investigate these concerns in a randomized trial designed to assess the prospects for scaling-up a contract teacher intervention in Kenya, previously shown to raise test scores for primary students in Western Kenya and various locations in India. The intervention was implemented in parallel in all eight Kenyan provinces by a non-governmental organization (NGO) and the Kenyan government. Institutional differences had large effects on contract teacher performance. We find a significant, positive effect of 0.19 standard deviations on math and English scores in schools randomly assigned to NGO implementation, and zero effect in schools receiving contract teachers from the Ministry of Education. We discuss political economy factors underlying this disparity, and suggest the need for future work on scaling up proven interventions to work within public sector institutions.

*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 for their leadership throughout the process. 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.

1 Introduction

The recent wave of randomized trials in development economics has catalogued a number of cost-effective, small-scale interventions proven to improve learning, health, and other welfare outcomes. This methodology has also provoked a number of criticisms regarding the generalizability of experimental findings, including concerns about external validity, general equilibrium effects, and the neglect of political economy in much of the evaluation literature (Acemoglu 2010, Deaton 2010, Heckman 1991, Rodrik 2009). These criticisms are particularly relevant when randomized trials of pilot projects run by well-organized and monitored NGOs are used as the basis for policy prescriptions at the national or global level. As noted by Banerjee and Duflo (2008), "what distinguishes possible partners for randomized evaluations is competence and a willingness to implement projects as planned. These may be lost when the project scales up. [...] Not enough effort has taken place so far in trying 'medium scale' evaluation of programs that have been successful on a small scale, where these implementation issues would become evident."

In this paper we employ the methodology of randomized trials to assess these substantive concerns about political and institutional constraints and measure precisely how treatment effects change when scaling up. We analyze a policy experiment in Kenya comparing the effectiveness of NGO and government implementation, and testing for the presence of heterogeneous treatment effects in a nationwide sample. The question of NGO versus government implementation is paramount to the formulation of national policy. Even in a fairly aid-dependent economy such as Kenya, the vast bulk of spending on primary education comes from domestic revenue channeled through the Ministry of Education, making the government the sole institutional actor capable of taking education polices to a national scale.[1]

[1]Government schools account for 90.2% of gross primary enrollment in Kenya. Furthermore, as of 2005 the Ministry's budget for primary education totalled $731 million (Otieno 2009); in contrast, NGOs working on education received just $4 million in international aid in 2009 (OECD 2012).

At the school level, this study replicates one of the most extensively tested, successful interventions to raise student learning in primary schools: the provision of contract teachers. Banerjee, Cole, Duflo and Linden (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 treatment schools of 0.15 and 0.13 standard deviations on math and language tests, respectively. In both cases, the additional teachers lead to significant learning gains despite salary costs that are a small fraction of civil service wages. Finally, of particular relevance for the present study given its geographic focus, Duflo, Dupas and Kremer (2009) show that exposure to a contract teacher 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.[2]

We report on the randomized evaluation of the pilot phase of a nationwide program which now employs over 18,000 teachers. The pilot was designed to test the Ministry of Education's ability to implement a fairly close variant of the NGO project described by Duflo et al. (2009) and to replicate the results across diverse conditions, spanning urban slums in Nairobi and nomadic communities in the remote Northeastern province. As part of the government's contract teacher pilot, 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

[2]See Bruns, Filmer and Patrinos (2011) for a summary of additional, non-experimental results on the impact of contract teachers, including Bourdon, Frolich 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.

the local affiliate of an international NGO, World Vision Kenya. The timing, salary levels, recruitment procedures and all other experimental protocols were held constant across the government and NGO arms of the evaluation.

While we find positive and significant effects of the program overall, these are concentrated entirely in schools where the contract teacher program was administered by an international NGO. Effects were significantly smaller and indistinguishable from zero in schools receiving contract teachers from the Ministry of Education. To understand the difference in outcomes between government and NGO implementation, we explore various ways in which the political economy of teacher employment in Kenya could have adversely affected the performance of teachers in the government treatment arm. Specifically, we discuss the role played by the teachers' union, which waged an intense political and legal battle that successfully altered the the contract teacher program in subsequent years in ways that may have undermined some of its incentive effects.

Examining the emerging literature on randomized trials in education in developing countries, and specifically the sub-set of those studies which measure impacts on test scores or other learning outcomes, it appears that working with governments places some constraints on the scope of interventions that can be tried.[3] NGO pilot programs have tested a wide range of interventions, particularly in India and Kenya, with a strong focus on accountability reforms and changes to teacher incentives.[4] Meanwhile, programs with government involve-

[3]Of the 31 studies we examined – all based on randomized trials in developing countries measuring impacts on test scores or other learning outcomes – 16 were conducted in Asia (13 of which in India), 11 in Africa (7 of which in Kenya) and 4 in Latin America. (See citations in subsequent footnotes.) Roughly half of the studies cite significant government involvement in project implementation (14 of 29), including all the Latin American studies, roughly half the Asian studies, and less than one third of the African studies.

[4]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 2010), and as already discussed, contract teachers or tutors (Banerjee et al. 2007, Muralidharan and Sundararaman 2010). Similarly in Kenya, NGO pilot programs have examined the impact of contract teachers and tracking students (Duflo,

ment have, by and large, tended to focus on increasing school inputs of various kinds, and are more likely to occur in Latin America and Asia, and less so in Africa.[5][6] This pattern is reminiscent of Heckman's (1991) concept of randomization bias, stemming from the self-selection of service providers into randomized trials. This paper attempts to overcome this self-selection of NGOs into randomized trials by involving government directly in the implementation of a school-level accountability reform, and demonstrates the potential magnitude of this bias. By comparing government and NGO implementation, our paper illustrates that RCTs can be a useful tool not only to identify cost-effective ways to enhance school performance, but also to identify key bureaucratic and political constraints that adversely effects governments' ability to scale up interventions that have been shown to work.

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 presents the main treatment effect estimates, comparing the relative effectiveness of NGO and Ministry 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 4.4 explores possible mechanisms explaining the government-NGO performance gap. Finally, Section 5 tests for heterogeneous

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.

[5]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 2010), 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 (2011), and school construction in Afghanistan (Burde and Linden 2010) and reforms to local school management in Madagascar (Glewwe and Maïga 2011).

[6]Notable exceptions to this pattern that we are aware of include the evaluation of World Bank-financed school management reform program in Madagascar, cited above (Glewwe and Maïga 2011).

treatment effects, finding no geographic differences or differences by initial class size, but some evidence that schools with lower initial test scores benefitted more from a contract teacher. Section 6 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).

The public education system is highly centralized. Officially all resources for the operation and maintenance of public schools flow through the Ministry of Education via two channels: non-salary expenditures deposited in school bank accounts, and teacher salaries paid directly to civil servants. (See Figure 1.) Each of these channels is problematic, as evinced by recent scandals involving embezzlement of school funds and payments to 'ghost teachers'.

2.1 School finance

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.[7] At the school level, FPE funds are held in a school bank account administered by a governing body known

[7]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.

as a school management committee (SMC). The SMC is chaired by the head teacher and comprised of representatives from the Ministry, parents from each grade, teachers, and in some cases local community or religious organizations.

Misappropriation of FPE funds was at the center of a major corruption scandal which emerged in 2009. An external audit commissioned by the Ministry of Education showed that actual funds disbursed to school bank accounts fell short of the allocated amount by $4.71 per pupil in 2005 and by smaller but significant amounts in other years. Press reports estimated that anywhere between $68 million and $590 million of the FPE budget had been misdirected between 2004 and 2008 (Teyie and Wanyama 2010), leading the President to suspend several top Ministry officials, and foreign donors including DfID and USAID to freeze aid disbursements in December 2009.

2.2 Civil service 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. In practice, schools also informally contract local teachers known as Parent-Teacher Association (PTA) teachers. In the sample of schools surveyed for this study, 83% of teachers were employed by TSC and the remaining 17% by PTAs. TSC teachers earned an average of $261 per month in 2009, compared to just $56 per month for PTA teachers.

The relatively high salaries of TSC teachers create an extreme form of labor market disequilibrium. On the demand side, the high salaries and the Ministry's limited budget lead to unfilled teacher vacancies. At the beginning of 2011 the Ministry of Education reported a shortage of 61,000 teachers (across roughly 20,000 primary schools) relative to its target of a 40:1 pupil-teacher ratio. On the supply side, high salaries attract a long queue of job applicants. TSC hires on the basis of an algorithm that primarily rewards seniority: the

first applicants to graduate from teacher training college are the first to be hired. In 2010, Ministry records show that most successful applicants to civil service positions had been in the job queue for 8 to 11 years. PTA teachers are often drawn from this queue of graduates.

These features contribute to limited accountability for civil service teachers vis-a-vis parents or school management committees. Salaries are paid directly from Nairobi to individual teachers' bank accounts. And because of the chronic teacher shortages, parents and schools have little incentive to pursue disciplinary action against teachers; if a teacher is reassigned or terminated, a school may wait months or years for a replacement.

2.3 Contract teachers

Motivated by a desire to fill teacher vacancies and regularize PTA teachers, in 2009 the Directorate of Basic Education within the Ministry of Education proposed an initiative to provide funds to schools to employ teachers on contract outside of the TSC system. A steering committee – including Ministry officials and the current authors – was formed to design a pilot program, evaluate its impacts, and report back to the Permanent Secretary.

Under pressure from the Ministry of Finance to spend funds as part of an economic stimulus package, the Ministry opted to scale-up the contract teacher program before the pilot was completed. Thus 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. In 2011 the Ministry succumbed to political pressure and agreed to allow the contract teachers to unionize and subsequently to hire all 18,000 contract teachers into the civil service at the end of their contracts.[8]

[8]From an evaluation perspective, an obvious concern is that the allocation of these 18,000 contract teachers contaminated the randomly allocated teachers from the pilot program. It is important to note that allocation of contract teachers to schools for the full-scale program – while not itself randomized – was done on the

2.4 Organizational structure of the Ministry and NGO

World Vision is an international NGO with affiliate offices in both donor and implementing countries. In implementing countries, World Vision's activities are organized in Area Development Programmes, geographic zones somewhat smaller than a district. In Kenya, World Vision ADPs are active in only a small fraction of the total districts. Outside of these areas, World Vision is not present – highlighting again the constraints to scaling up with a non-governmental service provider. Within the ADPs, World Vision employs permanent staff and paid "volunteers", who monitor and implement all World Vision program activities.

In the districts, the Ministry relies on its own local staff – the Education Officer, Staffing Officer, Quality Assurance and Standard Officers (QASOs) and the Teacher Assistant Centre (TAC) tutors – to monitor schools and teachers. In principle, the QASOs and TAC tutors should make routine visits to all schools. The Ministry's ability to directly call on DEOs to carry out specific tasks in their districts is limited by the fact that all communication to the district has to go via the Provincial Directors of Education.

Comparing these structures, it is also worth noting that salaries are higher in the NGO sector, and World Vision field offices are likely to be better equipped than District Education Offices with resources such as vehicles, fuel, generators and computers.

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.

basis of pupil-teacher ratios measured during the first quarter of 2010, i.e., prior to the random assignment. This ensured that the influx of 18,000 new teachers did not offset or in any way respond to the randomly allocated pilot component.

In order to disentangle the effect of the various contractual and programmatic arrangements, four variants of the basic contract teacher treatment were also randomly assigned: (i) government versus NGO implementation of the overall program, (ii) training for school management committees, (iii) local versus centralized recruitment and payment of contract teachers, and (iv) two alternative salary offers, equivalent to approximately $121 and $67 per month, respectively. Each of these four dimensions involved a binary choice, yielding four potential treatment cells and one pure control cell.

3.1 Program details

Contract teachers were randomly assigned to teach either grade 2 or 3.[9] As noted above, the contract teacher intervention combines both a class-size effect and the effect of changing teacher incentives. 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. For example, a school which, prior to the experiment, had a single civil service teacher instructing 70 grade 3 pupils would have been asked to split grade 3 into two classes, one taught by the pre-existing civil service teacher and the other taught by the contract teacher. As discussed below, compliance with these instructions was high but imperfect. Field monitors were able to ensure experimental teachers were assigned to the correct class, but had difficulty ensuring that other teaching staff were not reallocated to spread the teaching load more evenly, particulary in 2010. By the 2011 school year, reports from field visits suggested that compliance was fairly uniform on this front.

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,

[9]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. Thus there is some experimental variation in the length of direct exposure to the program within the treatment group.

based in part on the location of the offices of the partnering NGO.[10] Within each province, schools with a pupil-teacher ratio below the median were excluded from the sampling frame. Using this sampling frame of high pupil-teacher ratio schools, schools were chosen through simple random sampling within the selected districts.

The effects of the randomized interventions are measured by comparing baseline and follow-up academic assessments (exams) in math and English in 24 primary schools in each of Kenya's 8 provinces (192 total 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, with a sample of approximately 23,000 pupils. 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.

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. Figure 3 summarizes the partial factorial design of these various treatments. Each dimension presents a trade-off between what was deemed politically most feasible, versus what had previously been shown to be effective in small scale NGO trials. Ministry officials agreed to attempt the latter, more politically contentious forms of the project in hopes of producing experimental evidence that would help the Ministry make a political case for the technically-preferred

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

design.

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 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 explored 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 cell was more similar to the civil service model. Teachers were paid by the Ministry or World Vision headquarters in Nairobi, via direct deposits into personal accounts. In addition, teachers in the central cell were recruited from the district short-list of candidates from the previous round of civil service hiring. District Education Officers were instructed to hire from the pool of marginal rejects for civil service jobs. Since the ranking of Teacher Service Commission candidates is primarily based on time since graduation from teacher training college, marginal rejects would be more or less guaranteed a job in the next round off hiring.

At the time, the Ministry and teacher unions were actively debating whether or not the 18,000 contract teachers to be hired in the full scale-up would be guaranteed civil service employment at the end of their two-year contracts. The hypothesis to be tested in this cross-cut was whether an effective employment guarantee would dull dynamic incentives. Teachers hired locally constitute a relevant control group as they would on average be more junior with little or no prospect of graduating into civil service employment.

School management committee training To explore the importance of local accountability on teacher (and in turn, student) performance, and in line with Duflo et al. (2009), in

half of the treatment schools, members of the school management committee were invited 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, enrolment, number of classrooms, number of civil service teachers, number contract 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. The successful outcome of the randomization is reported in Table 1. Figure 2 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, in particular, baseline test scores for grades 2 and 3. Denote by Y_{ijt} the outcome of interest for pupil i in school j in period t. Let Z_j denote being randomly assigned treatment status, i.e. eligibility to receive an additional contract teacher. Let SMC_j denote the subset of treatment schools that are

randomly assigned to receive school management committee training. Finally, let $Gov_j = 1$ denote a treatment school where the intervention is implemented by the government and $\text{NGO}_j = 1$ a treatment school where the intervention is implemented by the NGO.

To examine whether the treatment and control schools are comparable prior to the intervention, we estimate

$$Y_{ij,t=0} = \alpha_0 + \beta_0 Z_j + \beta_0' Z_j \times \text{SMC}_j + \beta_0'' Z_j \times \text{Gov}_j + \beta_0''' Z_j \times \text{SMC}_j \times \text{Gov}_j + \varepsilon_{0ij,t=0}$$

using the baseline data. As seen from Table 2, none of the treatment dummies are significant, implying that test scores in treatment and control schools were indistinguishable prior to the intervention.

4 Comparative effectiveness of government and NGO programs

As noted in the introduction, a necessary step in scaling up any proven NGO education intervention in Kenya – as in many other settings – will be a transition to working with the government as monopoly supplier of public education at the national level. The experiment here is designed to address this central question of whether the Kenyan government replicate NGO impacts. Section 4.1 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 4.2 examines the most basic element of program compliance: successfully recruiting a contract teacher. Section 4.3 builds on this definition of compliance, defining "treatment" as the actual

presence of a contract teacher in a given school in a given month, and presents instrumental variables estimates of the impact of actual treatment status (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, while statistically significant, do nothing to explain differences in treatment effects.

4.1 ITT effects

We begin by estimating the average intention-to-treat (ITT) effect of school-level assignment to the contract teacher program, then proceed to compare the effects of the NGO and government treatment arms. The ITT effect is measured by the coefficient on the random assignment variable Z_{jt} in equation (1).

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

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 ($\mathbf{X}_{ijt}$): first, a simple cross-sectional OLS regression with no controls; second, controlling for initial tests 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, respectively. The point estimate is fairly consistent across all three specifications, at approximately 0.1 standard deviations, though marginally significant only in the lagged dependent variable model.

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

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

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

Across all specifications, the results consistently suggest 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 causal of effect of assignment to NGO versus government implementation of the project. The coefficient on Z_{jt} shows that NGO implementation raises scores by 0.16 to 0.19 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 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 0.04 in the lagged dependent variable model and -.02 in the fixed effects model, with standard errors of .081 and .088, respectively.

4.2 Compliance: teacher recruitment

The Z variable in the ITT analysis above distinguishes the 128 schools assigned to receive a contract teacher. In practice, schools had mixed success in recruiting contract teachers,

and 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, 56 (55) were successful in hiring a contract teacher at some point during the programme. However, teachers did not necessarily stay with the school for the entire duration of the programme and when a vacancy opened up, it was not always filled. As a consequence, out of the 18 months of the programme, schools in the government (NGO) arm actually employed a teacher for 11.59 (13) months on average. If we exclude schools that never employed a teacher from this calculation, the numbers rise to 13.25 and 15.13 months respectively.

Table 3 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 with-out controls for school characteristics.

We examine three experimental determinants of teacher labor supply. First, Table 3 shows that 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. 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 16% across specifications. Third, NGO implementation led to between 12 and 17% more months with a filled vacancy, relative to the government treatment arm, and this effect is significant across all specifications. 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 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.

4.3 ATT effects

Can differences in the probability of filling contract teacher vacancies explain the difference in government and NGO performance? We address this question by estimating average treatment on the treated (ATT) effects, examining whether NGO-government differences persist where the program was implemented successfully. Clearly, successful recruitment of a contract teacher is highly endogenous to factors such as the quality of school management which may also directly affect pupil performance. We assert, however, that the contract teacher program will affect student performance if and only if a teacher is successfully hired. Thus random assignment satisfies the exclusion restriction for a valid instrument for contract teacher presence, allowing us to estimate ATT effects for both the government and NGO program.[12]

As a benchmark, we present a naïve OLS regression of test scores on treatment status, where T_{jt} measures the number of months (out of a possible 18 months total duration of the program) that a contract teacher was in place in a given school.

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

Columns 4 to 6 of 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 between 0.1 to 0.14 standard deviations,

[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]Note that merely hiring a contract teacher might be considered a fairly minimal level of compliance. In Section 4.4 we discuss the possible effects of endogenous salary delays and teacher turnover. Crucially, these refinements to the concept of compliance do nothing to undermine the ignorability of our instrument here, as they require teacher presence as a minimum starting point.

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 18 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 18-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. 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) ar 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

$$Y_{ijt} = \alpha_4 + \beta_4 \hat{T}_{jt} + \gamma_4 \mathbf{X}_{ijt} + \varepsilon_{4ijt} \tag{4}$$

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

$$T_{jt} = \alpha_5 + \delta_5 Z_{jt} + \gamma_5 \mathbf{X}_{ijt} + \varepsilon_{5ijt}. \tag{5}$$

Results from estimating equation (4) 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 ATT effect ranges from 0.22 in the cross-section to 0.26 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 significant negative effect, with the point estimate implying zero ATT effect in the government

treatment arm.

4.4 Why did the government treatment arm fail?

The government's ambitious plan to employ 18,000 contract teachers nationwide posed a significant threat to the Kenyan National Union of Teachers. As Acemoglu, Johnson, Querubín and Robinson (2008) note, 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, which they refer to as a "seesaw effect". In this case, 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 the threat of a national strike. The political battle eventually altered the program in two key respects.

First, primary responsibility for employing contract teachers (both during the pilot and the full scale-up) was shifted away from its natural home within the government – the Teacher Service Commission, which manages recruitment and payroll for all civil service teachers – into other offices of the Ministry after the Teacher Service Commission refused to employ contract teachers. While implementation via the Teacher Service Commission posed its own challenges, the need to set up parallel systems elsewhere in the Ministry contributed to salary delays and other implementation challenges.

Second, by June 2011 the Ministry acquiesced to union demands to absorb the contract teachers into civil service employment at the end of their contracts. This result reflects a form of general equilibrium effect from scaling up: 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. In conversations with the research team, senior officers in the Ministry of Education not only acknowledged this dynamic, but endorsed it as a way for the Ministry to win support from the State House

and Ministry of Finance for more civil service teachers.

In the following paragraphs we explore the implications of these two "seesaw effects" – resulting in salary delays and changing career incentives – and highlight other, additional mechanisms that might explain what went wrong in the government implementation arm. Experimental variation in the program's implementation – e.g., the addition of school management committee training to improve local accountability, or reliance on local hiring – enables us to examine where and when the NGO-government gap was most severe, providing clues as to the underlying mechanisms at work. Nevertheless, we caution that the experiment does not provide a definitive test of these mechanism hypotheses, statistical power is weak for the analysis of cross-cuts within the government and NGO treatment arms separately, and thus the discussion here is somewhat speculative.

4.4.1 Salary delays and teacher turnover

As seen in Table 6, the parallel payroll system setup within the Ministry of Education for the contract teacher program - following the political compromise in which the Teacher Service Commission would not directly employ contract teachers - was associated with significant salary delays and, in turn, teacher turnover. The fourth and fifth rows of Table 6 summarize these delays in paying salaries. The average salary delay was 1 month in schools in the NGO implementation arm and more than twice as high – 2.33 months on average – in schools in the government implementation arm. In addition, there was large variation in the disbursement of salaries for schools where the intervention was administered by the Ministry of Education. The average maximum delay for that treatment arm was 5.56 months and 10% of teachers had to wait for their salaries for 10 months at some point. Teacher turnover in government-administered schools was also significantly higher, possibly as a result of salary delays.[13]

[13]We have no information about salary delays in the case of World Vision at this time.

4.4.2 Career incentives and teachers' expectations

The Ministry's concession - during the midst of this evaluation - to employ contract teachers as permanent civil servants at the end of their two-year term may have also dulled the incentive advantages of the contract teacher model in the government treatment arm.

We hypothesize that this erosion of contract teacher accountability was strongest in the sub-set of schools implemented by the Ministry of Education with central as opposed to local hiring. Contract teachers hired locally were not necessarily aware whether implementation was done by the NGO or the government, and had less reason to associate news reports from Nairobi about the fate of 'contract teachers' with their own career prospects.

An alternative hypothesis is that the prospect of permanent employment may have heightened long-run career incentives for contract teachers - particularly those employed directly by the District Education Officer in the government treatment arm.

Columns (4)-(6) in Table 5 show the marginal effect of local hiring in each treatment. The difference between local and central hiring is insignificant in all specifications in Table 5. However, the point estimates suggest the prospect of long-term civil service employment may have raised rather than dampened performance incentives. Pupil scores in central hiring schools are higher, particularly in the case of the government treatment arm. Results are consistent with the hypothesis that frequent interaction with the Ministry's District Education Officers or NGO officials increased incentives for good performance by raising expectations about future career prospects, though admittedly these results are suggestive at best given the lack of statistical significance.

4.4.3 Monitoring and accountability

Even abstracting from the political drama surrounding the contract teacher program, 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. 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).

In order to compensate for short-comings in top-down accountability, the school management committee training was designed to strengthen local accountability and monitoring among parents and community members. We hypothesize that this should moderate the positive effect of NGO implementation and narrow the gap between the government and NGO implementation arms. Columns (1)-(3) report in Table 5 report the effect of school management committee training on pupil test scores. While none of the coefficients are significant, the point estimates are in line with this hypothesis. In both the lagged-dependent variable and fixed effects specifications, the coefficient on school management committee training has a positive sign only in the government treatment arm.

4.4.4 Recruitment and teacher quality

The protocol for recruitment was the same for the government and the NGO. For the central treatment arm, recruitment was to be from the Teacher Service Commission short-list of applicants for civil service position, hiring from the pool of marginal rejects. In the local treatment arm, schools were at liberty to hire any candidate with minimum teaching qualifications and were not required to draw from the civil service short-list.

While the protocols were identical in theory, in practice, the NGO may have put more effort into recruiting high quality candidates. It is possible to test this hypothesis by comparing the observable characteristics of contract teachers hired in each treatment arm.[14]

[14]This section is preliminary and incomplete. Further data on characteristics of teachers in the two treatment arms is pending.

5 Heterogeneous effects

In addition to the institutional considerations raised above, a more traditional concern about the generalizability of RCT results is external validity. The broad geographic dispersion of our sample is helpful in addressing 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 2010). On the former question, the literature on medical trials commonly takes a data-driven approach based on boosting algorithms (Friedman, Hastie and Tibshirani 2000). Boosting is particularly well-suited to the design of optimal treatment regimens for a particular sub-group. 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.

$$Y_{ijt} = \alpha_6 + \beta_6 Z_{jt} + \beta_6^x \left(Z_{jt} \times \frac{\mathbf{X}_{jt} - \mu_x}{\sigma_x} \right) + \gamma_6 \mathbf{X}_{ijt} + \varepsilon_{6ijt} \tag{6}$$

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.

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. 2009).

Table 7 shows the results from estimating the heterogeneous ITT effects in equation (6). 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 is roughly 4 percentage points smaller in locations with a higher density of schools, contradicting our hypothesis – though this effect is 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 no relationship between scores on the national test and treatment effects. Column 4, however, shows a significantly negative relationship between initial test scores in the baseline and subsequent treatment effects. While the average ITT for schools with NGO-implementation was roughly $1/5^{th}$ of a standard deviation, column 4 implies this effect was only half as large in schools one standard deviation above the mean.

So far we have ignored the issues raised by conducting multiple comparisons. Testing m null hypotheses at a significance level of α, Boole's inequality predicts that at least one null

will be rejected with probability less than or equal to $m\alpha$. The Bonferroni correction limits this probability, known as the family-wise error rate (FWER), by testing each individual hypothesis against the corrected critical value $\alpha' = \alpha/m$. As Fink, McConnell and Vollmer (2011) show, the Bonferroni correction is quite conservative, in the sense of controlling Type I errors at the expense of more Type II errors (less power) vis-a-vis available alternatives. Benjamini and Hochberg (1995) and subsequent authors have proposed alternatives which minimize the false-discovery rate (FDR) rather than the FWER, defined as the proportion of the rejected null hypothesis which are erroneously rejected, leading to greater power.[15]

We apply Benjamini and Hochberg's (1995) method to the estimates in Table 7. The correction does not affect the coefficients or standard errors, but rather the critical value (p-value) used to establish statistical significance. As shown in Figure 6, this amounts to literally 'raising the bar' for statistical significance. In our particular example, the results are fairly unremarkable: only one interaction term in Table 7 was statistically significant at the 5% level when considered in isolation – the interaction of the Ministry of Education treatment arm with baseline test scores. This effect remains significant at the equivalent of the 5% level (now 0.625%) using corrected p-values.

6 Conclusion

As Reinikka and Svensson (2005) argue,

> "When scaling-up a specific program found to work in a controlled experiment run by a specific organization (often an NGO with substantial assistance from the research team), it is crucial also to have an understanding of the whole delivery chain. [...] Lack of attention to the service delivery system, and adjustment of

[15]For an intuitive exposition of the advantages of FDR versus FWER corrections see Fink et al. (2011). For a comprehensive classification of the corrections proposed to date, see Newson (2003).

policy accordingly, may imply effects very different from what a simple extrapolation of the estimates of the controlled experiment produces."

In this paper, we show that these concerns are of quantitative importance. We report on a randomized trial replicating earlier results showing that contract teachers significantly raise pupil test scores when implemented by an international NGO. These effects disappear entirely when the program is implemented with the bureaucratic structures of the Kenyan government.

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.

This lesson is relevant to debates on the generalizability of RCT results beyond development 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 would arguably fulfill the 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 results suggest that NGO involvement in the overall program management constitutes a significant departure from a scale-able model of government implementation.

Our concern with institutional context is particularly salient as economists adapt the methods of randomized trials – commonly used to evaluate technical innovations like new drugs or teaching methods – to examine tweaks to the incentives, contract types, and accountability structures under which the curriculum or medication is administered. These reforms are often politically contentious and, to operate at scale, must be implemented by public sector bureaucracies in weakly governed states. The fate of Kenya's contract teacher program is a reminder that in many cases, institution are not broken by accident. The discussion in Section 4.4 suggested that many of the obstacles to smooth implementation of the program by the Ministry of Education were not of a function of 'low capacity' or inherent bureaucratic inefficiency, but a result of the endogenous political economy response to a program that threatened vested interests.

In conclusion, we do not believe that our results justify abandoning government implementation to focus resources on NGOs, or in any way undermine the role of RCTs in policy research. Rather, we see this is a multi-stage process. Randomized trials with flexible NGO partners can identify organizational and institutional reforms at the school or clinic level as candidates for scale-up. The missing stage, which we have attempted to demonstrate in this paper, involves further experimentation with developing-country governments to identify - and in future, hopefully overcome - implementation constraints higher up the institutional hierarchy.

References

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

___ , Simon Johnson, Pablo Querubín, and James A. Robinson, "When Does Policy Reform Work - The Case of Central Bank Independence," *Brookings Papers on Economic Activity*, 2008, (1), 351–418.

Baird, Sarah, Craig McIntosh, and Berk Özler, "Cash or Condition? Evidence from a Cash Transfer Experiment," *World Bank Policy Research Working Paper*, 2010, *5259.*

Banerjee, Abhijit and Esther Duflo, "The Experimental Approach to Development Economics," *NBER Working Paper Series*, November 2008, (Working Paper 14467).

___ , 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.

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.

Benjamini, Y. and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, 1995, *Series B*, 289–300.

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. Frolich, 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," 2010.

Colclough, Christopher Otieno Wycliffe &, "Financing Education in Kenya: Expenditures, Outcomes and the Role of International Aid," *RECOUP Working Paper*, 2009, (25).

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

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Additional Resources versus Organizational Changes in Education: Experimental Evidence from Kenya," 2009.

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

___ , Rema Hanna, and Stephen Ryan, "Incentives Work: Getting Teachers to Come to School," 2010.

Fink, Gunther, Margaret McConnell, and Sebastian Vollmer, "Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures," 2011.

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," 2011.

___ **and Eugenie Maïga**, "The Impacts of School Management Reforms in Madagascar: Do the Impacts Vary by Teacher Type?," 2011.

___ **, 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 Large-Scale Experiments using Bayesian Additive Regression Trees," 2010.

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 Technical Working Paper Series*, July 1991, (107).

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, Edwar 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.

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.*

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.

Newson, Roger, "Multiple-test procedures and smile plot," *The Stata Journal*, 2003, *3* (2), 109–132.

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

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.

Rodrik, Dani, "The New Development Economics: We Shall Experiment, but How Shall We Learn?," in Jessica Cohen and William Easterly, eds., *What Works in Development? Thinking Big and Thinking Small*, Brookings, 2009.

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 rst 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.

Teyie, Andrew and Henry Wanyama, "Losses in FPE rise to Sh5.5 billion," *Nairobi Star*, 2010, *8 January.* Available online at http://allafrica.com.

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.*

7 Appendix: Figures and Tables

Table 1: Results of optimal multivariate matching algorithm

	Control	Treatment	Difference
Enrolment	43.33	53.26	9.935
			(7.418)
No. of classrooms	11.76	12.48	.715
			(1.046)
No. of civil service teachers	10.02	10.21	.195
			(1.002)
No. of contract teachers	1.90	2.27	.369
			(.347)
Average pay for contract teacher	2843	3393	550.103
			(531.535)
KCPE	239.48	235.083	-4.396
			(6.783)
Grade 1 English	.028	.074	.046
			(.166)
Grade 1 Maths	.060	.063	.003
			(.156)

Regressions based on 192 schools, collapsed at school level

Table 2: Differences in test scores in treatment and control schools prior to the intervention

	(1)	(2)	(3)
Z	.087	.039	.070
	(.083)	(.095)	(.094)
Z × Gov		.099	
		(.098)	
Z × SMC			.034
			(.099)
Obs.	6,264	6,264	6,264

Regressions based on 174 schools. Standard errors are clustered at the school level.

Table 3: Labor supply of contract teachers

	Linear Probability Model		Logit Model	
	(1)	(2)	(3)	(4)
High salary	.116	.087	.115	.089
	(.064)*	(.068)	(.064)*	(.068)
NGO implementation	.123	.166	.124	.170
	(.065)*	(.064)***	(.066)*	(.067)**
Local recruitment	.143	.162	.144	.157
	(.065)**	(.063)**	(.066)**	(.067)**
Geographic density		-.004		-.003
		(.002)**		(.002)*
Lagged KCPE score		.001		.002
		(.001)		(.001)
Pupil-teacher ratio		.003		.004
		(.002)		(.003)
Obs.	2,044	1,977	2,044	1,977

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 and 3 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 4: Treatment effects

	ITT			OLS			ATT		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pooling treatment arms:									
Z	.141	.115	.083						
	(.077)*	(.068)*	(.076)						
T				.139	.129	.122	.205	.166	.119
				(.089)	(.075)*	(.081)	(.113)*	(.098)*	(.108)
NGO vs gov't implementation:									
Z	.163	.190	.180						
	(.094)*	(.078)**	(.084)**						
$Z\times$ Gov	-.043	-.151	-.197						
	(.102)	(.083)*	(.085)**						
T				.167	.219	.238	.224	.259	.245
				(.111)	(.086)**	(.088)***	(.129)*	(.106)**	(.114)**
$T\times$ Gov				-.063	-.199	-.258	-.039	-.201	-.270
				(.133)	(.110)*	(.111)**	(.151)	(.121)*	(.122)**
Lag dependent variable		X			X			X	
School fixed effects			X			X			X
Obs.	8,711	8,154	14,975	8,711	8,154	14,975	8,711	8,154	14,975

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. Z represents an indicator variable for random assignment to any treatment arm; T is a continuous, and potentially endogenous, treatment variable measuring months of exposure to a contract teacher; *Gov* is an indicator variable for the Ministry of Education 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	.113	.099	.085	.150	.158	.135	.125	.117	.091
	(.086)	(.076)	(.087)	(.088)*	(.079)**	(.089)	(.081)	(.072)	(.081)
$Z\times$ Cross-cut	.057	.033	-.003	-.018	-.084	-.101	.065	-.009	-.036
	(.103)	(.084)	(.086)	(.102)	(.083)	(.086)	(.128)	(.092)	(.083)
NGO vs gov't implementation:									
Z	.129	.193	.219	.090	.185	.211	.162	.192	.188
	(.100)	(.088)**	(.101)**	(.098)	(.089)**	(.108)*	(.095)*	(.084)**	(.093)**
$Z\times$ Gov	-.033	-.185	-.272	.122	-.055	-.160	-.074	-.148	-.194
	(.128)	(.107)*	(.116)**	(.132)	(.117)	(.126)	(.113)	(.096)	(.102)*
$Z\times$ Cross-cut	.068	-.007	-.079	.147	.010	-.062	.003	-.010	-.034
	(.149)	(.114)	(.113)	(.147)	(.113)	(.113)	(.202)	(.131)	(.106)
$Z\times$ Cross-cut $\times$ Gov	-.022	.073	.153	-.329	-.183	-.064	.124	-.017	-.017
	(.206)	(.167)	(.169)	(.203)	(.163)	(.168)	(.255)	(.182)	(.161)
Lag dependent variable		X			X			X	
School fixed effects			X			X			X
Obs.	8,711	8,154	14,975	8,711	8,154	14,975	8,711	8,154	14,975

See notes for table 4. Columns 1, 4 and 7 use only the 2011 (follow-up) test data. Z represents an indicator variable for random assignment to any treatment arm; *Gov* is an indicator variable for the Ministry of Education treatment arm; SMC is an indicator variable for the SMC training 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: Compliance with the intervention protocol

	Control	Government	NGO
Schools that (ever) employed a teacher	0	56	55
Months of teacher	0	11.59	13
Months of teacher (conditional on employing a teacher)	0	13.25	15.13
Avg. months salary delay	0	2.33	NA
Avg. maximum months of salary delay	0	5.56	NA
Turnover (conditional on ever employing a teacher)	0	0.71	0.43
No of. obs	64	64	64

Table 7: Heterogeneous treatment effects

	(1)	(2)	(3)	(4)
$Z\times$ Gov	.045	-.017	-.017	.011
	(.089)	(.089)	(.089)	(.085)
$Z\times$ NGO	.223	.193	.181	.173
	(.086)***	(.084)**	(.084)**	(.084)**
$Z\times$ Gov $\times$ Density	-.039			
	(.068)			
$Z\times$ NGO $\times$ Density	-.043			
	(.057)			
$Z\times$ Gov $\times$ PTR		-.046		
		(.054)		
$Z\times$ NGO $\times$ PTR		.088		
		(.056)		
$Z\times$ Gov $\times$ KCPE			-.033	
			(.057)	
$Z\times$ NGO $\times$ KCPE			.046	
			(.054)	
$Z\times$ Gov $\times$ Y_1				-.185
				(.060)***
$Z\times$ NGO $\times$ Y_1				-.101
				(.055)*
Obs.	14,475	14,975	14,975	14,418

See notes for table 4. All equations include school fixed effects and standard errors are clustered at the school level.

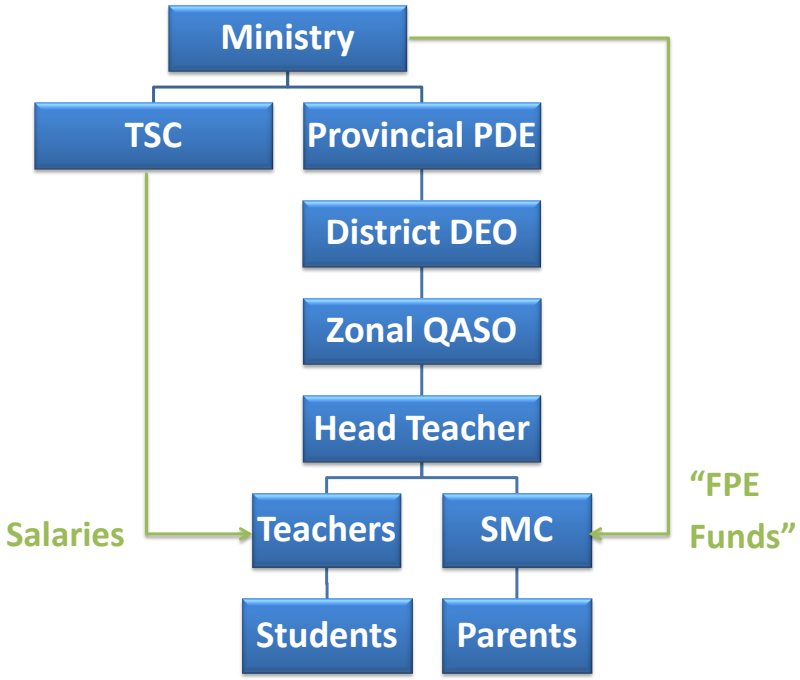

Figure 1: Ministry organization chart and resource flows to public schools.

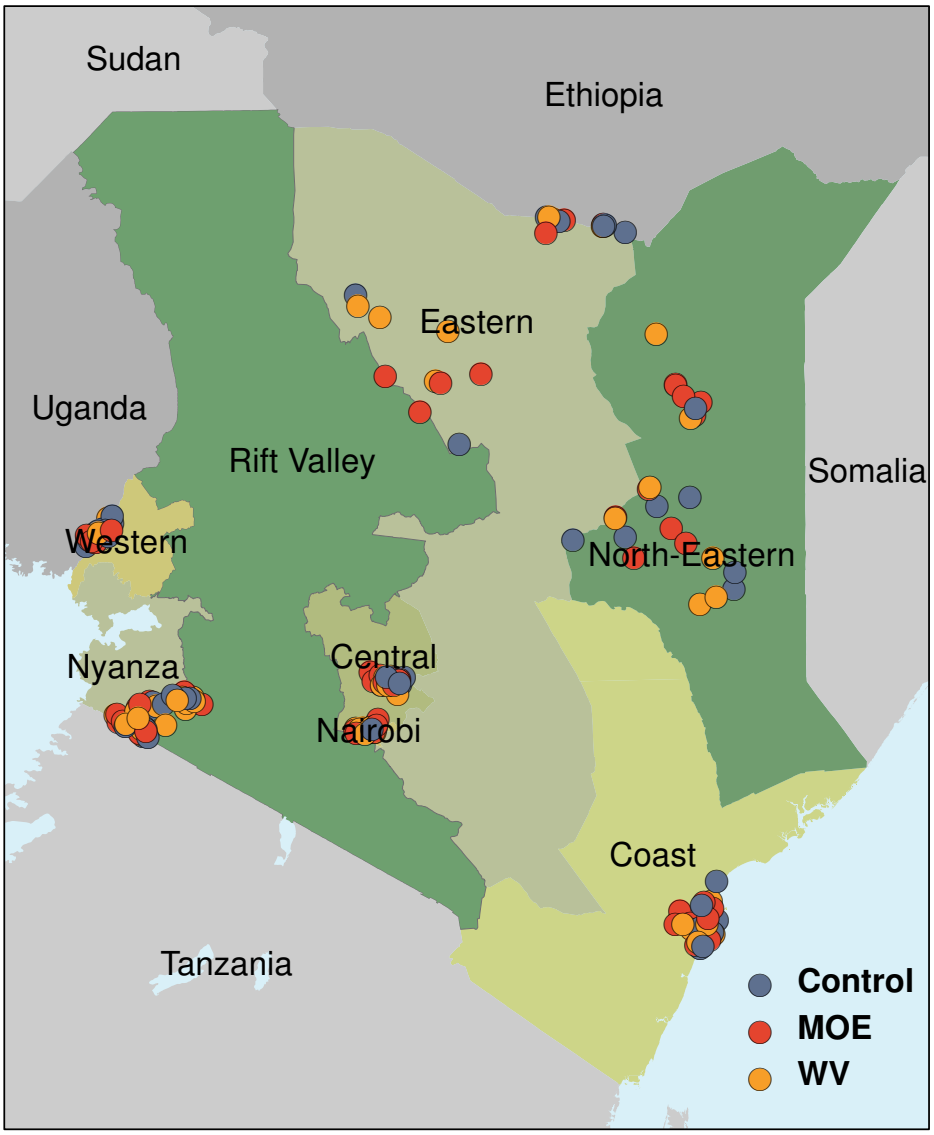

Figure 2: Treatment & control sites across Kenya's 8 provinces. (MOE and WV denote implementation by the Ministry of Education and World Vision, respectively.)

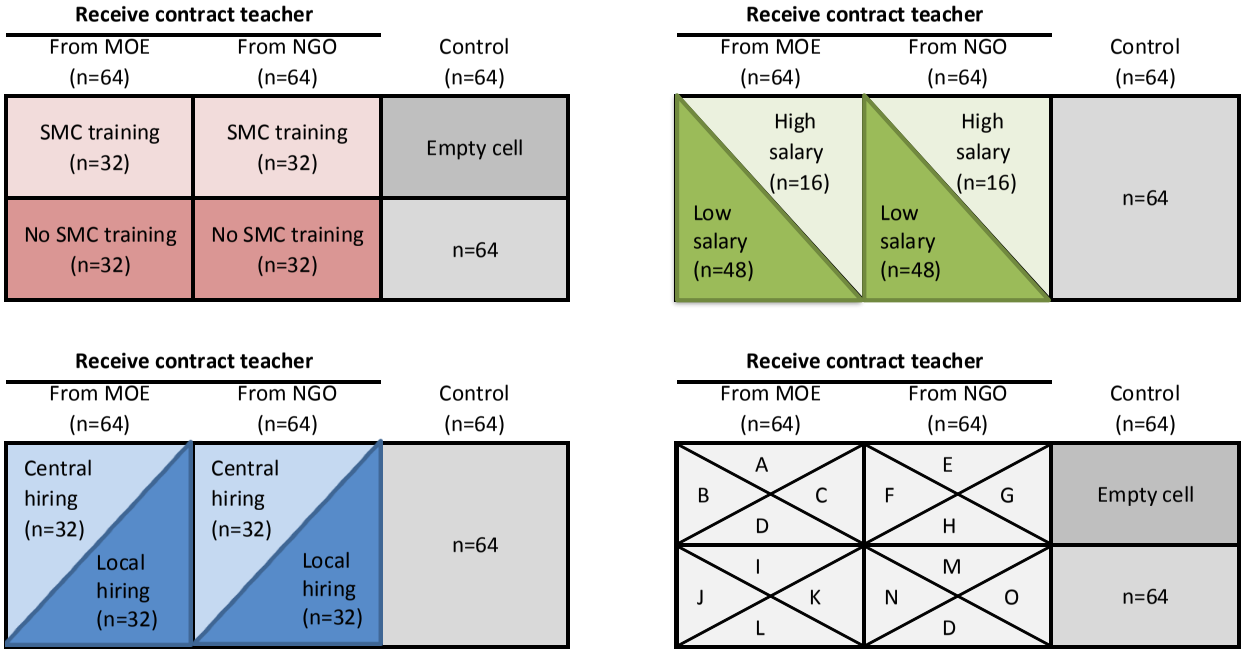

Figure 3: Treatment variations

A: MOE implementation, SMC training, central hiring, high salary. B: MOE implementation, SMC training, central hiring, low salary. C: MOE implementation, SMC training, local hiring, high salary. D: MOE implementation, SMC training, local hiring, low salary. E: NGO implementation, SMC training, central hiring, high salary. Etc.

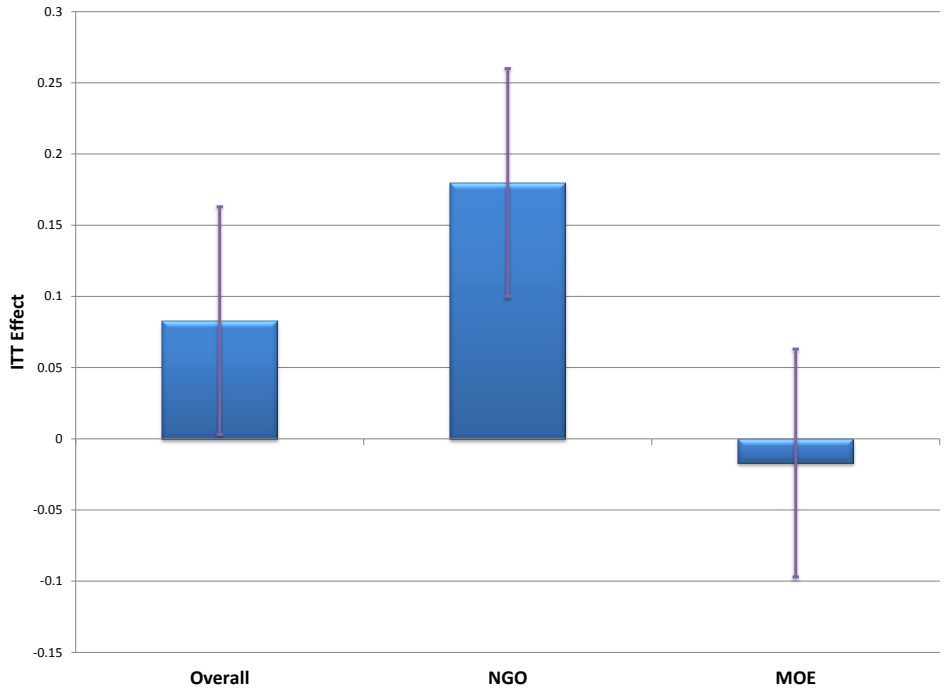

Figure 4: Intention-to-treat effects based on Column 2 of Table. Error bars show the mean effect plus or minus one standard error. 4.

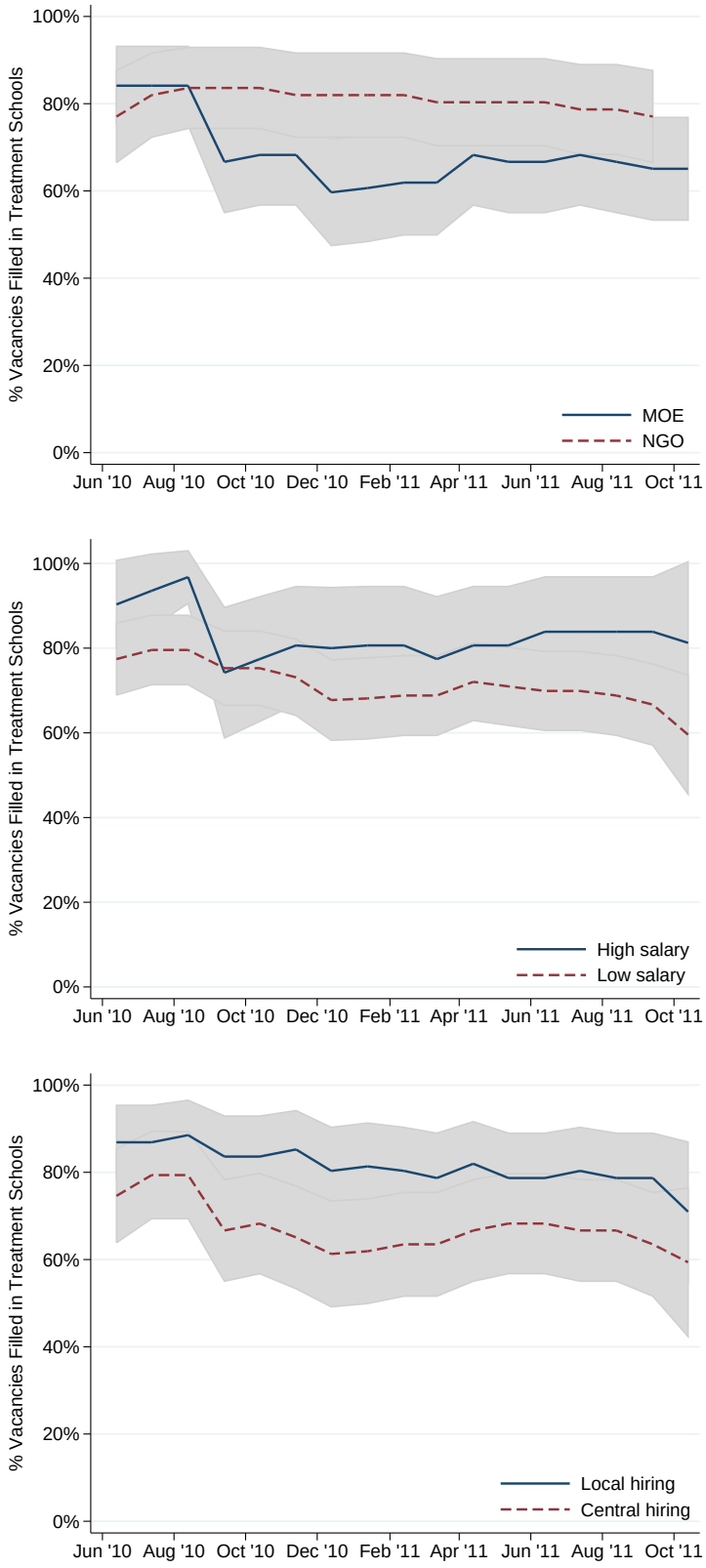

Figure 5: Proportion of contract teacher vacancies filled during evaluation

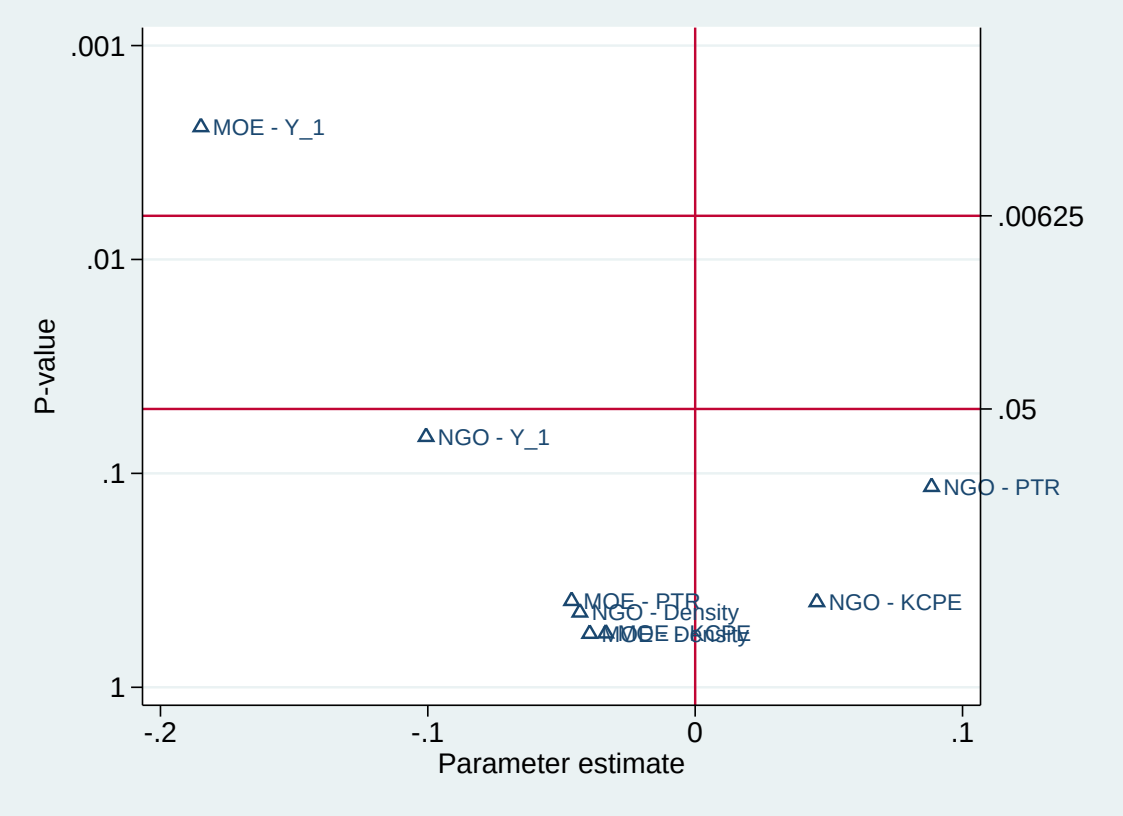

Figure 6: Heterogeneous treatment effects with p-values corrected for multiple comparisons. Each point represents a coefficient reported in Table 7. Points above the lower horizontal line are statistically significant when considered in isolation; points above the upper horizontal line remain significant with correct p-values.