Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania

Isaac Mbiti* Karthik Muralidharan† Mauricio Romero† Youdi Schipper‡
Constantine Manda‡ Rakesh Rajani‡

Abstract

Recent learning assessments have documented the low skill levels attained by pupils in Tanzanian schools. These low levels of learning are driven in part by limited accountability in the education system, which is reflected in the frequent absence of teachers from classrooms. This is further compounded by the resource constraints that schools face. In this study we conduct a randomized experiment to examine the effectiveness of increasing resources to schools relative to increasing teacher incentives, and the complementarity between teacher incentives and school resources. Specifically, we compare the student learning outcomes between schools that were randomly assigned to one of four different interventions: one in which we provide schools with extra resources through capitation (or per pupil) grants paid directly to the school bank account, one in which we provide teachers with a bonus based on the performance of their students on an externally administered exam, one in which schools received both programs, and the control group which received no support. Overall, we find that solely providing resources to schools does not improve learning outcomes. We also find that the teacher incentives did not significantly improve learning outcomes. However, we find learning outomces did significantly improve when teacher incentives were coupled with extra school resources.

1 Introduction

Over the past decade, developing countries have made significant investments aimed at increasing access at primary education. By employing a variety of programs such as conditional cash transfers, school feeding

*University of Virginia
†University of California - San Diego.
‡Twaweza

programs, school based health programs, and fee reductions the net enrollment rates in developing countries have increased dramatically over the past two decades. East African countries have seen sharp increases in their primary enrollment rates, in part due to their free primary education programs (Lucas and Mbiti (2012) in Kenya, Grogan (2009) in Uganda, and Valente (2015) in Tanzania). In Tanzania, the setting of this study, the net enrollment rate in primary school rose from 53% in 2000 to close to 90% in 2012 (World Bank, 2014). Yet despite Tanzania's progress toward achieving near universal primary school access, there is a growing concern that the quality of education may have been compromised. Recent independent nationwide schooling assessments highlighted the low levels of learning in Tanzanian schools, where less than one third of third graders could demonstrate competency in second grade numeracy or literacy (Uwezo, 2013).

The typical policy response has emphasized the need to alleviate resource constraints in schools often through capitation (or per-pupil block) grants to schools. While it is true that average per capita receipts at school level are less than 25 percent of policy mandated amounts (Twaweza, 2013), a large body of evidence suggests that increased monetary resources do not improve student learning outcomes (see McEwan (2015); Murnane and Ganimian (2014); Kremer, Brannen, and Glennerster (2013)). These systematic reviews further suggest that that most effective education interventions are ones that change students' classroom experience. Teacher incentives (or performance pay) could change the classroom experience, especially in low accountability settings such as Tanzania (Murnane & Ganimian, 2014). These low levels of accountability are reflected in the high levels of teacher absenteeism from classrooms, with public school teachers typically spending less than 50 percent of official instruction time in class (Uwezo, 2012; World Bank, 2012; Centre de Recherche Economique et Sociale, 2013). While papers such as Muralidharan and Sundararaman (2011) show that teacher incentives can improve learning outcomes, other sutdies, such as Fryer (2013) or Glewwe, Ilias, and Kremer (2010) find limited evidence of the efficacy of such programs. The differences in the effectiveness of teacher incentives is partly driven by the markedly different incentive designs employed in these studies. Muralidharan and Sundararaman (2011) evaluated a teacher performance pay program that rewarded teachers on the basis of student value added, whereas the programs in Fryer (2013) and Glewwe et al. (2010) rewarded teachers if their students demonstrated a certain pre-determined competency level. A drawback of rewarding teachers on the basis of their students' achieving a competency level (e.g. passing a test) is that teachers may not direct as much effort towards students who are far away from the threshold.[1] This may imply that the best students and the worst students are not well served by such incentive designs. On the other hand, these types of incentive programs are more commonly implemented as they are simpler

[1]No Child Left Behind is an example of a program that introduces threshold effects due to its focus on students achieving a certain qualification mark.

to understand, implement and scale-up, whereas programs that reward teachers for value-added are much harder to understand and scale-up as they require a complex student test-score tracking system.

Given the limited capacities of countries such as Tanzania to scale up more complex (and "more optimal") incentive programs, simple incentive programs that use a threshold design could potentially be a cost-effective approach to raise the productivity of the education sector. Such a program could encourage teachers to increase their efforts, which could result in decreased absenteeism and increased in time spent teaching. Moreover, such a program could encourage teachers, head-teachers and other administrators to use their available resources more effectively. For example research bySabarwal, Evans, and Marshak (2014) showed that textbooks are generally kept locked up in store rooms, rather than in active use by students in class. An incentive program could potentially encourage teachers to more effectively utilize textbooks by incorporating them in their daily lessons and even allowing the children to take them home. However, there is very limited evidence on the complementarities of inputs (or resources) and incentives. Given that programs (or interventions) occur in the context of an education system, a better understanding of the complementarities between proposed programs (or interventions) is extremely important for policymakers. Moreover, in a setting where providing school inputs is an important part of the policy makers' reality, it is important for an RCT to generate evidence on intervention arms that is relevant for business as usual (school grants) and provides potentially effective alternatives (bonus). We use a randomized control trial to compare the effectiveness of alleviating (monetary) resource constraints to the effectiveness of introducing teacher performance pay in public primary schools. The RCT is designed to further examine the complementarities between inputs and incentives. We sample a nationally representative set of 350 schools across 10 districts in Mainland Tanzania. We randomly allocate our set of 350 schools to four groups: Group 1 schools receive Capitation Grants (per pupil grants), Group 2 schools receive a simple performance pay program, Group 3 schools receive both grants and incentives, and Group 4 schools are our control schools.

Although primary schools in Tanzania teach several subjects from grade 1 to grade 7, our study only focuses on student performance in Math, Kiswahili and English in grades 1, 2 and 3. Consistent with previous studies, we find that merely increasing resources in schools does not lead to improved learning outcomes, even though the capitation grant program nearly doubled (non-teacher) spending per child in group 1 and group 3 treatment schools. We also find that the incentive program does not yield statistically significant gains in learning, although the coefficients are positive. We do, however, find that the combination of incentives and resources led to a positive and significant increase in learning. Relative to the control group, student test scores in combination schools increased by 0.20 SD. Additional statistical tests show that the treatment

effect in combination schools is larger than the sum of the estimated treatment effects for capitation grant schools and incentive schools. We argue that this suggests that there are complementarities (or synergies) between incentives and resources. Our findings illustrate the importance of designing RCTs to have sufficient power to detect complementarities. They also highlight the danger of ignoring complementarities in RCTs with multiple arms, especially RCTs with cross- cutting designs, as this may yield misleading findings.

2 Theoretical framework

In this section we propose a simple model of how teachers choose effort. The main goal of the model is to formalize the following intuition: If inputs (books, charts, etc.) and teacher effort are complements, then an increase in inputs increases effort, provided that effort is rewarded. If on the other hand, inputs and teacher effort are substitutes, is unclear whether an increase in inputs will increase effort. Specifically, if inputs decrease the marginal return of effort, then is possible that more inputs lead to lower teacher effort.

We assume that teachers choose how much effort to provide, e, and solve the following problem:

$$\max_{e} W + \lambda L - c(e)$$

s.t.

$$W = W_b + tL$$

$$L = f(e, I)$$

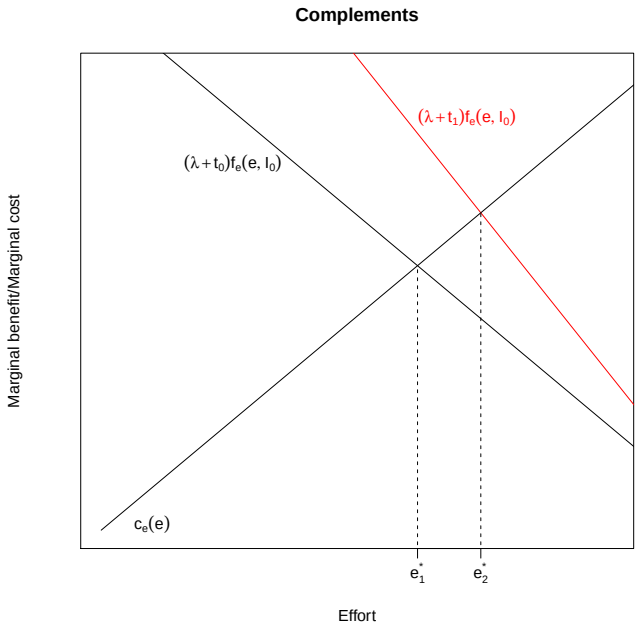
where W is their total salary which is equal to a base pay (W_b) plus a bonus (tL) that is proportional to the level of student learning L (for each "unit of learning" the teacher earns t). Teacher effort, together with other inputs (I), translates into learning via f, which is strictly increasing on both arguments ($f_e > 0$ and $f_I > 0$), and concave. Teachers earn direct utility from students learning via λL, and effort causes them some cost, c, which is increasing and convex ($c'(\cdot) > 0$ and $c''(\cdot) > 0$).

Notice that the FOC implies that the optimal level of effort (e^*) satisfies:

$$\underbrace{(t + \lambda) f_e(e^*, I)}_{\text{Marginal benefit}} = \underbrace{c_e(e^*)}_{\text{Marginal cost}}$$

If t increases, then the marginal benefit of effort increases for all levels of effort and therefore e^* increases (see Figure 1), and so do learning levels (L^*).

Figure 1: The effect of an increase in incentives on teacher effort



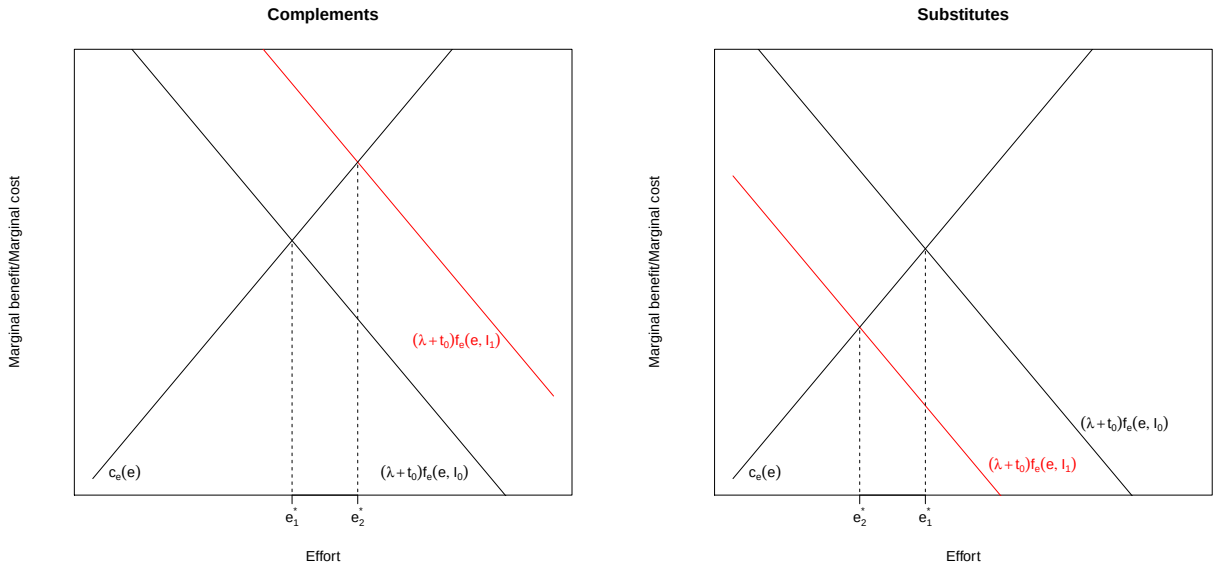
This figure shows how optimal effort (e^*) changes when incentives increase from t_0 to t_1. Notice that this figure is valid whether $f_{e,I}(e,I) > 0$ or $f_{e,I}(e,I) < 0$.

If I increases, depending on whether inputs and effort are complements ($f_{e,I}(e,I) > 0$) or substitutes ($f_{e,I}(e,I) < 0$), the marginal benefit of effort can either increase or decrease for all values of e. If $f_e(e^*, I)$ is increasing in I (i.e., inputs make effort more effective), then an increase in I increases e^*. If $f_e(e^*, I)$ is decreasing in I (i.e., inputs make effort less effective, which can happen if inputs substitute for teacher time), then an increase in I decreases e^* (see Figure 2). In this case is also ambiguous whether learning levels (L^*) increase with an increase in I since:

$$\frac{\partial L^*}{\partial I} = \underbrace{f_e(e^*, I)}_{>0} \underbrace{\frac{\partial e^*}{\partial I}}_{\gtrless 0} + \underbrace{f_I(e^*, I)}_{>0} \tag{1}$$

Note that is still possible for learning levels to increase even if $\frac{\partial e^*}{\partial I} < 0$. Thus, an increase in learning levels is compatible with effort and inputs being complements ($f_{e,I}(e,I) > 0$) or substitutes ($f_{e,I}(e,I) < 0$). Thus, if we had evidence that learning levels increase with an increase in inputs, this would not provide any insights on the relationship between between inputs and effort.

Figure 2: The effect of an increase in inputs on teacher effort



Both panels show how the optimal effort (e^*) changes when inputs increase from I_0 to I_1. The left panel shows the case when $f_{e,I}(e^*, I) > 0$ and the right panel the case when $f_{e,I}(e^*, I) < 0$.

If both I and t increase is unclear whether optimal effort increases. If $f_{e,I}(e^*, I) > 0$ then effort increases with an increase in inputs and incentives. This is intuitive, as increases in t always increase effor,t and increases in I also increase effort in this case. However, if $f_{e,I}(e^*, I) < 0$ then is unclear whether optimal effort will increase and will depende heavily on the exact shape of f, the change in t and the change in I (as well as λ). In this case is also ambiguous whether learning levels (L^*) increase with an increase in both I and t since:

$$\frac{\partial^2 L^*}{\partial t \partial I} = \frac{\partial e^*}{\partial t}\left[f_{ee}(e^*, I)\frac{\partial e^*}{\partial I} + f_{eI}(e^*, I)\right] + f_e(e^*, I)\frac{\partial^2 e^*}{\partial I \partial t} \tag{2}$$

$$= f_{eI}(e^*, I)\underbrace{\frac{\partial e^*}{\partial t}\frac{c_{ee}(e^*)}{c_{ee}(e^*) - f_{ee}(e^*, I)(t+\lambda)}}_{>0} + \underbrace{f_e(e^*, I)}_{>0}\frac{\partial^2 e^*}{\partial I \partial t} \tag{3}$$

If $f_{e,I}(e^*, I) > 0$ then $\frac{\partial^2 e^*}{\partial I \partial t} > 0$ and therefore $\frac{\partial^2 L^*}{\partial t \partial I} > 0$. If $f_{e,I}(e^*, I) < 0$ then learning levels can either increase or decrease. Notice that $\frac{\partial^2 e^*}{\partial I \partial t} < 0$ is a necessary, but not sufficient condition, for $\frac{\partial^2 L^*}{\partial t \partial I}$ to be negative. Thus, is possible for learning levels to increase with an increase in both inputs (I) and teacher incentives (t), even if effort and input are substitutes ($f_{e,I}(e, I) < 0$).

From a policy point of view the relevant parameters are $\frac{\partial^2 L^*}{\partial t \partial I}$, $\frac{\partial L^*}{\partial t}$, and $\frac{\partial L^*}{\partial L}$, and these parameters can be recovered from an experiment such as ours. However, if one is interested in understanding the production function of learning (f), even using in a very simple theoretical framework such as ours, this is not always possible. One can infer that $\frac{\partial^2 e^*}{\partial I \partial t} < 0$ if $\frac{\partial^2 L^*}{\partial t \partial I} < 0$. However, if $\frac{\partial^2 L^*}{\partial t \partial I} > 0$, is possible that teacher effort and inputs are substitutes (i.e., $f_{e,I}(e^*, I) < 0$) or complements (i.e., $f_{e,I}(e^*, I) > 0$) and teacher effort itself could be increasing or decreasing.

3 Experimental design

3.1 Context

Tanzania has dramatically expanded primary schooling in recent years, achieving close to universal enrollment. The net enrollment rate in primary school rose from 53% in 2000 to close to 90% in 2013 (with a historical high of 97% in 2008). This increase in enrollment followed an abolition of primary school fees in 2001. The previously collected fees were replaced with a capitation grant (or per-pupil block grant to schools) that would support the operation of schools.[2] The official policy stipulated that schools should receive approximately US$10 per enrolled student; this was later adjusted downward to TZS 10,000 (or around US$6.25). The policy also provided spending guidelines where 40% of these grants could be spent on textbooks, teaching guides and other reading materials, administrative expenses, 20% could be spent on chalk, exercise books, pens and pencils, 20% on minor construction and repair, 10% on examinations and test paper printing and 10% on administration. As teachers were assigned to schools and paid directly by the government, schools could not use capitation grant funds to pay teachers or hire teachers.

Despite official policy, the Tanzanian government has generally failed to provide that level of support to schools. From 2006-2010, schools were only allocated about TZS 6,000 (US$3.75) per student by government (?, ?). School finances were further constrained by significant leakages of the budgeted capitation funds. A recent World Bank study estimated that about 37% of budgeted capitation grant funds did not reach the school (World Bank, 2012). More recent survey evidence finds CG receipts per capita of only 20 percent of the official amount, with 34 percent of schools not receiving any CG funds(Twaweza, 2013). This leakage is in part driven by the disbursement system, where funds are first transferred to regional and local authorities who are then responsible for transferring the funds to schools.

[2]This change to per-pupil grants creates a an incentive for mis-reporting. There is some evidence that official figures ovserstate enrollment after funding shifted from user fees to per pupil government grants (Sandefur & Glassman, 2015).

As a result of the limited finances available to schools, the educational inputs and infrastructure at schools is often inadequate. The World Bank Service Delivery Indicators show that only 3% of schools have sufficient infrastructure (potable water, sanitation, and electricity) and 5 children (in grades 1, 2, and 3) shared a math textbook, while 2.5 children shared a reading book (World Bank, 2012). The increase in enrollment spurred by the introduction of the free primary school program has resulted in large class sizes and large pupil to teacher ratios (Valente, 2015). Class sizes in primary schools average 74 students, with almost 50 students per teacher (World Bank, 2012).

In addition to large class sizes and limited resources, there is also limited accountability in Tanzanian public primary schools. Teacher absence rates are high. Almost one in four teachers is absent from school on a given day (World Bank, 2012). Teacher effort also seems low even among those that are present in school: Over 50% of teachers who were present in school were absent from the classroom (World Bank, 2012) and children receive on average about 2 hours of instruction per day. These low levels of effort are mirrored by low self-reported motivation: 47% of teachers say that, if they could start over, they would not choose teaching as a career. Education analysts confirm that the teaching profession has an image problem and that teacher training colleges do not attract the brightest and most committed students.[3]

Despite indisputable accomplishments in getting children to schools, and perhaps unsurprisingly given the low quality of inputs and teacher effort, children often fail to attain proficiency in early grade reading and numeracy. Evidence from a variety of sources reports that Tanzanian learning outcomes are abysmal. In 2012, both the PSLE (Primary School Leaving Examination) and the CSEE (Certificate of Secondary Education Exam) had historically low pass rates of 31 percent and 34 percent, respectively. In parallel, the annual nationwide learning assessments carried out by Twaweza/Uwezo consistently show that less than one third of grade 3 students can read a simple story at a grade 2 level in Kiswahili (the national language and language of instruction) or successfully demonstrate grade 2 numerical skills. Performance in English was especially weak, with less than 12% of grade 3 students able to read at a grade 2 level in English (Uwezo, 2013; Jones, Schipper, Ruto, & Rajani, 2014).

Although teacher motivation as a policy term is not new in Tanzania, this usually refers to improvements in teacher welfare such as better housing or base salary increases. Teacher performance pay that links teacher payments to a performance indicator such as learning outcomes is a new policy idea, but not without some currency within government. Through the Big Results Now initiative, the government is piloting programs that reward teachers and schools that can deliver better learning outcomes. This policy shift is notable as

[3] A common saying in Tanzania is: Did you fail to get a job, even teaching ? (Umekosa ajira nyingine zote, hata ualimu)

it marks the first time that the Tanzanian education system is focusing on learning outcomes rather than educational inputs

3.2 Sampling and design

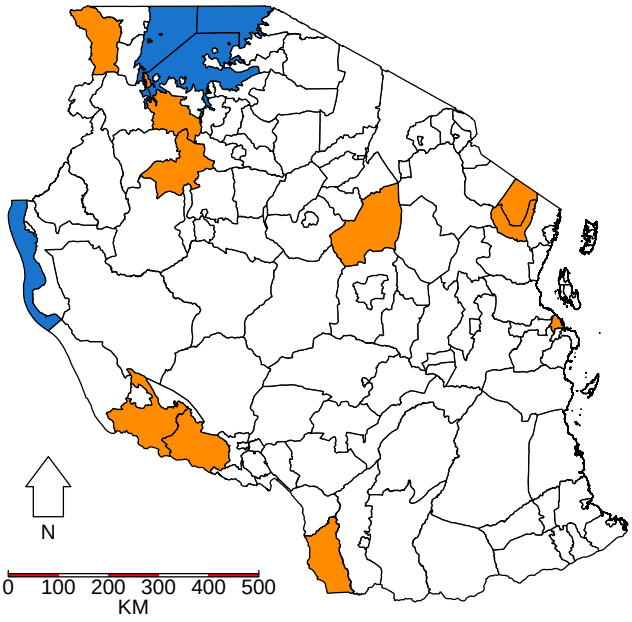
The KiuFunza project (as the RCT was known) featured 3 treatment arms and a control group and was implemented across a representative sample of 350 schools across 10 districts in Tanzania (see Figure 3). The treatment arms were:

1. A capitation grant (CG) to schools that provides them with block grants (the "input" treatment)
2. A "cash on delivery" (CoD) treatment to schools that provides teachers and head teachers with bonus payments conditional on the number of students who pass basic literacy and numeracy tests (the "incentive" treatment)
3. A combination (Combo) treatment arm where schools were provided with both the CG and the CoD treatments.

Each of the three treatments was assigned to 70 randomly-selected schools and an additional 140 schools served as a control group. Specifically, we randomly sampled 10 districts in Tanzania, and within each district we randomly sampled 35 schools. From each district 7 schools were randomly assigned to receive capitation grants (CG), 7 schools to receive teacher incentives ("Cash on Delivery" or COD), and 7 schools to receive both grants and incentives (Combo). The additional 14 schools did not receive anything and serve as the control group. The double sized control group increases our power to detect complementarities between CG and COD.

The intervention was managed by Twaweza, a Tanzanian NGO that focuses on citizen agency and public service delivery. Intervention schools were informed that the program would last for two years (2013 and 2014). Schools were informed about the program through community meetings. Program materials such as flyers and cartoon booklets were distributed to teachers and to students who were instructed to share the materials with their parents. Twaweza also worked closely with both central and regional government officials to ensure that there would be minimal interference with the program and the research efforts. All students in Grades 1, 2, and 3 in schools that received Cash on Delivery were tested in Kiswahili, English and Math at the end of the school year to determine teacher incentive payments. Tanzanian education professionals, following a similar structure as the Uwezo annual learning assessment, developed the subject tests for Grades 1, 2, and 3. The same schedule will be followed in 2014.

Figure 3: Districts in Tanzania from which schools are selected



3.2.1 Capitation Grant (CG) Arm

Schools assigned to receive the capitation grant were provided with the grants that mirrored the official Tanzanian capitation grant policy. Thus each school received TZS 10,000 ($\sim$ US$5) per student from Twaweza. Schools receiving these funds had to spend and account for the funds as outlined in the official policy (described above). As discussed above, capitation grant leakages were a major challenge faced by schools. In addition the timing of government grant disbursements was very unpredictable making financial planning by schools extremely difficult (Twaweza, 2013). In order to minimize leakage and enable better financial planning, Twaweza grants were transferred directly into school bank accounts in two tranches, the first at the beginning of the second term (around April) and the second at the beginning of the third term (around August/September).

Following the regular capitation grant policy, schools were also required to share revenue and expenditure information with the community and display summary financial statements in a public area in the school

(usually on a notice board in the school). On aggregate, approximately US$700,000 was disbursed to these schools each year in 2013 and 2014 (this includes schools in the CG and Combination arm). The size of the grants distributed to schools is approximately three times the pre-treatment per student expenditure (excluding teacher salaries). The CG treatment thus reflects a significant increase in the financial resources available to schools. The successful implementation of this intervention helped the government to review its capitation grant policy. The Government of Tanzania already committed in public to adopting the "direct CG" policy in 2014. The new Government of president John Magufuli, elected in October 2015, has started to implement the transfer of capitation grant funds directly to schools in January 2016.

3.2.2 Cash on Delivery (COD) Arm

The teacher performance pay program provided monetary bonuses to teachers contingent on the performance of their students. Given Twaweza's emphasis on early grade learning, the program was limited to teachers in grades 1, 2, and 3 and focused on numeracy (Mathematics) and literacy in English and Kiswahili. For each subject, eligible teachers earned a TZS 5,000 ($\approx$ \$ USD 3) for every student that passed a simple, externally administered, grade-appropriate assessment (based on the curriculum). Note that this payment was for absolute levels of learning, not gains in learning and that teachers were not penalized for students who did not pass. Additionally, the head teacher was paid TZS 1,000 ($\approx$ \$ USD 0.6) per subject each child passes. Notice that if a teacher taught all three subjects in a given grade, they received TZS 15,000 ($\sim$ US$9) for every student that passed all three assessments.

The program was announced to teachers in March of 2013. During an intervention baseline visit the details of the bonus were explained to the head teacher and focal subject and grade teachers. Flyers with a description of the bonus structure and frequently asked questions were handed out to teachers, and a booklet explaining the goals of the program and the visits were handed out to parents. A follow up visit in July 2013 reinforced the details of the program and provided an opportunity for questions and feedback. The high-stakes assessments were scheduled for the end of the school year. Understanding of the program was high as over 90% of teachers in the program could correctly calculate the bonus level in a hypothetical scenario. The simplicity of the incentive program enabled teachers to understand the program. The simple "threshold design" utilized here was also simpler to implement and is arguably more scalable in resource (both human and financial) constrained settings such as Tanzania. However, threshold designs are not optimal incentive designs as they can encourage teachers to focus their attention on students close to the passing threshold. Thus students who are far below the passing threshold and those who are far above the threshold would not

be well served by such an incentive design. Despite the limitation of this design it is important to note that it is routinely used in programs such as No Child Left Behind (Neal & Schanzenbach, 2010).

In order to ensure the fidelity of the implementation, we created several versions of the high stakes end of year test. We also took student photos at enrollment to prevent identity fraud. We also conducted the end of year testing in a sample of control schools and also conducted a low stakes audit test (or research test) on a sample of 30 students (10 students in each grade) in all 350 schools in our sample. Teachers (and head teachers) earned about $150,000 in bonuses each year (this includes teachers in the COD only and Combo arm). These bonuses were paid directly into teacher bank accounts or through mobile money transfers.

3.2.3 Combination arm

Schools assigned to the combination arm received both capitation grants and teacher incentives. As discussed earlier we increased the size of the control group in order to increase our ability to detect complementarities between incentives and resources.

3.3 Data, balance and attrition

From each school, we sample 10 students from each grade focal grade (grade 1, 2 and 3) to create a student panel of 10,500 students who we follow over the course of this study. This provides us with a panel of student test scores. From this set of students, we randomly sample 3500 students to conduct household surveys. These surveys collect information on educational expenditures, general household characteristics and non-financial educational inputs at the household (such as helping with homework). We survey all teachers (about 1500) who teach focal grade (grade 1, 2, 3) and focal subjects (Math, English and Swahili). We measure teacher effort, teacher time use and teaching strategies . We also obtain teacher opinions on incentive programs. We further conduct 350 head teacher/school level surveys. We collect information about the school facilities, the teaching roster, input availability and expenditures. Unlike most school inputs, textbooks are easily assignable to grades and subjects. We therefore collect information on textbook purchases at the grade subject level. In addition to this survey data, we also utilize administrative data from the program, including test scores from the Twaweza test that is used to assess students for the COD program.

It is important to note there are two sets of tests performed to measure student learning levels. The intervention test is taken by all students in grades 1, 2 and 3 in COD and Combo schools to calculate teacher payments. Additionally, we tests 30 students in all schools which allows us to estimate treatment effects

using a low stakes exam. The intervention (high-stakes) test carried out by Twaweza is used to calculate the incentive payments, but the impact of the programs is estimated using using the low-stakes exam.

The baseline was conducted in early 2013, followed by a midline in August 2013 and an end line in October 2013. The pay for performance testing occurred in November 2013. A similar calendar was followed in 2014 (see Figure 4). Prior to analysis the data, we filled a pre-analysis plan on the AEA registry.

Figure 4

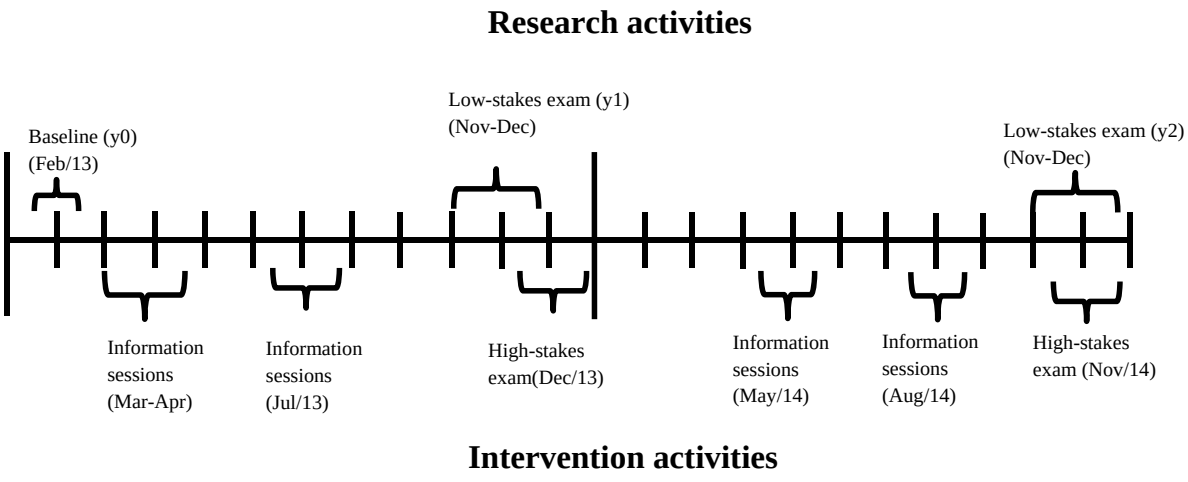


Table 1 shows that observable characteristics of students, households, schools, and teacher are balanced across our treatment arms. About half of the students in our sample are male, and the average age of students (in 2013) is 8 years. We normalize test scores relative to the mean in the control group in each grade- subject and find no difference at baseline. Households spent about US$8 on the focal child's education and the majority of these households had floors made of earth/mud (rather than cement or more durable materials). Panel C shows that schools had limited infrastructure and had large enrollments with on average 50 students per teacher. Less than 10% of schools tracked students and 40% of schools had multiple shifts, where half the grade would attend school in the morning and the other half would attend from the late morning. Schools (and households) were mostly rural. Panel D shows that about 60% of teachers in our sample were female and had about 15 years of experience overall and 7 years of experience, with nearly 40% of them not having a teaching certificate.

Table 1: Summary statistics across treatment groups

	Combo	CG	COD	Control	p-value (all equal)
Panel A: Students					
Male	0.50	0.49	0.50	0.50	0.99
	(0.0095)	(0.010)	(0.0085)	(0.0075)	
Age	8.94	8.96	8.94	8.97	0.96
	(0.050)	(0.053)	(0.047)	(0.039)	
Swahili test score	0.055	-0.019	0.065	0.00014	0.78
	(0.067)	(0.067)	(0.082)	(0.048)	
Math test score	0.070	0.010	0.059	0.00016	0.77
	(0.061)	(0.064)	(0.072)	(0.045)	
English test score	-0.017	-0.011	0.020	0.00010	0.97
	(0.047)	(0.049)	(0.063)	(0.037)	
Panel B: Households					
Household size	6.247	6.332	6.436	6.352	0.805
	(0.134)	(0.141)	(0.137)	(0.0940)	
Asset Index (PCA)	-0.0144	-0.0585	0.0259	0.0234	0.927
	(0.101)	(0.106)	(0.115)	(0.0778)	
Expenditure in education (2013)	11371.7	10382.4	11876.8	13218.6	0.208
	(1053.8)	(975.5)	(1236.8)	(935.8)	
Floor made out of earth/mud	0.657	0.659	0.660	0.668	0.994
	(0.0365)	(0.0367)	(0.0367)	(0.0249)	
Panel C: Schools					
Infrastructure Index (PCA)	-0.10	0.061	-0.086	0.065	0.65
	(0.13)	(0.14)	(0.16)	(0.084)	
Electricity	0.13	0.14	0.13	0.14	0.99
	(0.040)	(0.042)	(0.040)	(0.029)	
Single shift	0.60	0.59	0.64	0.63	0.89
	(0.059)	(0.059)	(0.058)	(0.041)	
Students/Teachers	54.8	58.8	55.5	60.2	0.56
	(2.63)	(3.09)	(2.53)	(3.75)	
Track students	0.071	0.10	0.071	0.093	0.88
	(0.031)	(0.036)	(0.031)	(0.025)	
Urban	0.16	0.13	0.17	0.15	0.91
	(0.044)	(0.040)	(0.045)	(0.030)	
Enrolled students	739.1	747.6	748.5	712.4	0.89
	(48.4)	(51.9)	(51.7)	(30.4)	
Panel D: Teachers					
Male	0.37	0.32	0.34	0.35	0.89
	(0.040)	(0.041)	(0.035)	(0.026)	
In what year were you born?	1975.0	1975.3	1975.0	1975.5	0.87
	(0.69)	(0.81)	(0.67)	(0.45)	
In what year did you start teaching?	1999.0	1999.0	1998.8	1999.4	0.91
	(0.72)	(0.84)	(0.70)	(0.45)	
Travel time from house to school	20.4	17.3	21.4	20.4	0.57
	(1.99)	(2.20)	(2.28)	(1.61)	
Teaching Certificate	0.75	0.74	0.75	0.73	0.95
	(0.030)	(0.029)	(0.027)	(0.017)	

This tables presents the mean and standard error of the mean (in parenthesis) for several characteristics of students (Panel A), households (Panel B), schools (Panel C) and teachers (Panel D) across treatment groups. Column 4 shows the p-value from testing whether the mean is equal across all treatment groups (H_0 := mean is equal across groups). The household asset index is the first component from a Principal Component Analysis of the following assets: Mobile phone, watch/clock, refrigerator, motorbike, car, bicycle, television and radio. The school infrastructure index is the first component from a Principal Component Analysis of indicator variables for: Outer wall, staff room, playground, library, and kitchen. Standard errors are clustered at the school level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2 shows that attrition is balanced across treatment arms and low (we were able to track around 90% of students both years).

Table 2: Student attrition

	Yr 1	Yr 2
CG	0.0054	0.0052
	(0.012)	(0.0096)
COD	0.022*	0.00038
	(0.011)	(0.011)
Combo	0.00025	0.0029
	(0.012)	(0.010)
N. of obs.	10496	10496
Mean control	0.87	0.90
Combo-COD-CG	-0.027	-0.0027
p-value (H_0:Combo-COD-CG=0)	0.12	0.87

The independent variable is whether we were able to test the student at endline of year 1 (Column 1) and at endline of year 2 (Column 2). Standard errors are clustered at the school level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Results

4.1 School expenditure and funding

Table 3 shows how schools receiving the grants spent the funds across broad categories. Overall there are no statistically significant differences in spending behavior between combo schools and regular capitation grant schools. In the first year schools spent about 80% of the grant and saved the remainder. Schools were more restrained in spending the grant in year two. Given the uncertainties of government funding (both in terms of timing and amount), this "precautionary saving" behavior by school is reasonable.

Table 3: How are schools spending the money?

	Combo	CG	Diff
Year 1			
Total	8350.5	8076.0	274.5
	(254.7)	(318.4)	
$ Admin./Student	1995.2	1773.1	222.2
	(139.0)	(148.3)	
$ Student/Student	450.5	622.5	-172.0
	(82.64)	(94.69)	
$ Teaching Aid/Student	5803.9	5620.1	183.7
	(205.9)	(285.0)	
$ Teacher/Student	2.742	0	2.742
	(1.968)	(0)	
$ Construction/Student	98.13	60.35	37.78
	(51.42)	(36.58)	
Year 2			
Total	5609.6	6047.3	-437.6
	(352.1)	(352.6)	
$ Admin./Student	2023.3	2069.7	-46.41
	(167.9)	(199.2)	
$ Student/Student	409.0	456.3	-47.25
	(65.03)	(82.08)	
$ Teaching Aid/Student	3110.0	3448.1	-338.2
	(259.9)	(268.6)	
$ Teacher/Student	0	3.364	-3.364
	(0)	(3.364)	
$ Construction/Student	67.31	69.76	-2.444
	(39.29)	(61.16)	

Mean expenditure per student. *Admin:* administrative cost(including staff wages), rent and utilities, and general maintenance and repairs. *Student:* food, scholarships and utilities (notebooks, pens, etc.) *Teaching aid:* classroom furnishings, textbooks, maps, charts, blackboards, practice exams, etc. *Teachers:* salaries, bonuses and teacher training. Standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To estimate the effect that each intervention had on total school expenditure we estimate the following equation:

$$Y_{sdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 Combo_s + \gamma_t + \gamma_d + X_s\beta_2 + \varepsilon_{sd}, \tag{4}$$

where Y_{sdt} is the outcome of interest in school s in district d at time t. COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, and $Combo_s$ indicates whether the schools received both cash on delivery and a capitation grant. γ_d is a set of district fixed effects, γ_t is a set of time fixed effects, and X_s is a set of

school characteristics at baseline (facilities, teacher per student, and school committee characteristics). The coefficients of interest are the α's. Additionally, we test $\alpha_3 - \alpha_2 - \alpha_1 = 0$ and $\alpha_3 - \alpha_2 = 0$. The first to see if the effect on Combo schools is more than the sum of the parts (i.e., whether the interaction between CG and COD is significant), the second to test whether CG and Combo schools have different expenditure patterns.

Table 4 presents the results from estimating the previous equation using school expenditure, funding by other sources, and household expenditure as the outcomes of interest. Column 1 examines total school expenditures over the two years the program lasted. Spending at Combo schools and CG schools was approximately two times that of control group schools. In general spending patterns between Combo and CG schools were similar. The only (statistically significant) difference in expenditure is in teaching aids (textbooks, charts, maps, blackboards, practice exams, etc.), on which Combo school spend over 1,000 TZS more per student during the first year (see Table 15 in Appendix B). This is consistent with Combo schools spending more of their resources to supports teachers who are eligible for the COD bonus. Additionally, most of these teaching aids are a stock that depreciates slowly over time, and therefore there is no need for schools to invest their resources on these again during the second year. A potential drawback of capitation grant programs that are provided by NGOs is that households, governments and other stakeholders may cut back their support, thereby undermining the program (Das et al., 2013; Pop-Eleches & Urquiola, 2013). Column 2 in Table 4 shows that there were moderate cutbacks in all schools (CG, COD, and Combo) of about TZS 300 each year but this was not statistically significant (See Table16 in Appendix B for details). Finally, Column 3 in Table 4 shows that households in CG schools cut back their expenses by about TZS 2,000, while households in Combo schools and COD schools did not (See Table17 in Appendix B for more details). Overall, the minor reductions in resource support from other support did not fully offset the Twaweza grant, thus the capitation grants did lead to overall increases in educational support and spending.

Table 4: Expenditure and funding

	School Expenditure	Other funding	Household expenditure
CG	4922.4***	-390.8	-2240.3*
	(709.0)	(562.5)	(1321.3)
COD	-613.9	-597.8	659.5
	(534.6)	(533.4)	(1299.6)
Combo	5457.6***	-207.3	269.2
	(709.7)	(659.7)	(1767.1)
N. of obs.	699	699	6709
Mean control	5241.9	5427.5	27012.4
Combo-COD-CG	1149.1	781.3	1850.0
p-value (H_0:Combo-COD-CG=0)	0.20	0.28	0.63

Results for estimating equation 4 for school expenditure, school funding, and household reported expenditure in education. Column (1) has school expenditure as the dependent variable. Column (2) has total funding received by the school (from other sources besides our own transfers). Column (3) has household level data on expenditure in education. All regressions are done including data for both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5 shows how schools' expenditure on textbooks varied across treatments and grades. Focal grades (FG) refers to grades 1, 2 and 3. Swahili, Math and English teachers in focal grades are eligible for bonuses in COD and Combo schools. Two important results stand out. First, overall textbook expenditure increases in CG and Combo, but there is no differential overall expenditure across both treatment arms. Second, textbook expenditure is lower in focal grades, however, Combo schools spend more than CG schools on textbooks for these grades. This is consistent with Combo schools spending more of their capitation grant resources to supports teachers who are eligible for the COD bonus.

Table 5: Expenditure on textbooks

	Textbook expenditure
CG	1720.3***
	(213.9)
COD	-131.8
	(112.3)
Combo	1457.9***
	(200.7)
FG	-359.3***
	(51.3)
CG × FG	-469.7***
	(174.0)
COD × FG	87.6
	(97.7)
Combo × FG	69.6
	(239.6)
N. of obs.	4880
Mean control	696.3
Mean control (FG)	498.7
Combo-COD-CG	-130.6
p-value (H_0=Combo-COD-CG=0)	0.66
Combo-COD-CG (FG)	680.3**
p-value (H_0=Combo-COD-CG (FG)=0)	0.011

Results for estimating equation 4 for textbook expenditure. The regression includes data for both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.2 Test scores

To estimate the effect that each intervention had on students test scores we estimate the following equation:

$$Z_{isdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 Combo_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_s\beta_2 + \varepsilon_{isd}, \quad (5)$$

where Z_{isd} is the test score of student i in school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, γ_d is a set of district fixed effects, γ_w is a set of week fixed effects[4], γ_g is a set of grade fixed effects, X_i is a series of student characteristics (age, gender and grade), X_s is a set of school and teacher characteristics (facilities, teacher per student, school committee characteristics, teacher's age, experience, qualifications, and gender). The coefficients of interest are the α's. As before, we

[4]The survey test was performed before the intervention test, but the timing is balanced across treatment arms. The week fixed effects should the increase the precision of our estimates.

Table 6: Effect on test scores

	Year 1				Year 2			
	Math	Swahili	English	Focal Subjects	Math	Swahili	English	Focal Subjects
CG	-0.051	0.0039	-0.0099	-0.022	0.0040	-0.019	0.0088	-0.0033
	(0.039)	(0.041)	(0.036)	(0.037)	(0.048)	(0.050)	(0.041)	(0.044)
COD	0.036	0.054	0.068*	0.063*	0.062	0.0072	0.028	0.039
	(0.039)	(0.039)	(0.038)	(0.036)	(0.045)	(0.048)	(0.037)	(0.041)
Combo	0.092**	0.12***	0.084**	0.12***	0.19***	0.20***	0.094*	0.19***
	(0.044)	(0.040)	(0.042)	(0.040)	(0.046)	(0.044)	(0.048)	(0.044)
N. of obs.	9142	9142	9142	9142	9439	9439	9439	9439
Combo-COD-CG	0.11*	0.059	0.026	0.076	0.12*	0.21***	0.057	0.16**
p-value (H_0:Combo-COD-CG=0)	0.075	0.33	0.67	0.18	0.074	0.0029	0.37	0.014

Results for estimating equation 5 for different subjects at both follow-ups. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

test $\alpha_3 - \alpha_2 - \alpha_1 = 0$ to see if the effect on Combo schools is more than the sum of the parts (i.e., whether the interaction between CG and COD is significant).

Table 6 below shows the main results of our study. First, capitation grants (CG) do not raise test-scores. This is true across subjects and years. We also see that the performance pay program (COD) is not effective in raising test-scores on its own. However, we see that the combination of both programs significantly raises learning outcomes. This is true across all subjects, however the effect on Kiswahili and Math is twice as large as the effect on English after two years. For robustness and to mitigate concerns of multiple testing, we create summary index of focal subjects. [5] We see that the combination arm increases test scores by 0.17 SD in the first year, and over 0.3 SD in the second year. We formally test whether the treatment effect of the combination arm is greater than the sum individual CG and COD arms and report the P-Value of that test. Our tests show that there is significant evidence of the complementarities of resources and incentives especially in Swahili and Math.

We examine if there are positive (or negative) spillovers into non-incentivized subject (science) in Table 7. We do not find any evidence that the gains reported in Table 6 came at the expense of other subjects or grades. Overall the results suggest there are likely positive spillovers as students may be able to better understand other subjects as a result of improvements in literacy and numeracy. We also examine if our interventions affected performance in the Grade 7 national exit examination. One potential concern is that schools may invest most of their funds to support Grade 7 since school reputations are based on performance in this exam. However we do not see any evidence that any of our interventions significantly affect performance in

[5] We take the first component from a Principal Component Analysis (PCA) on the scores of the three subjects.

the national exit exam.

Table 7: Effect on test scores

	Science	Grade 7 National Exam 2013		
		Pass rate	Average score	Test takers
CG	-0.026	-0.026	-2.12	2.16
	(0.044)	(0.024)	(1.75)	(3.04)
COD	-0.040	-0.018	-1.26	2.31
	(0.040)	(0.025)	(2.00)	(2.99)
Combo	0.047	0.0016	0.067	3.18
	(0.041)	(0.026)	(2.10)	(3.08)
N. of obs.	18581	337	337	337
Mean control group		0.49	101.0	74.0
Combo-COD-CG	0.11*	0.045	3.45	-1.29
p-value (H_0:Combo-COD-CG=0)	0.067	0.24	0.26	0.78

Column (1) estimates equation 5 for Science using data for both follow-ups, and therefore coefficients represent the average treatment effect across both years. Column (2)-(4) use data from the national exit examination as dependent variables: pass rates, average test scores, and number of test takers. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.3 Teacher effort

We use self-reported data from teacher to examine differences in teacher effort. Table 8 shows that teachers in COD and Combo schools were more likely to offer tutoring, and teachers in COD schools gave out over 1.5 more tests on average per school year. Teachers in CG and Combo schools are more likely to report their teaching inputs to be above average, and less likely to have been switch into their grade recently. In the second year most of these effects are no longer statistically significant, but the point estimates for tutoring and remedial teaching are similar. Table 9 examines time use by teachers. Teachers in Combo schools devoted more time to extra classes. We argue that these results suggest that teacher behavior did change in response to the incentives.

Table 8: Focal years tutoring, tests and remedial

	Tests	Tutoring	Inputs	Moved grades
CG	-0.52	-0.0069	0.065***	0.0017
	(0.66)	(0.019)	(0.018)	(0.012)
COD	1.53**	0.041*	0.043**	-0.022**
	(0.63)	(0.023)	(0.021)	(0.011)
Combo	0.25	0.051**	0.16***	-0.035***
	(0.53)	(0.024)	(0.019)	(0.012)
N. of obs.	2923	2968	3572	3571
Mean of Dep. Var.	9.10	0.091	0.79	0.074
Combo-COD-CG	-0.76	0.016	0.049	-0.015
p-value (H_0:Combo-COD-CG=0)	0.41	0.64	0.086*	0.39

Results for estimating any treatment effects on teacher behavior. All data is self-reported. Column (1) has the number of test per period as the dependent variable. Column (2) has a dummy variable that indicates whether the teacher provided any extra tutoring to students as the dependent variable. Column (3) uses a dummy variable equal to one if teacher indicates teaching inputs are "above average" as the dependent variable. Finally, Column (4) uses a dummy variable equal to one if the teacher just started teaching a focal-grade subject combination recently. All regressions are done including data for both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Teacher Time Use

	Preparing class (Mins)	Teaching (Mins)	Extra classes (Mins)	Socializing (Mins)	Time at school (Hrs)
CG	0.47	1.24	1.85	0.38	0.044
	(2.12)	(4.34)	(2.43)	(2.10)	(0.11)
COD	0.37	-9.47**	3.01	1.62	0.14
	(2.05)	(4.77)	(2.47)	(2.50)	(0.086)
Combo	-0.73	1.25	4.31*	2.82	0.080
	(2.06)	(4.39)	(2.36)	(2.14)	(0.11)
N. of obs.	3030	3030	3030	3030	3033
Mean of Dep. Var.	43.3	152.4	26.8	37.9	7.71
Combo-COD-CG	-1.57	9.48	-0.54	0.81	-0.11
p-value (H_0:Combo-COD-CG=0)	0.63	0.15	0.89	0.82	0.46

Results for estimating any treatment effects on teacher time use. All data is self-reported. Column (1) estimates the effect on the time (in m
Column (2) on the time (in minutes) spent teaching regular classes, Column (3) on the time (in minutes) spent teaching extra classes, Colum
spent socializing, and Column (5) on the total number of hours spent at the school. All regressions are done including data for both follow-
represent the average effect over both years. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.4 Heterogeneity

We explore heterogeneous treatment effects by student gender, age and baseline test scores in Table 10. Overall there is limited evidence of heterogeneous effects by these characteristics. Although we note that girls performance in English improves less in Combo and COD schools.

Table 10: Heterogeneity by student characteristics

	Gender	Age	Lag test score
Panel A: Math			
	Gender	Age	Lag test score
CG*Covariate	0.043	-0.0078	-0.025
	(0.045)	(0.016)	(0.037)
COD*Covariate	-0.032	-0.012	-0.026
	(0.040)	(0.015)	(0.039)
Combo*Covariate	-0.075*	-0.020	-0.013
	(0.041)	(0.015)	(0.038)
N. of obs.	18581	18581	18581
Panel B: Swahili			
	Gender	Age	Lag test score
CG*Covariate	0.065	0.012	0.028
	(0.040)	(0.015)	(0.045)
COD*Covariate	-0.029	0.0039	-0.038
	(0.042)	(0.015)	(0.040)
Combo*Covariate	-0.057	-0.025*	-0.048
	(0.044)	(0.014)	(0.042)
N. of obs.	18581	18581	18581
Panel C: English			
	Gender	Age	Lag test score
CG*Covariate	0.012	-0.019	-0.034
	(0.031)	(0.012)	(0.041)
COD*Covariate	-0.053	-0.019	0.011
	(0.034)	(0.013)	(0.036)
Combo*Covariate	-0.072**	-0.020	0.0077
	(0.034)	(0.013)	(0.045)
N. of obs.	18581	18581	18581

The independent variable is the standardized test score. Each regression has a different covariate interacted with the treatment dummies. The column title indicates the covariate interacted. Baseline score is the standardized test score at the beginning of the first year; Male is a equal to one if the student is male; Age is the age in years of the student. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

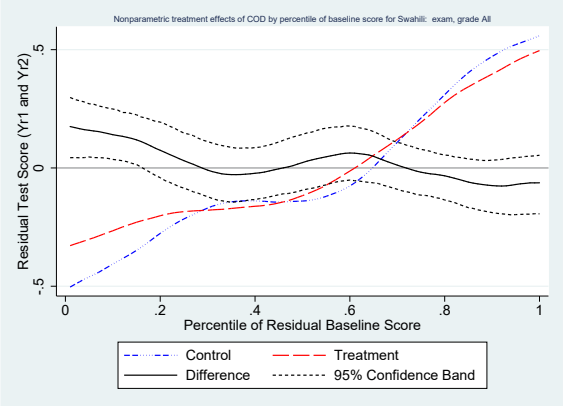
As discussed earlier, threshold incentive designs can lead teachers to focus on students near the passing level, perhaps at the expense of students who are in the tails of the distribution. We explore the potential for these effects using non-parametric analysis. We do a locally weighted regression of the end line test scores on the baseline score of students. Specifically, we estimate the following equation

$$Z_{it} = \alpha_0 + \alpha_1 F(Z_{i,t=0}) + \varepsilon_{it},$$

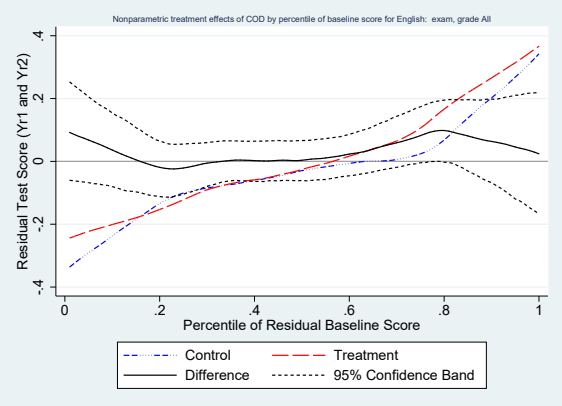
where F is the CDF of the baseline scores of students. The pointwise treatment effect is calculated as $g(x;T) = f(x;T) - f(x;Control)$ and the confidence intervals are estimated using bootstrapping. This enables us to estimate how the treatment effect varies for students with different initial abilities or knowledge.

We find that neither CG nor COD have treatment effects on students regardless of initial learning levels (See Figures 5a-5c and 6a-6c). However, we find evidence that the effects of the program in Combo schools were concentrated among students in the "middle" of the distribution, especially for Math and Swahili (See Figures 7c and 7a). This is consistent with the program introducing threshold effects, where teachers only focus on students near the qualification threshold. Given the low levels of English language skills, the treatment effects for students (in the Combo arm) in English were concentrated in the right tail (see Figure 7b). This is again consistent with the notion that teachers would only focus on students who were at the margin of passing.

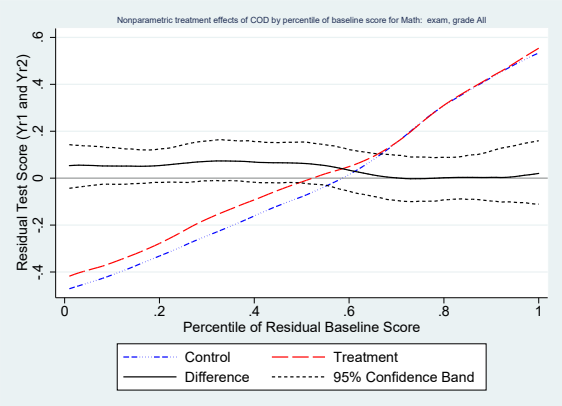
Figure 5: Non-parametric treatment effects of COD by percentile of baseline score



(a) Swahili

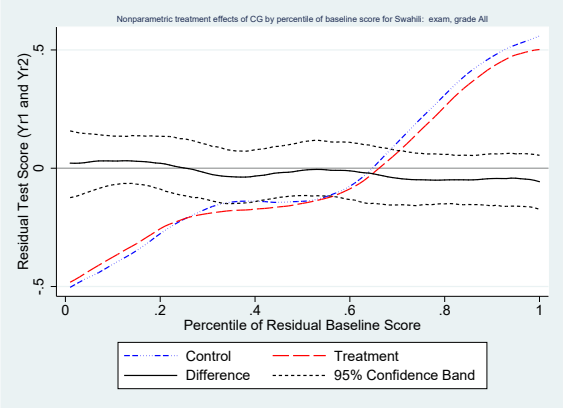


(b) English

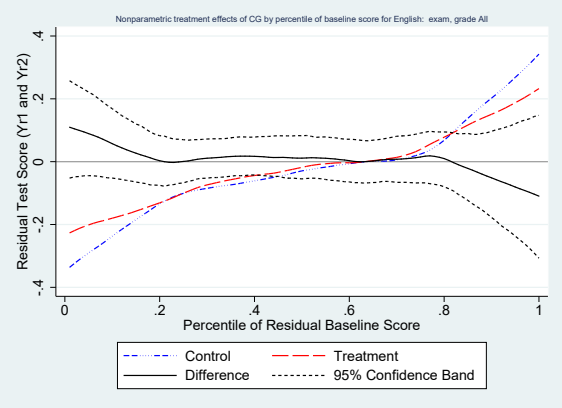


(c) Math

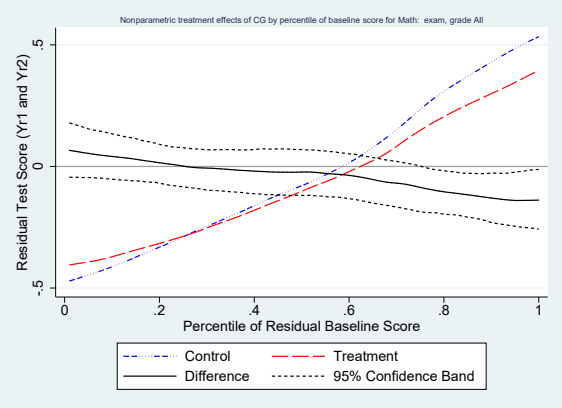
Figure 6: Non-parametric treatment effects of CG by percentile of baseline score



(a) Swahili

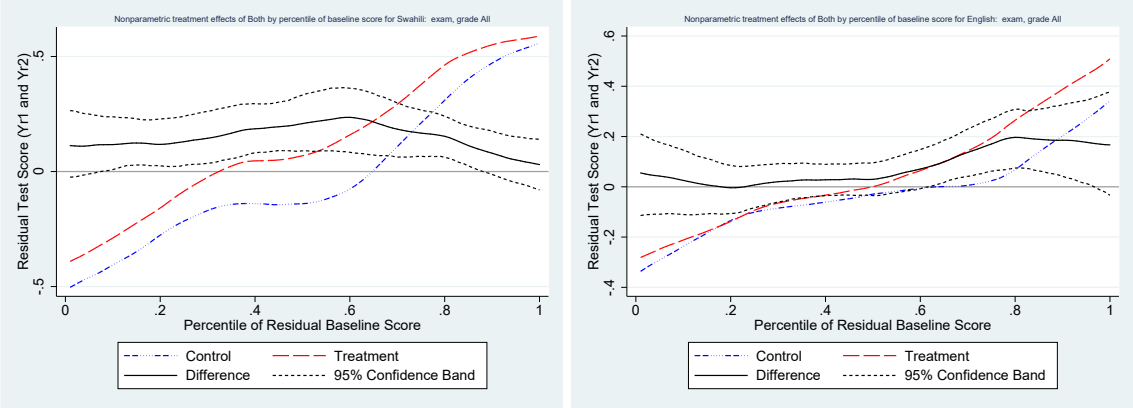


(b) English



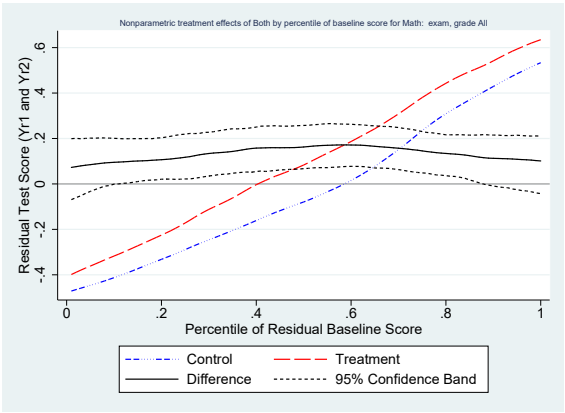
(c) Math

Figure 7: Non-parametric treatment effects of Combo by percentile of baseline score



(a) Swahili

(b) English



(c) Math

4.5 Low-stakes Vs high-stakes test

As mentioned before, two sets of tests performed to measure student learning levels. The intervention test was taken by all students in grades 1, 2 and 3 in COD and Combo schools to calculate teacher payments. Additionally, in 40 (4 per district) randomly selected control schools, all students in grades 1, 2 and 3 took the intervention test. This allows us to compare the results from a high-stakes (intervention) test, and from a low-stakes (research) test. For the low-stakes test 30 students in every schools were randomly selected to be tested. We were able to match around the results from the high and the low-stakes exam for about one third of the sample of students tested using the low-stakes exam.

Table 11 shows the comparison in the results when using the low- and the high-stakes exam. Its important to note that the comparison is made only over the students that we were able to match across both exams, and over a sample of questions that overlap across the two exams. See Appendix C for more details. As

can be seen, altough the point estimates for COD and Combo tend to be higher on the high stakes exam (except for Swahili), they are not statistically different from each other (except for the interaction effect in Swahili). We believe the difference between these the points estimates is mainly due to a test-day effect. When enumerators visit schools to do the low-stakes exams, schools activities mostly remain the same since only a small number of students is tested. On the other hand, when Twaweza visits schools to perform the high-stakes exams, most of the school activities are canceled since every student is tested. Additionally, teachers might motivate students to do well on the high-stakes exam.

Table 11: High- and low-stakes exams

	Low-stakes	High-stakes	Difference
Panel A: Math			
	(1)	(2)	(3)
COD	0.076	0.14**	0.068
	(0.060)	(0.057)	(0.053)
Combo	0.23***	0.27***	0.039
	(0.064)	(0.059)	(0.058)
N. of obs.	3465	3465	6930
Combo-COD	0.16	0.13	-0.029
p-value	0.0029***	0.0047***	0.56
Panel B: Swahili			
	(1)	(2)	(3)
COD	0.069	0.083	0.014
	(0.065)	(0.059)	(0.051)
Combo	0.27***	0.20***	-0.069
	(0.066)	(0.059)	(0.048)
N. of obs.	3465	3465	6930
Combo-COD	0.20	0.11	-0.083
p-value	0.00014***	0.011**	0.012**
Panel C: English			
	(1)	(2)	(3)
COD	-0.055	0.0082	0.063
	(0.067)	(0.063)	(0.067)
Combo	0.065	0.15**	0.090
	(0.070)	(0.073)	(0.077)
N. of obs.	3465	3465	6930
Combo-COD	0.12	0.15	0.027
p-value	0.025**	0.012**	0.67

Treatment effect in the low-stakes and high-stakes test, estimated for the same pool of students in a set of comparable questions (across tests). Column (3) reports whether the difference in the estimated treatment effects are different. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.6 Cross-cut designs in RCTs

Cross-cut designs are often employed in field experiments as a cost-saving measure. Typically, these designs are characterized by several randomly (and independently) assigned treatments, and subjects (or units) are allowed to receive multiple interventions. For example the researchers may randomly assign one group of subjects to receive treatment 1 and another group treatment 2, and another group to receive both treatments. By pooling together all units that received treatment 1 and all units that received treatment 2 in a regression framework, the conventional thinking is that these designs could provide a cost-effective way to increase the effective sample size and statistical power.[6] However, implementing such a research design could lead to incorrect statistical inference. As Kremer (2003) puts it: "Conducting a series of evaluations in the same area allows substantial cost savings. Once staff are trained, they can work on multiple projects. Since data collection is the most costly element of these evaluations, cross-cutting the sample reduces costs dramatically.... This tactic can be problematic, however, if there are significant interactions between programs".

The potential for significant interaction effects between treatments poses several problems for inference, which are typically unaccounted for. To illustrate this issue, let us consider a cross-cutting design with two interventions T_1 and T_2, and the following Data Generating Process (DGP):

$$Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_1 \times T_2 + \varepsilon, \tag{6}$$

where ε is the error term which satisfies $\mathbb{E}(\varepsilon) = 0$, $V(\varepsilon) < \infty$, $\mathbb{E}(\varepsilon T_1) = 0$, and $\mathbb{E}(\varepsilon T_2) = 0$. We also assume that the randomizations of T_1 and T_2 are independent from each other, which implies $\mathbb{E}(T_2 T_1) = 0$.

If we know that $\beta_3 = 0$ then we can consistently estimate β_1 and β_2 by simply estimating the following equation:

$$Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \underbrace{U}_{\beta_3 T_1 \times T_2 + \varepsilon} \tag{7}$$

since $\mathbb{E}T_1 U = 0$ OLS will give us consistent estimates of β_1. We can obtain consistent estimates of β_2 in a similar manner using OLS. However, if $\beta_3 \neq 0$ then $\mathbb{E}T_1 U \neq 0$, which and OLS no longer yields consistent estimators of β_1 and β_2 in equation 7. These are the sort of problems that Kremer (2003) refers to.

In many instances we would also be interested obtaining consistent estimates of β_3 by estimating 6. However, introducing the treatment 3 dummy in the estimating equation would reduce the statistical power

[6] To be precise the regression equation would be $Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \varepsilon$, where T_1 and T_2 are treatment 1 and treatment 2 dummies.

for estimating β_1 and β_2 compared to the parsimonious equation 7 . Intuitively, if we estimate equation 6, the identifying variation for β_1 comes from individuals that receive T_1 but not T_2. However, If we estimate 7 then the identifying variation comes from all individuals that get T_1 (regardless of whether they get T_2) which increases power relative to equation 6.

In practice, researchers typically employ a two- step procedure to determine whether to estimate equation 6 or 7. If a researcher estimates equation 6, and decides to keep his estimates of β_1, β_2, and β_3 or to estimate equation 7 to get new estimates of β_1 and β_2 (and assume $\beta_3 = 0$), then the finite-sample and asymptotic distribution of these estimators are complicated and highly non-normal(Leeb & Pötscher, 2005, 2006, 2008), making the usual t-statistics and p-values highly misleading.

Many cross-cut designed RCTs are not adequately powered to test for interactions. This leaves researches in somewhat of a cross-road. They can either assume that $\beta_3 = 0$ and estimate equation 7 to *hopefully* get consistent estimates of β_1 and β_2, or they can estimate equation 6 and lose some of the identifying variation for β_1 and β_2 even if β_3 is truly zero. A third alternative is to do model selection and use Bonferroni-style correction to adjust critical values (McCloskey, 2012). Intuitively, if one is certain that β_3 is within a certain range, one can study the behavior of $\hat{\beta_1}$ (under the null distribution) for each possible value of β_3 and keep the highest critical value (to be conservative). One can take this argument a step further, and do this for an asymptotically valid confidence set for β_3 (taken from the first step of the model selection), and adjust the critical values accordingly (McCloskey, 2012). Intuitively, the further the confidence interval onβ_3 is from zero, the smaller the adjustment on the critical values.. An important caveat, is that this is not a simple "standard error" adjustment, as the asymptotic distribution of the estimators no longer resembles a normal distribution as n approaches infinity. Therefore, McCloskey (2012) Bonferroni-style correction leads to new critical values, not to new standard errors.

To illustrate the effect of model selection on critical values we apply McCloskey (2012)'s Bonferroni-style correction to our own data. To do so, we re-estimate our main results as if we had done model selection. In other words, we estimate equation 6 where T_1 is CG and T_2 is COD, and if we cannot reject the null hypothesis that $\beta_3 = 0$, then we estimate equation 7. We adjust our critical values for β_1 and β_2 in each case. Table 12 shows the results from estimating equation 6, which is isomorphic to Table 6, but explicitly estimates the interaction term (β_3). To do inference on the null that $\beta = 0$, and in order to apply McCloskey (2012)'s Bonferroni-style correction, we will use as a test statistic the coefficient itself ($\hat{\beta}$). The critical value, without any correction, with significance at the 10% level is $\approx s.e.(\hat{\beta})z_{1-\alpha/2}$, where $s.e.(\hat{\beta})$ is the standard error of $\hat{\beta}$ and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the normal distribution. Notice that is isomorphic to using

$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$ as the t-statistics and $z_{1-\alpha/2}$ as the critical value. The adjusted critical values depend on whether one performs consistent or conservative model selection.[7]. When performing conservative model selection the selection threshold fix and do not depend on sample size (e.g., $s.e.(\hat{\beta})z_{1-\alpha/2}$). In consistent model selection the threshold grows as the sample size increases (e.g., the Hannan-Quinn information criterion $s.e.(\hat{\beta})\sqrt{\ln(n)}$). The former is more common in applied work, but leads to lower powered post-selection test's in general (McCloskey, 2012).

Table 12: Effect on test scores

	Year 1				Year 2			
	Math	Swahili	English	Focal Subjects	Math	Swahili	English	Focal Subjects
CG	-0.051	0.0039	-0.0099	-0.022	0.0042	-0.019	0.0092	-0.0031
	(0.065)	(0.068)	(0.059)	(0.061)	(0.078)	(0.082)	(0.068)	(0.072)
COD	0.035	0.054	0.068*	0.063*	0.063	0.0072	0.029	0.039
	(0.063)	(0.064)	(0.063)	(0.060)	(0.074)	(0.080)	(0.061)	(0.068)
CG X COD	0.11*	0.059	0.026	0.076	0.12*	0.21***	0.056	0.16**
	(0.099)	(0.098)	(0.100)	(0.093)	(0.11)	(0.12)	(0.10)	(0.10)
N. of obs.	9142	9142	9142	9142	9439	9439	9439	9439

Results for estimating equation 5 for different subjects at both follow-ups. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.6.1 Conservative model selection

Table 13 shows the results from applying conservative model selection (using $s.e.(\hat{\beta})1.64$ as the threshold). The columns in table 12 where $\hat{\beta_3} < s.e.(\hat{\beta})1.64$ are re-estimated using equation 7, and β_3 is assumed to be zero. In parenthesis is the critical value without correction ($s.e.(\hat{\beta})1.64$), in square parenthesis the adjusted critical value from McCloskey (2012)'s Bonferroni-style correction after conservative model selection. Notice that without adjustment the effect of COD is now significant (at the 10%) level for Swahili (Yr1), Index (Yr1) and English (Yr1 and Yr2). However, once we adjust critical values only English (Yr1) remains significant, as in table 12. Importantly, the adjusted critical values tend to be, in general, smaller than in the original regression.

[7]In conservative model selection the selection threshold does not depend on the sample size. In consistent model selection, the threshold grows as the sample size increases.

Table 13: Effect on test scores

	Year 1				Year 2			
	Math	Swahili	English	Index	Math	Swahili	English	Index
CG	-0.049	0.03	-0.0003	0.012	0.0033	-0.02	0.034	-0.0066
	(0.066)	(0.05)	(0.052)	(0.084)	(0.079)	(0.082)	(0.056)	(0.13)
	[0.079]	[0.084]	[0.068]	[0.13]	[0.097]	[0.1]	[0.082]	[0.15]
COD	0.036	0.079	0.081	0.14	0.064	0.0072	0.056	0.079
	(0.063)	(0.049)*	(0.049)*	(0.08)*	(0.074)	(0.079)	(0.054)*	(0.12)
	[0.078]	[0.077]*	[0.076]*	[0.13]*	[0.092]	[0.098]	[0.072]	[0.15]
CG X COD	0.1	0	0	0	0.12	0.21	0	0.26
	(0.1)*				(0.11)*	(0.12)*		(0.18)*
N. of obs.	9141	9141	9141	9141	9436	9436	9436	9436

Non-adjusted critical values (for a significance level of 10%) in parenthesis. Adjusted critical values (for a significance level of 10%) in squarte parenthesis.
* significant at the 10% level.

4.6.2 Consistent model selection

Table 14 shows the results from applying consistent model selection (using $\sqrt{\ln(n)}$ as the threshold). The columns in table 12 where $\hat{\beta}_3 < \sqrt{\ln(n)}$ are re-estimated using equation 7, and β_3 is assumed to be zero. In parenthesis is the critical value without correction ($s.e.(\hat{\beta})1.64$), in square parenthesis the adjusted critical value from McCloskey (2012)'s Bonferroni-style correction after consistent model selection. Notice that without adjustment the effect of COD is now significant for every subject (in both years), and that CG is significant for Swahili (Yr2) and Index (Yr2). However, after the adjustment

Table 14: Effect on test scores

	Year 1				Year 2			
	Math	Swahili	English	Index	Math	Swahili	English	Index
CG	-0.0035	0.03	-0.0003	0.012	0.055	0.07	0.034	0.1
	(0.052)	(0.05)	(0.052)	(0.084)	(0.061)	(0.061)*	(0.056)	(0.097)*
	[0.097]	[0.078]	[0.059]	[0.15]	[0.11]	[0.15]	[0.076]	[0.21]
COD	0.081	0.079	0.081	0.14	0.12	0.1	0.056	0.19
	(0.049)*	(0.049)*	(0.049)*	(0.08)*	(0.057)*	(0.06)*	(0.054)*	(0.092)*
	[0.094]	[0.075]*	[0.062]*	[0.15]	[0.11]*	[0.16]	[0.071]	[0.21]
CG X COD	0	0	0	0	0	0	0	0
N. of obs.	9141	9141	9141	9141	9436	9436	9436	9436

Non-adjusted critical values (for a significance level of 10%) in parenthesis. Adjusted critical values (for a significance level of 10%) in squarte parenthesis.
* significant at the 10% level.

5 Conclusion

In this paper we report findings from a large education RCT aimed at improving learning in early grades. Consistent with other findings, we show that merely increasing school resources do little to improve learning outcomes. We also find that a simple incentive program yield insignificant but positive impacts on learning. We find that test scores in schools that received both programs were significantly higher. Moreover, we find strong evidence of complementarities between inputs/resources and incentives. We further find that the increases in learning (in the combo schools) were concentrated among students near the passing threshold. This is further evidence of the importance of incentive design in promoting student learning. Finally, we highlight the potential danger in inference in ignoring complementarities between programs. Ignoring complementarities in cross-cutting experimental designs may yield biased estimates and lead to the scale up and adoption of ineffective programs. We thus argue that researchers should adequately design studies to account for complementarities.

References

Centre de Recherche Economique et Sociale. (2013). *Service delivery indicators: Pilot in education and health care in tanzania* (Tech. Rep.). World Bank.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2013). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics*, *5*(2), 29-57. Retrieved from http://www.aeaweb.org/articles.php?doi=10.1257/app.5.2.29 doi: 10.1257/app.5.2.29

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from new york city public schools. *Journal of Labor Economics*, *31*(2), 373-407. Retrieved from http://www.jstor.org/stable/10.1086/667757

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, 205–227.

Grogan, L. (2009). Universal primary education and school entry in uganda. *Journal of African Economies*, *18*(2), 183–211.

Jones, S., Schipper, Y., Ruto, S., & Rajani, R. (2014). Can your child read and count? measuring learning outcomes in east africa. *Journal of African Economies*. Retrieved from http://jae.oxfordjournals.org/content/early/2014/06/12/jae.eju009.abstract doi: 10.1093/jae/eju009

Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: Some lessons. *The American Economic Review*, *93*(2), pp. 102-106. Retrieved from http://www.jstor.org/stable/3132208

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*(6130), 297–300.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21-59. Retrieved from http://www.jstor.org/stable/3533623

Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 2554–2591.

Leeb, H., & Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, *24*(02), 338–376.

Lucas, A. M., & Mbiti, I. M. (2012). Access, sorting, and achievement: the short-run effects of free primary education in kenya. *American Economic Journal: Applied Economics*, 226–253.

McCloskey, A. (2012). Bonferroni-based size-correction for nonstandard testing problems. *Available at SSRN 2171912*.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries a meta-analysis of randomized experiments. *Review of Educational Research*, *85*(3), 353–394.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. *Journal of Political Economy*, *119*(1), pp. 39-77. Retrieved from http://www.jstor.org/stable/10.1086/659655

Murnane, R. J., & Ganimian, A. J. (2014). *Improving educational outcomes in developing countries: Lessons from rigorous evaluations* (Tech. Rep.). National Bureau of Economic Research.

Neal, D., & Schanzenbach, D. W. (2010, February). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, *92*(2), 263–283. Retrieved from http://dx.doi.org/10.1162/rest.2010.12318

Pop-Eleches, C., & Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, *103*(4), 1289-1324. Retrieved from http://www.aeaweb.org/articles.php?doi=10.1257/aer.103.4.1289 doi: 10.1257/aer.103.4.1289

Sabarwal, S., Evans, D. K., & Marshak, A. (2014). The permanent input hypothesis: the case of textbooks and (no) student learning in sierra leone. *World Bank Policy Research Working Paper*(7021).

Sandefur, J., & Glassman, A. (2015). The political economy of bad data: Evidence from african survey and administrative statistics. *The Journal of Development Studies*, *51*(2), 116-132. Retrieved from http://dx.doi.org/10.1080/00220388.2014.968138 doi: 10.1080/00220388.2014.968138

Twaweza. (2013). *Capitation grants in primary education: A decade since their launch, does money reach schools?* (Tech. Rep.). Author.

Uwezo. (2012). *Are our tanzania learning? literacy and numeracy in tanzania* (Tech. Rep.). Author. Retrieved from http://www.twaweza.org/uploads/files/UwezoTZ-ALA2014-FINAL-EN.pdf (Accessed on 05-02-2016)

Uwezo. (2013). *Are our children learning? numeracy and literacy across east africa* (Tech. Rep.). Author. Retrieved from http://www.twaweza.org/uploads/files/UwezoTZ2013forlaunch.pdf (Accessed on 05-12-2014)

Valente, C. (2015). Primary education expansion and quality of schooling: Evidence from tanzania.

World Bank. (2012). *Tanzania service delivery indicators* (Tech. Rep.). Service Delivery Indicators.

A Theoretical Model - Additional Proofs

The FOC implies that the optimal level of effort (e^*) satisfies:

$$\underbrace{(t+\lambda)f_e(e^*,I)}_{\text{Marginal benefit}} = \underbrace{c_e(e^*)}_{\text{Marginal cost}}$$

and we get the following level of learning $L^* = f(e^*, I)$. By means of the implicit function theorem we get that

$$\frac{\partial e^*}{\partial t} = -\frac{f_e(e^*,I)}{f_{ee}(e^*,I)(t+\lambda) - c_{ee}(e^*)} \tag{8}$$

$$\tag{9}$$

Notice that $\frac{\partial e^*}{\partial t} > 0$ since $f_{ee} < 0$ and $c_{ee} > 0$. In other words, an increase in the piece rate bonus increases teacher effort. This in turns increases learning since:

$$\frac{\partial L^*}{\partial t} = f_e(e^*,I)\frac{\partial e^*}{\partial t} > 0 \tag{10}$$

Now, lets see what happens as we increase the inputs I:

$$\frac{\partial e^*}{\partial I} = -\frac{f_{eI}(e^*,I)(t+\lambda)}{f_{ee}(e^*,I)(t+\lambda) - c_{ee}(e^*)} \tag{11}$$

$$\tag{12}$$

Therefore $\frac{\partial e^*}{\partial I} > 0$ if $f_{eI}(e^*, I) > 0$ and $\frac{\partial e^*}{\partial I} < 0$ if $f_{eI}(e^*, I) < 0$. In turn, this means that learning might not increase with an increase in inputs since

$$\frac{\partial L^*}{\partial I} = f_e(e^*,I)\frac{\partial e^*}{\partial I} + f_I(e^*,I) \tag{13}$$

Notice however that its possible that $\frac{\partial L^*}{\partial I} > 0$ even if $\frac{\partial e^*}{\partial I} < 0$. Now, lets see what happens as we increase the piece rate t and the inputs I to learning

$$\frac{\partial^2 e^*}{\partial t \partial I} = \frac{f_{ee}(e^*,I)\frac{\partial e^*}{\partial I} + f_{eI}(e^*,I) + (t+\lambda)\left(f_{eee}(e^*,I)\frac{\partial e^*}{\partial I}\frac{\partial e^*}{\partial t} + f_{eeI}(e^*,I)\frac{\partial e^*}{\partial t}\right) - c_{eee}(e^*)\frac{\partial e^*}{\partial I}\frac{\partial e^*}{\partial t}}{c_{ee}(e^*) - f_{ee}(e^*,I)} \tag{14}$$

$$\frac{\partial^2 L^*}{\partial t \partial I} = \frac{\partial e^*}{\partial t}\left[f_{ee}(e^*, I)\frac{\partial e^*}{\partial I} + f_{eI}(e^*, I)\right] + f_e(e^*, I)\frac{\partial^2 e^*}{\partial I \partial t} \quad (15)$$

$$= \frac{\partial e^*}{\partial t}\left[-f_{ee}(e^*, I)\frac{f_{eI}(e^*, I)(t+\lambda)}{f_{ee}(e^*, I)(t+\lambda) - c_{ee}(e^*)} + f_{eI}(e^*, I)\right] + f_e(e^*, I)\frac{\partial^2 e^*}{\partial I \partial t} \quad (16)$$

$$= \frac{\partial e^*}{\partial t} f_{eI}(e^*, I)\frac{c_{ee}(e^*)}{c_{ee}(e^*) - f_{ee}(e^*, I)(t+\lambda)} + f_e(e^*, I)\frac{\partial^2 e^*}{\partial I \partial t} \quad (17)$$

$$(18)$$

B Additional tables

Table 15: Total expenditure

	$ Total.	$ Admin.	$ Student	$ Teaching Aid	$ Teacher	$ Construction
CG	4922.4***	1873.8***	114.5	2723.7***	16.2	194.1
	(709.0)	(357.7)	(419.7)	(251.6)	(63.9)	(254.8)
COD	-613.9	-109.2	-375.3	-282.2*	-13.5	166.3
	(534.6)	(169.1)	(345.1)	(148.1)	(40.1)	(299.6)
Combo	5457.6***	2052.7***	159.5	3420.2***	18.0	-192.9
	(709.7)	(315.3)	(443.6)	(317.8)	(47.2)	(201.2)
N. of obs.	699	699	699	699	699	699
Mean control	5241.9	1752.9	775.7	2029.9	138.7	544.6
Combo-COD-CG	1149.1	288.1	420.3	978.7**	15.3	-553.3
p-value (H_0:Combo-COD-CG=0)	0.20	0.50	0.33	0.014	0.84	0.14

Mean expenditure per student. Standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 16: Substitution from other sources

	Total	Government CG	Government Other	Local Government	NGOs	Parents	Other
CG	-390.8	11.0	124.1	-80.7	82.4	-513.4	-14.3
	(562.5)	(297.4)	(182.9)	(96.2)	(62.7)	(417.5)	(76.3)
COD	-597.8	-313.2	35.0	100.9	-18.4	-488.4	86.4
	(533.4)	(285.7)	(150.0)	(190.9)	(30.4)	(387.5)	(97.1)
Combo	-207.3	-103.1	449.5	-126.8	62.1	-530.3	41.4
	(659.7)	(322.1)	(344.1)	(91.0)	(74.9)	(450.2)	(78.9)
N. of obs.	699	699	699	699	699	699	699
Mean control	5427.5	3438.2	104.3	192.7	10.7	1565.5	116.0
Combo-COD-CG	781.3	199.0	290.4	-147.1	-1.81	471.4	-30.7
p-value (H_0:Combo-COD-CG=0)	0.28	0.63	0.36	0.45	0.98	0.30	0.80

Mean expenditure per student. Standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 17: Household expenditure

	Total Expenditure	Fees	Textbooks	Other books	Supplies	Uniforms	Tutoring	Transport	Others
CG	-2240.3*	-584.7	-118.7*	29.9	91.9	-430.1	-793.0	-164.9	-253.4*
	(1321.3)	(501.0)	(71.7)	(58.3)	(179.8)	(514.3)	(583.5)	(222.9)	(149.4)
COD	659.5	-230.9	-44.0	13.1	365.8*	196.7	467.7	-106.3	-69.1
	(1299.6)	(495.9)	(77.6)	(40.9)	(197.6)	(474.8)	(597.0)	(229.9)	(171.0)
Combo	269.2	278.9	156.0	28.5	-406.5	268.1	-175.7	-45.7	214.3
	(1767.1)	(649.7)	(105.3)	(83.0)	(276.8)	(757.6)	(956.3)	(283.2)	(226.9)
N. of obs.	6709	6717	6717	6717	6717	6717	6717	6717	6717
Mean control	27012.4	2962.7	367.0	139.2	4491.9	13144.2	3741.7	351.7	1797.5
Combo-COD-CG	1850.0	1094.5	318.8	-14.4	-864.3	501.5	149.6	225.5	536.9
p-value (H_0:Combo-COD-CG=0)	0.63	0.45	0.15	0.92	0.13	0.73	0.93	0.74	0.26

Mean expenditure per student. Standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C Low- and high-stakes exams

Table 18 shows the results from using the low and the high stakes exam. Several adjustments must be made in order to make the results comparable. Columns 1-5 show the results using the low-stakes exam, while Columns 6-8 show the results using the high-stakes exam. The first column replicates the results from table 6; the second column restricts the sample to those schools that were tested using both the high and the low-stakes exam; the third replaces week fixed effects, for flexible time controls between both exams; since the intervention (high-stakes) exam was much shorter and had a more narrow set of questions than the low-stakes exam, the fourth column uses only the questions for which the difficulty between the low- and the high-stakes exam overlap; the fith column restricts the sample to the students that we were able to match across both exams. The sixth column is identical to column the previous column, but uses the high-stakes exam. Since we are unable to use student lagged-test scores as controls for all students tested in the high-stakex exam, we use the average lagged test score per school as a control in the seventh column; finally, column eight uses the full sample of students tested using the high-stakes exam.

Table 18: High- and low-stakes exams

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Math								
COD	0.064	0.041	0.094	0.087	0.066	0.14**	0.14**	0.14***
	(0.045)	(0.059)	(0.062)	(0.066)	(0.067)	(0.057)	(0.057)	(0.054)
Combo	0.19***	0.21***	0.24***	0.23***	0.21***	0.27***	0.27***	0.26***
	(0.047)	(0.061)	(0.063)	(0.066)	(0.069)	(0.059)	(0.059)	(0.059)
N. of obs.	9436	4859	4859	4859	3465	3465	3465	4078
Combo-COD[-CG]	0.12*	0.17***	0.14***	0.15***	0.15***	0.13***	0.13***	0.12***
p-value	0.077	0.0012	0.0061	0.0054	0.0075	0.0047	0.0047	0.0073
Panel B: Swahili								
COD	0.0072	0.011	0.042	0.073	0.089	0.083	0.083	0.056
	(0.048)	(0.068)	(0.068)	(0.066)	(0.066)	(0.059)	(0.059)	(0.060)
Combo	0.20***	0.22***	0.25***	0.26***	0.26***	0.20***	0.20***	0.19***
	(0.044)	(0.064)	(0.066)	(0.062)	(0.065)	(0.059)	(0.059)	(0.061)
N. of obs.	9436	4859	4859	4859	3465	3465	3465	4078
Combo-COD[-CG]	0.21***	0.21***	0.21***	0.18***	0.17***	0.11**	0.11**	0.13***
p-value	0.0029	0.00011	0.00012	0.00020	0.00065	0.011	0.011	0.0039
Panel C: English								
COD	0.033	-0.015	-0.021	-0.022	-0.056	0.0082	0.0082	0.0061
	(0.039)	(0.056)	(0.056)	(0.056)	(0.068)	(0.063)	(0.063)	(0.059)
Combo	0.097*	0.10	0.095	0.045	0.012	0.15**	0.15**	0.15**
	(0.051)	(0.062)	(0.062)	(0.061)	(0.071)	(0.073)	(0.073)	(0.069)
N. of obs.	9436	4859	4859	4859	3465	3465	3465	4078
Combo-COD[-CG]	0.052	0.12**	0.12**	0.067	0.068	0.15**	0.15**	0.14**
p-value	0.44	0.017	0.012	0.16	0.23	0.012	0.012	0.012
Student Controls	Yes	Yes	Yes	Yes	Yes	Yes	No	No
School Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week F.E.	Yes	Yes	No	No	No	No	No	No
Timing Controls	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Test	Survey	Survey	Survey	Survey	Survey	Intervention	Intervention	Intervention
Schools	All	Common	Common	Common	Common	Common	Common	Common
Questions	All	All	All	Common	Common	Common	Common	Common
Students	All	All	All	All	Common	Common	Common	All

Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$