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

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

July 21, 2018

Abstract

The idea that complementarities across policies can yield increasing returns from joint implementation has been posited in several economic settings. Yet there is limited, well-identified evidence of such complementarities in practice. We present results from a randomized experiment across a representative sample of 350 schools in Tanzania that studied the impact of providing schools with (a) unconditional school grants, (b) bonus payments to teachers based on student performance, and (c) both of the above. At the end of two years, we find (a) no impact on student test scores from providing school grants, (b) some evidence of positive effects from offering performance-linked bonuses to teachers, and (c) significant positive effects on learning from providing both programs. Most importantly, we find strong evidence of complementarities between the two programs, with the effect of joint provision being significantly greater than the sum of the individual effects. Our results suggest that accounting for complementarities between inputs and incentives could substantially improve the effectiveness of public spending on education.

JEL Classification: C93, H52, I21, M52, O15
Keywords: school grants, teacher performance pay, complementarities, education policy, Tanzania

*We are grateful to Joseph Mmbando who superbly oversaw the implementation team. We thank Oriana Bandiera, Prashant Bharadwaj, Julie Cullen, Gordon Dahl, Taryn Dinkleman, Eric Edmonds, Caroline Hoxby, David Figlio, Kelsey Jack, Kirabo Jackson, Jason Kerwin, Prashant Loyalka, Craig McIntosh, Adam Osman, Imran Rasul, Mark Rosenzweig, Abhijeet Singh, Tavneet Suri, Rebecca Thornton and several seminar participants for comments. In addition, we would like to acknowledge the support of Bryan Plummer and J-PAL Africa staff. Erin Litzow, Jessica Mahoney, Kristi Post, and Rachel Steinacher provided excellent on-the-ground research support through Innovations for Poverty Action. We also thank Ian McDonough for additional research support. The data collection was conducted by the EDI Tanzania team including Respichius Mitti, Andreas Kutka, Timo Kyessy, Phil Itanisia, Amy Kahn and Lindsey Roots, and we are grateful to them for their excellent data collection efforts. We received IRB approval from Innovations for Poverty Action, Southern Methodist University, UC San Diego, and University of Virginia. The protocol was also reviewed and approved by the Tanzania Commission for Science and Technology (COSTECH). A randomized controlled trials registry entry and the pre-analysis plan are available at: https://www.socialscienceregistry.org/trials/291.

†University of Virginia; J-PAL; IZA: imbiti@virginia.edu
‡University of California, San Diego; NBER; J-PAL: kamurali@ucsd.edu
§University of California, San Diego: mtromero@ucsd.edu
¶Twaweza: yschipper@twaweza.org
‖Yale University: constantine.manda@yale.edu
**Twaweza: rakeshrajani@gmail.com

1 Introduction

Improving education quality in low-income countries is a top priority for the global human development agenda (United Nations, 2015), with governments and donors spending over a hundred billion dollars annually on education (World Bank, 2017). Yet, developing country education systems face several challenges, and have found it difficult to convert increases in spending and enrollment into improvements in student learning (World Bank, 2018). Some of these challenges include resource scarcity in schools, poor student health and nutrition, low student attendance, low human capital of teachers and parents, mismatch between curriculum/pedagogy and student learning levels, and low levels of teacher effort and accountability.[1]

One implication of the multiple constraints described above is that policies that address these individually may have limited impact on learning outcomes if other binding constraints are not alleviated. Thus, the impact of policies that alleviate these constraints simultaneously may be greater than the aggregate impact of addressing each constraint individually. This possibility has influenced the design of social programs in both developed and developing countries.[2] However, while the idea of complementarities across policies to improve human welfare has been a central theme in development economics (Johnston & Mellor, 1961; Ray, 1998; Banerjee & Duflo, 2005), there is limited well-identified evidence of such complementarities in practice.

This paper tests for the presence of complementarities across education policies using a large-scale randomized evaluation. Our study is set in Tanzania, where two widely-posited constraints to education quality are a lack of school resources, and low teacher motivation and effort (World Bank, 2012). We study the individual impact of two programs, each designed to alleviate one of these constraints, and also study the impact of providing these programs jointly. The first program aimed to alleviate resource constraints by providing schools with grants that nearly *tripled* the per-student resources available to them (not including infrastructure and teacher salaries). The second one aimed to improve teacher motivation and effort by providing teachers with performance-based bonuses — based on the number of their students who passed basic tests of math, Kiswahili (local language), and English. A teacher with average enrollment could earn up to 125% of monthly base pay as a bonus.

[1] Each of these challenges has been extensively documented in multiple developing country settings. See Glewwe and Muralidharan (2016) and Mbiti (2016) for reviews and references to primary sources.

[2] Examples include Head-start in the US (which provides a combination of education, nutrition, and health services for early-childhood development) and anti-poverty graduation programs in several developing countries (which provide ultra-poor households with a combination of physical capital, human capital, and ongoing engagement and support) (Banerjee et al., 2015; Bandiera et al., 2017).

We conducted the experiment in a large nationally-representative sample of 350 public schools (and over 120,000 students) across 10 districts in mainland Tanzania. We randomly allocated schools to four groups (stratified by district): 70 received unconditional school grants, 70 received the teacher performance pay program, 70 received *both* programs, and 140 were assigned to a control group. The study was powered adequately to test for complementarities, and we gave the same importance to testing for complementarities as testing for the main effects of the two programs.[3] All programs were implemented by Twaweza, a leading Tanzanian non-profit organization.

We report four sets of results. First, the school grant significantly increased per-student expenditure in treated schools. Consistent with prior findings (as in Das et al. (2013)) we find evidence of crowding out of school and household spending in treated schools. After this reduction, there was still a near doubling of *net* school-level spending per student in treated schools (excluding teacher salaries). However, this increase in spending had no impact on student learning outcomes on low-stakes tests (conducted by the research team) in math, Kiswahili, or English after both one and two years.

Second, we find mixed evidence on the impact of teacher performance pay on student learning. On low-stakes tests conducted by the research team, we find that student test-scores in treated schools were modestly higher than those in the control group, but typically not significant. However, we find significant positive treatment effects on the high-stakes tests administered by Twaweza. After two years, students in treated schools were 37%, 17%, and 70% more likely to pass the Twaweza tests in math, Kiswahili, and English — the outcome that teacher bonuses were based on. Overall, scores on high-stakes tests were 0.21σ higher in treated schools after two years. As specified in our pre-analysis plan, the analysis in this paper is mainly based on the low-stakes tests.[4] We present results on high-stakes tests to enable comparison with other studies on teacher performance pay (that report results using high-stakes tests), and defer discussion of the differences in results on the two sets of tests and their implications to Section 5.2.

Third, students in schools that received both inputs and incentives had significantly higher test scores (relative to the control group) in *all* subjects on *both* the low-stakes and high-stakes tests. After two years, composite test scores were 0.23σ higher on the low-stakes tests, and 0.36σ higher on the high-stakes tests. Student passing rates on the latter were 49%, 31%, and 116% higher in math, Kiswahili, and English.

[3]Trial registry and pre-analysis plan available at https://www.socialscienceregistry.org/trials/291

[4]Our pre-analysis plan focuses on the low-stakes tests because we only collected data on learning outcomes in *all* treatment groups for the low-stakes tests (high-stakes tests were not conducted in the Grant schools since they were not needed for program implementation). Thus, all tests of complementarity (which was a central topic of interest for this study) are based on the low-stakes tests.

Fourth, and most important, we find strong evidence of complementarities between inputs and incentives. At the end of two years, test score gains in the Combination schools were significantly greater than the sum of the gains in Grant and Incentives schools in *each* of the three subjects (math, Kiswahili, and English). Using a composite measure of test-scores across subjects, the "interaction" effect was equal to 0.18σ ($p < 0.01$). These complementarities are quantitatively important: point estimates of the impact of the Combination treatment are over three times greater than the sum of the impact of the Grant and Incentives treatments after one year, and over *five times greater* after two years.

To help interpret our results, we present a simple theoretical framework that specifies an education production function and a teacher's optimization problem regarding how much effort to exert. The key insights from the model are the following: first, the observed effects of policy changes will depend not just on the production function but also on changes in effort induced by the policy change. Second, even if there are complementarities in the production function between inputs and effort, if teachers act like agents in standard economic models (with disutility from effort and no intrinsic motivation), then the optimal response to an increase in inputs may be to reduce effort, which may attenuate impacts on learning. Third, the introduction of financial incentives will typically raise the optimal amount of teacher effort when inputs increase, yielding complementarities between inputs and incentives in improving learning outcomes.

We make several contributions to research and policy. Our first and most important contribution is to *experimentally* establish the existence of complementarities across policies aiming to improve human capital, which (to the best of our knowledge) has not been shown to date. Despite strong interest in complementarities (Bleakley, 2010), credible evidence is limited as observational studies require two sources of exogenous variation (or "two lightning strikes" according to Almond and Mazumder (2013)). Recent studies have examined complementarities between a variety of policy pairs, such as home environment and school quality, grade retention and school accountability, and Head-start and public school spending (Malamud, Pop-Eleches, & Urquiola, 2016; Geng, 2018; Johnson & Jackson, 2017). However, the lack of experimental variation in these studies requires exogeneity to be established for *both* sets of policies, which is a non-trivial challenge.

Turning to experiments, several studies have employed factorial (or cross-cutting) designs that could in principle be used to test for complementarities. However, due to budget and sample size constraints, these studies have typically been under-powered to detect economically meaningful complementarities. In practice, they often *assume away* complementarities to increase power in estimating the effects of the main treatments of interest (see Kremer (2003); Muralidharan, Romero, and Wuthrich (2018) for

reviews). Other experiments have evaluated basic and augmented versions of a program and study variants A, and A + B; but not A, *B*, and A + B, which would be needed to test for complementarities (for instance, see Pradhan et al. (2014); Kerwin and Thornton (2017)). Finally, experimental studies of teacher incentive programs find larger effects in schools with more resources, but this evidence is only suggestive of complementarities because of lack of random assignment of the inputs (see Muralidharan and Sundararaman (2011b); Gilligan, Karachiwalla, Kasirye, Lucas, and Neal (2018)).

The closest experimental study that was explicitly designed to test for complementarities in human capital formation is Attanasio et al. (2014) which studies the effects of providing (1) nutrition supplements, (2) stimulation programs, and (3) both of them, on early childhood development in Colombia, and finds no evidence of complementarities across the two programs studied.[5]

Second, our results and theoretical framework help to clarify two important points regarding the study of complementarities in human capital formation. While much of the theoretical literature focuses on *production function* complementarities (Heckman, 2007; Cunha & Heckman, 2007), the possibility of behavioral responses makes it difficult to identify these empirically. Thus, even well-identified studies (including ours) will estimate *policy* and not production-function complementarities. Moreover, even if there are production-function complementarities between two sets of inputs, there may not be policy complementarities from providing both because the former may be offset by a reduction in agent effort. In contrast, combining inputs and incentives is more likely to increase agent effort. Thus, there are more likely to be policy complementarities between interventions that improve inputs and agent effort (which is what we find).

Third, we contribute to the broader literature on teacher incentives. While there is generally mixed evidence on the effectiveness of teacher incentives, the patterns in the results suggest that such policies are more effective in developing countries (Ganimian & Murnane, 2014). Our results are consistent with this view and with results from Lavy (2002, 2009); Glewwe, Ilias, and Kremer (2010); Muralidharan and Sundararaman (2011b); Duflo, Hanna, and Ryan (2012); Contreras and Rau (2012); and (Muralidharan, 2012) who find that various forms of performance linked pay for teachers in low- and middle-income countries improved student test scores.[6]

[5] Behrman, Parker, Todd, and Wolpin (2015) study the impacts of providing (1) student incentives, (2) teacher incentives, and (3) both of them, on learning of high school students in Mexico. However, they do not test for complementarities because the variants of student and teacher incentives provided in the combined treatment arm (3) were not the same as those in the individual treatment arms (1) and (2).

[6] The claim that our results are consistent with prior evidence is based on results using our high-stakes tests because most of these studies (except Duflo et al. (2012)) report impacts on high-stakes tests.

Finally, our results suggest that a likely reason for the poor performance of input-based education policies in developing countries is the absence of adequate teacher incentives for *using* resources effectively. Several randomized evaluations have found that augmenting school resources has little impact on learning outcomes in developing countries (see for example Glewwe, Kremer, and Moulin (2009); Blimpo, Evans, and Lahire (2015); Das et al. (2013); Pradhan et al. (2014); Sabarwal, Evans, and Marshak (2014)). Our results replicate the results on the non-impact of providing additional school inputs, but also show that the inputs can improve learning when combined with teacher incentives.[7]

The idea that there may be complementarities between resources and incentives is gaining policy traction globally. Donors such as the World Bank are increasingly using results-based-financing schemes in education (as proposed by Birdsall, Savedoff, Mahgoub, and Vyborny (2012)), and several US states are exploring similar reforms that link parts of school financing to performance on statewide tests (Collier, 2016; Mesecar & Soifer, 2016; Calefati, 2016). Our results provide empirical support for such policy approaches, and suggest that accounting for complementarities between inputs and incentives could substantially improve the effectiveness of public spending on education.

2 Theoretical Framework

We present a simple model of how changes in inputs and incentives translate into changes in teacher effort and student learning outcomes. The model has three goals: first, it clarifies that the impact of an education intervention on learning outcomes will depend on both the production function *and* behavioral responses by teachers. In other words, experiments will typically identify the "policy effect" of an intervention and not the "production function" parameters (Todd & Wolpin, 2003). Second, it highlights that it is only under the implicit (and usually unstated) assumption that teachers are intrinsically motivated that increasing inputs should be expected to improve test scores. In contrast, if teachers behave like agents in standard economic models (with disutility of effort and no intrinsic utility from their job), then increasing inputs may lead to a *reduction* of effort and no change in learning, *even if* there are production function complementarities between inputs and teacher effort. Finally, if there are complementarities between effort and inputs in the production function, then providing incentives to teachers may *raise* the optimal level of effort when inputs are increased, giving rise to policy

[7] Prior studies have presented plausible *ex post* rationales for the lack of impact of additional resources including poor implementation, household substitution, and inputs being mis-targeted (such as providing textbooks to students who could not read). Our results suggest that these constraints may not bind if teachers are suitably motivated to use school resources better.

complementarities between providing inputs and incentives.

Formally, we model teachers' choice of effort (e) as solving the following problem:

$$\max_{e} U_i(e) = W + \lambda_i \Delta L - c_i(e) \tag{1}$$

subject to

$$W = S + b\Delta L \tag{1a}$$

$$\Delta L = f(e, I) \tag{1b}$$

$$\Delta L \geq \underline{\Delta L} \geq 0 \tag{1c}$$

where W is total earnings, which is equal to a base salary (S) plus a bonus ($b\Delta L$) proportional to gains in students' learning ΔL (b is typically zero in practice). λ_i is a measure of the teacher's intrinsic utility from improving student learning. Teacher effort, together with other inputs (I), translates into learning gains via f, which is strictly increasing in both arguments ($f_e > 0$ and $f_I > 0$), concave in each argument ($f_{ee} < 0$ and $f_{II} < 0$), and features complementarity between effort and inputs ($f_{eI} > 0$). Effort entails a cost, c_i, which is increasing and convex ($c_i'(\cdot) > 0$ and $c_i''(\cdot) > 0$). We allow λ_i and c_i to vary across teachers (indexed by i) to account for teacher heterogeneity. Finally, we assume that learning gains cannot be negative and have to be over a minimum level ($\underline{\Delta L}$). This can be interpreted as the minimum level of learning (including that taking place outside the school) required for teachers to not be sanctioned by parents or supervisors.[8]

Let $e_{min}(I)$ be the effort required to achieve $\underline{\Delta L}$ at a level of inputs equal to I (i.e., $f(e_{min}, I) = \underline{\Delta L}$). Let $e^*_{mc}(I)$ be the effort at which the marginal cost of effort is equal to its marginal benefit (i.e., $(\lambda_i + b) f_e(e^*_{mc}, I) = c_i'(e^*_{mc})$). Thus, the level of effort chosen will be $e^*(I) = \max(e_{min}(I), e^*_{mc}(I))$.

With the structure above, Figure 1a illustrates how the optimal level of teacher effort would vary with $b + \lambda_i$ at two different levels of inputs ($I_1 > I_0$). Figure 1b shows the corresponding learning gains. In the absence of incentives or intrinsic motivation (i.e., $b + \lambda_i = 0$), it is Equation 1c that binds, and $e^*(I) = e_{min}(I)$. Thus, if $b + \lambda_i = 0$, then the marginal cost of effort is above the marginal benefit in equilibrium.[9] Effort does not change as b increases up to the point where the marginal benefit ($b + \lambda_i$) is equal to the

[8] $\Delta L \geq \underline{\Delta L} \geq 0$ can also be motivated by intrinsic motivation considerations with teachers experiencing disutility if outcomes are too low. This is a variant of Holmstrom and Milgrom (1991) where teachers have a minimum outcome threshold as opposed to a minimum effort threshold below which they experience disutility. In this case, $\underline{\Delta L}$ would also vary by teacher.

[9] If $\lambda_i > 0$ the qualitative results do not change as long as λ_i is low enough that Equation 1c binds, leading to $e^*(I) = e_{min}(I)$.

marginal cost of providing effort. This corresponds to the flat region to the left of the thresholds κ_0 and κ_1 in Figure 1a.

In the absence of incentives and for low values of λ_i (such that $b + \lambda_i$ is near zero), an increase in inputs will lead teachers to re-optimize and decrease the effort they exert. The intuition is straightforward: if inputs increase, teachers can achieve the required minimum $\underline{\Delta L}$ with lower effort. This is consistent with evidence from multiple settings showing that teachers in developing countries reduce effort when provided with more resources.[10] Since the binding constraint for effort continues to be Equation 1c, the increase in inputs would lead to a reduction of effort to the point that allows $\underline{\Delta L}$ to be achieved, and there would be no net gain in learning as seen in Figure 1b.

Thus, in the absence of incentives for improving learning outcomes, the relationship between extra inputs and improved test scores will depend on the *distribution* of intrinsic motivation (λ_i) in the population of teachers. In settings where λ_i is high for most teachers, improving school inputs may improve test scores.[11] Increasing inputs lowers the threshold (from κ_0 to κ_1 in Figure 1a) that $b + \lambda_i$ needs to exceed for Equation 1c to not bind, and for effort to increase (because $f_{eI} > 0$). This is another channel through which increasing inputs could increase teacher effort and test scores (as seen in Figure 1a, where $\kappa_1 < \kappa_0$ when $I_1 > I_0$). However, in settings where λ_i is low for most teachers (such as in many developing countries with high levels of teacher absence), this may be less likely (since $\lambda_i + b = 0$ may still be below κ_1).

If additional inputs are combined with performance-linked pay that increases b, then the distribution of $b + \lambda_i$ is shifted to the right, and for any given distribution of λ_i it is more likely that teachers are shifted to the right of κ_1 and find it optimal to increase effort.[12] Further, as discussed above, to the right of κ_1, the optimal amount of effort is higher at higher levels of inputs (i.e., $e_I^*(I_1) > e_I^*(I_0)$ if $b + \lambda_i > \kappa_1$). Thus, as long as Equation 1c is not binding, the complementarity in the production function ($f_{eI} > 0$) will also yield complementarities in the policy effects.

[10] For instance, Duflo, Dupas, and Kremer (2015) find that providing a randomly selected set of primary schools in Kenya with an extra contract teacher led to an *increase* in absence rates of teachers in treated schools. Muralidharan and Sundararaman (2013) find the same result in an experimental study of contract teachers in India. Finally, Muralidharan, Das, Holla, and Mohpal (2017) show, using panel data from India, that reducing pupil-teacher ratios in public schools was correlated with an increase in teacher absence.

[11] For instance, Jackson, Johnson, and Persico (2016) find positive effects of school spending on education outcomes in the US, but default teacher effort in the US may be higher than in developing countries.

[12] While it is possible that the provision of incentives for performance may crowd out intrinsic motivation (Deci & Ryan, 1985; Fehr & Falk, 2002), it is also possible that the opposite is true and that incentives can crowd in intrinsic motivation by reinforcing the value of the task (Mullainathan, 2005). Empirical evidence from education in developing countries suggests that performance-based pay *increases* teachers' motivation (Muralidharan & Sundararaman, 2011a). We assume therefore that λ_i and b are additively separable.

We do not formally test the model above because intensity of teacher effort is difficult to measure accurately. We include the model to provide an intuitive and parsimonious framework to interpret our experiment and results, as well as existing results in the literature. Note also that teacher effort in the model need not be restricted to classroom effort. It can also include working with parents to provide inputs or effort at home.

3 Context and Interventions

3.1 Context

Our study is set in Tanzania, which is the sixth largest African country by population, and home to over 50 million people. Partly due to the abolishment of school fees in public primary schools in 2001, Tanzania has made striking progress towards universal primary education with net enrollment growing from 52% in 2000 to over 94% in 2008 (Valente, 2015). Yet, despite this increase in school enrollment, learning levels remain low. In 2012, nationwide learning assessments showed that less than one-third of grade 3 students were proficient at a grade 2 level in Kiswahili (the medium of instruction) literacy, or in basic numeracy. Proficiency in English (the medium of instruction in secondary schools) was especially limited, 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).

Despite considerable public spending on education,[13] budgetary allocations to education (and actual funds received by schools) have not kept pace with the rapid increases in enrollment. As a result, inadequate school resources are a widely-posited reason for poor school quality. In 2012 only 3% of schools had sufficient infrastructure (clean water, adequate sanitation, and access to electricity) and in grades 1, 2, and 3 there was only one math textbook for every five children (World Bank, 2012). Class sizes in primary schools average 74 students, with almost 50 students per teacher (World Bank, 2012).

A second challenge for education quality is low teacher motivation and effort. A study conducted in 2010 found that nearly one in four teachers were absent from school on a given day, and over 50% of teachers who were present in school were absent from the classroom (World Bank, 2012). The same study reported that on average, children receive only about 2 hours of instruction per day (less than half of the scheduled instructional time). Self-reported teacher motivation is also low: 47% of teachers surveyed in our data report that they would not choose teaching as a career if they could start over again.

[13] About one-fifth of overall Tanzanian government expenditure is devoted to the education sector, over 40 percent of which is allocated to primary education (World Bank, 2015).

3.2 Interventions and Implementation

The interventions studied in this paper were implemented by Twaweza, an East African civil society focusing on citizen agency and public service delivery. Through its Uwezo program, Twaweza has conducted large-scale independent citizen-led measurement of learning outcomes in East Africa from 2009 (see for example Uwezo (2017). Having documented the challenge of low levels of learning through the Uwezo program, Twaweza conducted extensive discussions with education stakeholders (including teachers' unions, researchers, and policy makers) and identified that the two most widely cited barriers to improving learning outcomes were inadequate school resources, and poor teacher motivation and effort.

Following this process, Twaweza formulated a program that aimed to alleviate these constraints and study their impact on learning outcomes. The program was called KiuFunza ("Thirst for learning" in Kiswahili) and was implemented in a representative sample of schools across Tanzania over two years (2013 and 2014). Twaweza also worked closely with government officials to ensure smooth implementation of the program and evaluation. The interventions are described below.

3.2.1 Capitation Grant (Grants) Program

Schools randomly selected for the capitation grants (CG) intervention received TZS 10,000 (∼US$6.25 at the time of the study) per student from Twaweza. For context, GDP/capita in Tanzania in 2013 was ∼US$1,000 and the per-student grant value was ∼0.6% of GDP/capita, a sizeable amount. While, the guidelines for expenditure were similar to that of the government's own capitation grant program, there were three key differences. First, the per capita Twaweza grant was larger than the average Government grant receipt.[14] Second, the Twaweza grants were sent directly to the school bank account to minimize diversion and leakage. Third, Twaweza communicated clearly with schools about the size of each tranche and expected date of receipt to enable better planning for optimal use of the resources.

Twaweza announced the grants early in the school year (March) during a series of meetings with school staff and community members, including parents and announced that the program would run for two years (2013 and 2014). Twaweza also distributed pamphlets and booklets that explained the program to parents, teachers, and community

[14]The value of the Twaweza grant was set at the official policy level. In practice, the average school received only around 60 percent of the government's stipulated grant value, and many received much less than that (World Bank, 2012). Reasons included inadequate budgetary allocations, diversion of funds for other uses by local governments, and delays in disbursements.

members. Funds were transferred to school bank accounts in two scheduled 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). Typically, head teachers and members of the school board decided how to spend the grant funds, but schools had to maintain financial records of their transactions and were required to share revenue and expenditure information with the community by displaying summary financial statements in a public area in the school.

Overall, Twaweza disbursed ∼US$350,000/year to the 70 schools in the Grant treatment arm, in effect fully implementing the Government's Capitation Grant policy. The size of the grants distributed to schools was ∼2-3 times the pre-treatment school-level spending per student (excluding teacher salaries and household spending), and the Grants treatment represented a significant increase in the resources available to schools.[15]

3.2.2 Teacher Performance Pay (Incentives) Program

The teacher performance pay program provided cash bonuses to teachers based on the performance of their students on independent learning assessments conducted by Twaweza. 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 of these subjects, an eligible teacher earned a TZS 5,000 (∼ US$3) bonus for each student who passed a simple externally administered, grade-appropriate assessment based on the national curriculum. Additionally, the head teacher was paid TZS 1,000 (∼ US$0.6) for each subject test a student passed.[16]

The term used by Twaweza for the teacher-incentive program was "Cash on Delivery (CoD)" to reinforce the contrast between the approaches that underlay the two programs — with the CG program being one of unconditional school grants, and the teacher incentive program being one where payments were contingent on outcomes.[17] The communication to schools and teachers emphasized that the aim of the CoD program was to motivate teachers and reward them for achieving better learning outcomes.

An advantage of the simple proficiency-based (or "threshold" based) incentive scheme used by Twaweza is its transparency and clarity. As pay-for-performance schemes are

[15]For example, if schools spent all of their grants on books, the funds would be sufficient to purchase about 4,000 textbooks per school (∼ 4-5/student), given the average grant size of ∼ US$5,000 per school.

[16]Twaweza included head teachers in the incentive design to make them stakeholders in improving learning outcomes. It is also likely that any scaled up teacher incentive program would also feature bonuses for head-teachers along the lines implemented in the KiuFunza project.

[17]Twaweza used the term CoD as a local version of a concept developed in the context of foreign aid by Birdsall et al. (2012).

relatively novel in Tanzania, Twaweza prioritized having a bonus formula that would be easy for teachers to understand. Bonuses based on passing basic tests of literacy and numeracy are also simpler to implement compared to more complex systems based on calculating student and teacher value added.

There are also important limitations to such a threshold-based design. It may encourage teachers to focus on students close to the passing threshold, neglecting students who are far below or far above the threshold (Neal & Schanzenbach, 2010). In addition, such a design may be unfair to teachers who serve a large fraction of students from disadvantaged backgrounds, who may be further behind the passing standard. While Twaweza was aware of these limitations, they took a considered decision to keep the formula simple in the interest of transparency, simplicity of explaining to teachers, and ease of implementation.[18] Further, since the bonuses were based on achieving basic functional literacy and numeracy, they were not too concerned about students being so far behind the threshold that teachers would ignore them.

Twaweza announced the program to teachers in March 2013 and explained the details of the bonus calculations to the head teacher and teachers of the target grades (1-3) and subjects (math, Kiswahili, and English). Pamphlets with a description of the bonus structure and answers to frequently asked questions were handed out to teachers, and booklets explaining program goals were distributed to parents. A follow-up visit in July 2013 reinforced the details of the program and provided an opportunity for questions and feedback. Teachers understood the program: over 90% of those participating in the program were able to correctly calculate the bonus level in a hypothetical scenario.

The high-stakes assessments that were used to determine the bonus payments were conducted at the end of the school year (with dates announced in advance), and consisted of three subject tests administered to all pupils in grades 1, 2 and 3. To ensure the integrity of the testing process, Twaweza created *ten* versions of the high-stakes tests, and randomly assigned these to students within a classroom. To prevent teachers from gaming the system by importing (or replacing) students, Twaweza only tested students enrolled at baseline (and took student photos at baseline to prevent identity fraud). Since each student enrolled at baseline had the potential to pass the exam, there would be no gains from preventing weaker students from taking the exam. All tests were conducted by and proctored by independent enumerators. Teacher bonuses were paid directly into their bank accounts or through mobile money transfers.

[18] In the US, the early years of school accountability initiatives such as No Child Left Behind focused on measures based on *levels* of student learning rather than value-addition for similar reasons.

3.2.3 Combination Arm

Schools assigned to the combination arm received *both* the capitation grant and teacher incentive programs discussed above with identical implementation protocols.

4 Research Design

4.1 Sampling and Randomization

We conducted the experiment in a *nationally representative* sample of 350 public schools across 10 districts in mainland Tanzania.[19] We first randomly sampled 10 districts from mainland Tanzania, and then randomly sampled 35 schools within each of these districts to get a sample of 350 schools (Figure 2). Within each district, 7 schools were randomly assigned to receive capitation grants, 7 schools to receive teacher incentives, and 7 schools to receive both grants and incentives. The remaining 14 schools did not receive either program and served as our control group.

4.2 Data

Our analysis uses several pieces of data collected from schools, teachers, students, and households over the course of the study. Enumerators collected data on school facilities, input availability, management practices, and school income and expenditure.[20] While most categories of school expenditure are difficult to map into specific grades, we collected data on textbook expenditures at the grade and subject level since this is a substantial expenditure item that can be easily assigned to a specific grade.

Enumerators also surveyed all teachers (about 1,500) who taught in focal grades (grades 1, 2, 3) and focal subjects (math, English and Kiswahili), and collected data on individual characteristics such as education and experience as well as effort measures such as teaching practices. They also conducted head teacher interviews.

For data on student learning outcomes, we sampled and tested 10 students from each focal grade (grades 1, 2 and 3) within each school, and followed these 30 students over the course of the study. We refer to these as low-stakes (or non-incentivized) tests as they are used purely for research purposes. From this set of 10,500 students, we randomly

[19]The combination of random assignment and representative sampling provides externally validity to our results across Tanzania (see Muralidharan and Niehaus (2017) for a more detailed discussion).

[20]Data on school expenditures were collected by reviewing receipts, accounting books, and other accounting records, following the expenditure tracking surveys developed and used by the World Bank (Reinikka & Smith, 2004; Gurkan, Kaiser, & Voorbraak, 2009)

sampled 10 students from each school (five from each of grades 2 and 3) to conduct household surveys. These 3,500 household surveys were used to collect information on household characteristics, educational expenditures, and non-financial educational inputs at the household (such as helping with homework).[21]

We also use data from the high-stakes (or incentivized) tests conducted by Twaweza that were used to determine teacher bonuses. These tests were taken by all students in grades 1, 2, and 3 in incentive and Combination schools (where bonuses had to be paid). Twaweza did not conduct these tests in Grant schools, but they did conduct them in a sample of 40 control schools to enable the computation of treatment effects of the incentive programs on the high-stakes tests. However, we only have student level test-scores from the second year of the evaluation as the Twaweza teams only recorded aggregated pass rates (needed to calculate bonus payments) in the first year.

Figure 3 presents a timeline of the project, with implementation related activities listed below the line, and research related activities above the line. The baseline survey was conducted in February 2013, followed by an endline survey (with low-stakes testing) in October 2013. The high-stakes tests by Twaweza were conducted in November 2013. A similar calendar was followed in 2014. The trial registry record and the pre-analysis plan are available at: https://www.socialscienceregistry.org/trials/291.

4.3 Summary Statistics and Validity

The randomization was successful and observable characteristics of students, households, schools, and teachers are balanced across our treatment arms; as are the normalized baseline test scores in each grade-subject (Table 1). Table 1 also provides summary statistics on the (representative) study population. The student gender ratio is balanced, and the average student is 9 years old (Panel A). The schools are mostly rural (85%), mean enrollment is ∼730, and class sizes are large – with an average of over 55 students per teacher (Panel C).[22] Teachers in our sample were ∼2/3 female, ∼40 years old, had ∼15 years of experience, and ∼40% of them did not have a teaching certificate (Panel D).

Attrition on the low-stakes tests conducted by the research team is balanced across treatment arms and is low — we were able to track around 90% of students in both years (last two rows of Table 1: Panel A). On the high-stakes tests, there is no differential

[21] Because most of the survey questions focused on educational expenditures, including expenditures in the previous school year, we did not survey first-grade students in the first year of the study as they were typically not attending school in the previous year. In the second year of the study, the second graders (the initial cohort of first graders) were sampled for the household survey.

[22] Thus, total enrollment in study schools was over 250,000 (350 x ∼730). Total enrollment in the focal grades for the study (grades 1, 2, and 3) was a little over 120,000 students.

student attendance in Incentive schools relative to the control group, but attendance in Combination schools was significantly higher (Table A.3). We therefore present bounds of treatment effects on high-stakes tests, using the approach of Lee (2009).

4.4 Empirical Strategy

Our main estimating equation for school-level outcomes takes the form:

$$Y_{sdt} = \alpha_0 + \alpha_1 Grants_s + \alpha_2 Incentives_s + \alpha_3 Combination_s + \gamma_d + \gamma_t + X_s\alpha_4 + \varepsilon_{sdt}, \quad (2)$$

where Y_{sdt} is the outcome of interest in school s in district d at time t. $Grants_s$, and $Incentives_s$ are indicator variables for school s receiving the capitation grant, and teacher incentive programs respectively. $Combination_s$ indicates if a school s received both programs. γ_d and γ_t are district (strata) and year fixed effects, and X_s is a set of school-level controls to increase precision. We use a similar specification to examine teacher-level outcomes. All standard errors are clustered at the school-level.

We use a similar estimating equation to study effects on learning outcomes:

$$Z_{isdt} = \delta_0 + \delta_1 Grant_s + \delta_2 Incentives_s + \delta_3 Combination_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_g + X_i\delta_4 + X_s\delta_5 + \varepsilon_{isd}, \quad (3)$$

where Z_{isd} is the normalized test score of student i in school s in district d at time t (normalized with respect to the control-group distribution on the same test). $Z_{isd,t=0}$ are normalized baseline test scores, γ_d and γ_g are district (strata) and grade fixed effects. X_i is a series of student characteristics (age, gender and grade), and X_s is a set of school and teacher characteristics. We also report robustness to dropping the school-level controls.

We focus on test scores in math, English, and Kiswahili as our primary outcomes, and also study impacts on science (not a focal subject) to test if gains in focal subjects were achieved at the cost of other subjects (multi-tasking). To mitigate concerns about the potential for false positives due to multiple hypothesis testing across subjects, we also create a composite summary measure of test scores, by taking the first component from a Principal Component Analysis (PCA) on the scores of the three subjects.

Since high-stakes tests were only conducted in incentive schools, combination schools, and a random set of 40 control schools, we cannot estimate the full comprehensive specification above. Furthermore, because the high-stakes exam is conducted only at the end of the year, we do not have baseline test scores or other student-level controls. Finally, student-level data on high-stakes tests were only available in the second year. As mentioned earlier, we prioritize results using low-stakes tests but present results on high-stakes tests to enable comparison with the literature.

For clarity of exposition and interpretation, we first present the impacts of the grant and incentive treatments individually (using only the intervention and the control group). We then present the impacts of all interventions estimated jointly, and test for complementarity: specifically, we test $H_0 : \delta_3 - \delta_2 - \delta_1 = 0$.

5 Results

5.1 Capitation Grant Program

5.1.1 How Were Grants Spent?

Table 2 presents descriptive statistics on how Grant schools spent their extra funds. Textbooks and classroom teaching aids (like maps, charts, blackboards, chalk, etc.) were the largest category of spending, jointly accounting for ∼65% of average spending over the two years. Administrative costs, including wages of non-teaching staff (e.g., cooks, janitors, and security guards) accounted for ∼27% of spending. Smaller fractions (∼7%) were allocated to student support programs such as meal programs, and very little (∼1%) was spent on construction and repairs. There were essentially no funds allocated to teachers, as stipulated by the program rules.[23]

Schools also saved some of the grant funds (∼20% and ∼40% of grant value in the first and second year). Since schools knew that the Grant program would end after two years, and government funding streams are uncertain (both in terms of timing and amount), we interpret this as "precautionary saving" and/or "consumption smoothing" behavior by schools (Sabarwal et al., 2014). The possibility of outright theft was minimized by the careful review of expenditures conducted by the Twaweza team (and the prior announcements that such audits would take place).

5.1.2 Did Grants Offset other Spending?

Table 3 examines the extent to which receiving the Grant program led to changes in other sources of spending. Column 1 summarizes the total extra spending from the capitation grant in grant schools. Schools that received Twaweza capitation grants saw a reduction in school expenditure from other sources (Column 2). Aggregating across both years, schools receiving the Grants program saw a reduction in other school spending

[23]Since teacher salaries are paid directly by the government, the capitation grant rules do not allow these funds to be used for teacher salaries. The Twaweza CG program had the same guidelines.

of ∼2,400 TZS per child, which is around a third of the additional spending enabled by the Grant program (Panel C - Columns 1 and 2).

Since average school spending per child in the control group was ∼5,200 TZS, spending the full grant value of 10,000 TZS would have tripled the school-level spending per child. After accounting for savings and offsetting reductions in school spending, there was still a significant net increase in total school spending per child of ∼4,700 TZS - almost double the expenditure relative to the control group (Panel C - Column 3).

Next, we examine changes in household spending. Column 4 shows the household offsets and Column 5 shows the total net per-child spending, accounting for both school and household spending. Consistent with the results documented by Das et al. (2013), we see an insignificant reduction in household spending by ∼1,000 TZS per child in the first year, and a larger significant reduction of ∼2,200 TZS per child in the second year (p=0.07).[24] These spending cuts were from assorted fees, textbooks, and food (Table A.2).[25] Taken together, the reductions in school and household spending attenuated the impact of the Twaweza grant on per-student spending, but did not fully offset it. On net, Grant schools saw a significant average increase in per-student spending of ∼3,100 TZS/year (Panel C, Column 5), a 60% increase over mean school-spending per student, enough to buy 3 textbooks per student per year.

5.1.3 Did Grants Improve Learning?

Despite the significant and meaningful increases in per-pupil funding discussed above, there was no difference in test scores between Grant and control schools in low-stakes tests of math, English or Kiswahili in either year of our study. Point estimates of impact on a composite measure of test scores were -0.03σ after one year and 0.01σ after two years (both insignificant; Table 4). Offsets are unlikely to be the main reason for our results, as we do not see any impacts of the grant on test scores even in the first year, when the net increase in spending per student in Grant schools was three times greater than in the second year (Table 3, Column 5). Overall, our results are consistent with and add to a large body of research that finds that merely increasing school resources rarely improves student learning outcomes in developing countries (including Glewwe et al. (2009) in Kenya, Blimpo et al. (2015) in Gambia, Das et al. (2013) in India, Pradhan et al. (2014) in Indonesia, and Sabarwal et al. (2014) in Sierra Leone).

[24] Das et al. (2013) posit that this is likely explained by the grants being unanticipated in the first year, and anticipated in the second one. Similar reasons may apply in our setting as well.

[25] Households spend ∼5 times more per child than schools. Nearly 70% of this spending is on uniforms, tutoring, and food - which are typically not covered by the school (see Table A.2 for details).

5.2 Teacher incentives

On the low-stakes tests administered by the research team, we find that test scores in Incentive schools are modestly higher than those in the control group, but typically not significant (Table 5: Panel A). The composite treatment effect at the end of the first year was 0.06σ (p=0.09), and at the end of two years it was 0.03σ (not significant).

However, students in Incentive schools were significantly more likely to pass the high-stakes Twaweza tests (the metric that bonuses were based on). At the end of two years, they were 37%, 17%, and 70% more likely to pass the Twaweza tests in math, Kiswahili, and English (all significant). These correspond to a 7.7, 7.3, and 2.1 percentage-point increase in the passing rate relative to the mean control group passing rate of 21%, 44%, and 3% in these subjects (Table A.1). Pass rates were also higher on all three subjects after the first year (though not significant in English). On normalized test scores, students in Incentive schools scored 0.17σ, 0.12σ, 0.12σ higher on math, Kiswahili, and English (all significant), and 0.21σ higher on the composite measure (Table 5: Panel B).[26].

We now consider possible possible reasons for the difference in estimated impacts across the two sets of tests. As mentioned in Section 3.2.2, Twaweza employed strict security protocols for the high-stakes test, including having *ten* different versions of the test paper that were randomized across students in the same class, and having independent proctors present for every test. So, the likelihood of cheating was minimized.

A second possibility is differences in test timing. On average, low-stakes tests were conducted $\sim$3 weeks before high-stakes test in both years. Since schools often conduct reviews and practice exams in this period, the superior performance on the high-stakes tests could reflect this additional preparation (which would have had to be more intense in the incentive schools). However, the performance on the low-stakes test does not seem to vary as a function of the number of days between the two tests (Table A.5).

A final possibility is differences in student effort and testing conditions across the two sets of tests. During the low-stakes test, only a small (but representative) sample of students were tested while the rest of the school functioned as if it were a regular school day. On the other hand, Twaweza intervention testing was conducted in a more visible manner, where all other non-academic school activities were canceled to allow all grade 1, 2, and 3 students to take the test in as quiet an environment as possible. In addition, many schools opted to use the Twaweza exams as the official end of year exam

[26]Note that we only have student-level data on the high-stakes tests in the second year. In the first year, Twaweza only recorded if students passed each test, which was the only metric needed to calculate teacher bonuses. Hence, we can estimate effects on passing the Twaweza test in both years, but can only calculate effects on normalized test scores in the second year

for grades 1, 2, and 3. Further, qualitative interviews suggest that teachers were more likely to have emphasized the importance of this test to students (since bonus payments depended on performance on these tests). Hence, students and teachers were likely to have been more motivated by the Twaweza exams.

Taken together, we conjecture that the main reason for the variation in estimated treatment effects is the differences in student effort and testing conditions across the two sets of tests. The estimated difference in the two sets of tests of 0.10-0.15σ, is exactly in line with recent experimental estimates that quantify the role of day of test student effort on measured test scores (Levitt, List, Neckermann, & Sadoff, 2016).

The demonstration that test-taking effort is a salient component of measured test scores by Levitt et al. (2016) presents a conundrum for education researchers as to what the appropriate measure of human capital should be for assessing the impact of education interventions. On one hand, low-stakes tests may provide a better estimate of a true measure of human capital that does not depend on external stimuli for performance. On the other hand, test-taking effort is costly, and students may not demonstrate their true potential under low-stakes testing, in which case, an 'incentivized' testing procedure may be a better measure of true human capital.

We focus on the low-stakes tests because these are the only tests that allow us to study the impact of *both* interventions and test for complementarities between them (since the high stakes tests were not carried out in Grant schools). Further, our pre-analysis plan prioritized the use of low-stakes tests.

Yet, given recent evidence on the importance of test-taking effort for measured test scores, and the fact that most existing studies of teacher incentives have reported results based on the high-stakes tests, some readers may prefer to focus on the estimates from the high-stakes tests (especially for cost-effectiveness calculations and comparing with existing studies). We therefore present both sets of results for completeness.

5.3 Combination of Capitation Grant and Teacher Incentives

After one year, relative to the control group, students in Combination schools scored .10σ higher on the low-stakes tests in all three focal subjects ($p < 0.05$ in all cases), and scored 0.12σ higher on the composite measure (Table 6-Panel A). After two years, they scored 0.20σ, 0.21σ, 0.18σ higher on math, Kiswahili, and English ($p < 0.01$ in all cases), and scored 0.23σ higher on the composite measure of learning (Table 6-Panel A).[27]

[27]These results include students who were only treated for one year (e.g., third graders in the first year of the program and first graders during the second year), and students who were treated in both years (e.g., first and second graders during the first year of the program). Appendix Table A.6 shows the results

Turning to the high-stakes test scores, at the end of the second year, students in Combination schools scored 0.25σ, 0.23σ, 0.22σ higher on math, Kiswahili, and English ($p < 0.01$ in all cases), and scored 0.36σ higher on the composite measure (Table 6-Panel B).[28] Pass rates (that bonuses were based on) were also higher. At the end of two years, students in Combination schools were 49%, 31%, and 116% more likely to pass the Twaweza-administered high-stakes test in math, Kiswahili, and English ($p < 0.01$ in all cases; (Table A.1). These correspond to a 10.3, 13.6, and 3.5 percentage-point increase relative to the control means of 21%, 44%, and 3%. Pass rates were also higher on all three subjects after the first year (though not significant in English).

Thus, regardless of whether we use the high-stakes tests (conducted by Twaweza) or the low-stakes tests (conducted by the research team), students in schools that received both programs had significantly higher test scores than those in control schools.

5.4 Complementarities Across Programs

Using the low-stakes tests (that were conducted in *all* schools), we find strong evidence of complementarities between the grant and incentive programs. Specifically, after two years, the impact under the Combination program is *significantly greater than the sum* of the impacts of the Grant and Incentive programs on their own, with this difference being significant for *every* subject and also for the composite measure of learning (α_4 in Table 6-Panel A). The point estimate for complementarities is also positive for all subjects after one year, but not always significant.

These complementarities are quantitatively important. Point estimates on the composite measure of learning for the Combination treatment are over three times the size of the sum of the impact of the Grant and Incentives treatments in the first year, and over five times greater in the second year. In short, school inputs appear to be quite effective when teachers have incentives to use them effectively, but not otherwise. Conversely, motivated teachers (either intrinsically or through incentives) can be much more effective with additional educational inputs.

While we cannot test for complementarities on the the high-stakes tests (because these were not conducted in Grant schools), we see suggestive evidence of similar complementarities here as well using two different approaches. First, if we assume that the impact of the Grant program on its own is zero (based on Table 4), then we can interpret

focusing on the panel of students who were exposed to the interventions in both years. We find very similar results among this group.

[28] Due to the differential attendance rates between Combination and control schools on the high-stakes tests (Table A.3), we estimate Lee (2009) bounds on the treatment effects and find that the treatment effect is still positive and significant for every subject as well as the composite measure of learning (Table A.4).

the significant difference on the high-stakes tests between Combination and Incentive schools as evidence of complementarities (β_5 in Table 6-Panel B).[29] A second approach is to compare the difference between Combination and Incentive schools (which reflects the impact of the "Grant" *and* the "complementarities") on both the high-stakes and low-stakes tests. We cannot reject that this difference is zero (β_5 - α_5 in last row of Table 6-Panel C). In other words, the estimated effects of the "Grant plus complementarities" are similar across the low- and high-stakes tests.

The experimental evidence of complementarities across education policies is our most important and original result. This has (to the best of our knowledge) not been shown experimentally to date, though there is suggestive prior evidence of complementarity between teacher incentives and inputs in prior work. For instance, Muralidharan and Sundararaman (2011b) and Muralidharan (2012) find greater impact of teacher performance pay in cases where teachers have higher education and training, suggesting complementarity between inputs (teacher knowledge) and incentives. More recently, Gilligan et al. (2018) conduct a randomized evaluation of a teacher performance pay program in Uganda and find that there was no impact on learning in schools that had no textbooks, but that there was a significant positive impact in schools with textbooks (consistent with our findings in neighboring Tanzania).

Yet, this prior evidence is only suggestive because teacher education and training, or textbooks are not randomly assigned and may be correlated with other omitted variables. In contrast, the current study features random assignment of *both* treatments *and* their interaction, and is explicitly powered to test for complementarities. This allows us to experimentally demonstrate the presence and importance of complementarities among education policies - especially between input and incentive based policies (as also suggested by our theoretical framework).

5.5 Multi-tasking and Diversion of Effort/Resources

An important concern with teacher performance-pay schemes is the risk that such programs will encourage teachers to focus on incentivized subjects at the cost of other subjects or activities; a classic case of the multi-tasking problem (Holmstrom & Milgrom, 1991). On the other hand, if programs that reward gains in math and language are able to improve literacy and numeracy skills, they may promote student learning even in other non-incentivized subjects. Thus, the impact of performance-pay on non-incentivized outcomes will depend on the extent to which the effort needed to improve incentivized

[29] Note that this difference is significant even after Lee-bounds based adjustment of confidence intervals for differential attrition (β_4 in Table A.4)

and non-incentivized outcomes are complements or substitutes (see Muralidharan and Sundararaman (2011b) for a more detailed discussion).

We test for these possibilities by looking at impacts on science, a non-incentivized subject that was included in our battery of low-stakes student assessments. Results on science are consistent with those on the other subjects, with no impact in the Grant and Incentives treatments, and positive impacts in Combination schools (Table 7). Further, mirroring the patterns we see on the incentivized subjects, we find evidence of complementarities between grants and incentives in science learning in the second year. Overall, the results suggest that teacher incentives on math and language in this setting did not hurt learning in other subjects, and may have even helped it when the gains in math and language were significant (as was the case in Combination schools).

In the case of the capitation grant program, the value of the school grant was based on the total enrollment across all grades (with the same per-student value of 5,000 TZS). However, it is possible that schools may have spent the funds unequally across grades. In particular, since performance on the grade 7 primary-school exit exam is an externally salient metric that governments and parents focus on, schools may have chosen to divert some of the grant to students in later grades (especially grade 7). We test for cross-grade diversion/spillovers by examining impacts on student performance on the Primary School Leaving Examination (PSLE) taken by students in Grade 7, and find no evidence of any impact of our treatments on this metric, both in terms of average scores or pass rates (Table 7-Columns 3-6). Consistent with the incentive program not being implemented outside grades 1-3 (though the grants were provided to all grades), we find no evidence of complementarities between interventions in the grade 7 outcomes.

5.6 Mechanisms

We report impacts on teacher effort using survey-based measures of teacher attendance, and teacher self-reports. For the most part, we do not detect meaningful impacts on these outcomes (Table 8). Teacher absence rates are unchanged (consistent with Muralidharan and Sundararaman (2011b)), and we find little systematic evidence of impact on self-reported data on the number of tests given, or provision of remedial teaching. Teachers in both Incentive and Combination schools are more likely to report providing extra tutoring, but the coefficient is only significant for Combination schools. However, the intensity of teaching effort is difficult to measure well through surveys and observations. Further, given the high cost of data collection, we prioritized collecting data on

expenditure and outcomes rather than teaching activities.[30]

A different way of measuring teacher effort (especially between Combination and Grant schools) is to see if they were effective at *reducing the offset* of school and household resources. We test for this possibility and report the results in Table 9. In both years, we see that the increase in net expenditure (Table 9–Column 5) was higher in Combination schools than in the Grant schools. The contrast is stronger in the second year, when parents in grant schools cut back their spending, whereas there are no parental offsets in combination schools (p =0.11; last row of Panel B, Column 4). This is consistent with increases in (unobservable) teacher (and head teacher) effort in combination schools. In particular, teachers (and head teachers) could lobby and encourage parents to continue to financially support their children.

Further evidence of mechanisms is seen in the patterns of textbook expenditures. Table 10 compares school expenditure on textbooks for students in Grades 4 to 7 (non-incentivized grades) versus those in Grades 1 to 3 (the incentivized grades) across both Grant and Combination schools. Consistent with receiving extra resources, textbook expenditures increased across all grade groups in both grant and combination schools (but not in Incentive schools). However, Grant schools spent nearly 40% more on textbooks in higher grades, while Combination schools spent similar amounts across all grades (Column 3). Testing for equality in the differences in relative spending across the treatments, we find that Combination schools spent significantly more per student (543 TZS) on textbooks in incentivized grades (relative to non-incentivized grades) compared to schools that only received the Grants (p <0.05).

Overall, while our direct measures of teacher effort are limited, the indirect evidence from patterns of expenditure across Grant and Combination schools suggests that teachers in Combination schools may have exerted more effort to ensure that an increase in resources translated into improvements in learning as well (for incentivized grades).

5.7 Heterogeneity

We examine heterogeneity of program impacts by non-parametrically plotting treatment effects as a function of baseline test scores (which are a good summary statistic of all prior inputs into human capital creation). As a summary measure, we focus on the composite measure of human capital across subjects, using the low-stakes tests (since

[30]For instance, we did not conduct classroom observations. In addition to cost, this decision was also informed by prior work showing considerable Hawthorne effects in measuring teacher classroom behavior (Muralidharan & Sundararaman, 2010), rendering such measures unreliable for measuring treatment effects on teacher effort.

these are the tests for which we have baseline scores). We show results separately by treatment and year, with bootstrapped 95% confidence intervals around the estimated treatment effect at each percentile of the baseline test-score distribution (Figure 4).

Consistent with the overall zero effects in Grants schools, we find no significant effect at any part of the baseline test-score distribution, though weaker students seem to have benefited more in the second year. Students in Incentive schools scored higher than those in control schools at nearly all points in the baseline distribution, but effects are typically not significant. Finally, students in Combination schools did better than those in the control schools at every point in the baseline score distribution, with the effects being significant at all points in the distribution in the second year.

Since the incentive formula rewarded teachers based on the number of students who passed a threshold, teachers in Incentive and Combination schools may have focused more on students near the passing threshold (as shown by Neal and Schanzenbach (2010) in the US). We therefore test for heterogeneity of effects as a function of distance of student test-scores from the passing threshold. Since the passing score varies by grade, and subject, we define the "distance from the threshold" as the absolute value of the difference in a students' own percentile and the percentile of the passing threshold (this allows us to pool across grades and subjects for power). Overall, we find no evidence of differential treatment effects as a function of either the average or the square of distance from the passing threshold and report the results in Table A.8).[31]

Next, we test for heterogeneity by student, teacher, and school characteristics using Equation 2, and adding interactions of the treatment with each covariate. As above, we use the low-stakes tests, and focus on the composite index of test scores. The interaction coefficients of interest are reported in Table 11, with columns 1-3, 4-6, and 7-9 focusing on heterogeneity by student, teacher, and school characteristics respectively.

Overall, the treatments seem to have helped disadvantaged students more. In Combination schools (where treatment effects are positive and significant), girls, and those with lower initial test scores gain more. Results are not as robust for the Grant and Incentive schools, but are broadly consistent (columns 1-3). We find little evidence of heterogeneity by measures of teacher age, gender, or salary (columns 4-6), and some suggestive evidence of heterogeneity by school characteristics (columns 7-9). On the lat-

[31]This is a robust result. Since this was a dimension on which we expected to find some heterogeneity (as seen in our pre-analysis plan), we tested for this possibility using several possible functional forms and definitions of "distance from the passing threshold", but we never reject the null of no heterogeneity along this dimension. This result validates Twaweza's hypothesis (which informed the design of the Incentive program) that differential targeting of students by teachers was unlikely given the very low absolute levels of learning seen in this setting and the modest gains needed to achieve a passing score.

ter, schools scoring higher on an index of facilities show higher gains when they receive teacher incentives (Column 7). This is consistent with our experimental findings on the complementarities of resources and incentives.

We also find suggestive evidence of greater effects of receiving school grants (in both Grant and Combination schools), when schools are better managed (as measured by a management practices survey administered to the head teacher). These results are consistent with growing recent evidence on the importance of school management in the education production function (see Bloom, Lemos, Sadun, and Van Reenen (2015); Lemos, Muralidharan, and Scur (2018)). They are also consistent with our theoretical framework (with better management proxying for higher baseline levels of effort). However, since we did not pre-specify this hypotheses, we simply report the results for completeness and leave it to future work to explicitly test for complementarities between management quality and school resources.

5.8 Cost Effectiveness

The cost of the capitation grant program including the administrative cost of transferring the money and conducting the audits was 7.13 USD per student. The cost of the teacher incentive program, inclusive of the administrative cost of implementing the program and testing all the students was 7.10 USD per student. Finally, the cost of the Combination program was 13.29 USD per student.[32] All estimates of costs include both the direct costs (value of grants and incentives) as well as the implementation costs (test design and implementation, communications, audit, etc.) of each program. Table A.9 provides a breakdown of the direct and implementation costs of all three programs.

Our results using low-stakes tests suggest that neither the Grant nor Incentive programs were effective on their own, and that only the Combination program was effective (and hence cost effective). In Combination schools, we estimate that the cost of increasing test scores by 0.1σ per student was USD 5.78.

We also perform cost-effective calculations using estimated treatment effects from the high-stakes exams for comparability with existing studies. Using these estimates, the cost of increasing test scores by 0.1σ per student was USD 3.38 in Incentive schools and USD 3.69 in Combination schools. The similarity in cost effectiveness, despite the complementarities between inputs and incentives, is driven by the fact that the larger test score gains in Combination schools also led to larger bonus payments.

[32]The Combination program's cost is not equal to the sum of the cost of Grant and Incentives programs since there were some administrative economies of scale in implementing the programs together .

A bonus is a different way of compensating teachers. Hence, in the medium-term, it may be possible to implement teacher incentive programs at a lower cost by doing so in the context of regular salary increases. For instance, a scheduled across-the-board 10% increase in teacher salaries could be replaced with a 5% across-the-board increase and a further 0-10% increase based on performance.[33] In such a scenario, the main long-term cost of a teacher incentive program is the administrative cost of implementing the program (including costs of independent measurement and recording of student learning) and *not* the cost of the bonus itself.[34] Using the administrative costs in this study, the cost of increasing test scores by 0.1σ per student would be USD 2.18 in Incentive schools and USD 1.27 in Combination schools.

Overall, these estimates compare well with the estimated cost effectiveness of several other interventions to improve education in Africa. For instance, some of the interventions with positive impacts on learning reviewed by Kremer, Brannen, and Glennerster (2013) include: a conditional cash transfer in Malawi, with a cost of USD 100 per 0.1σ gain per student (Baird, McIntosh, & Özler, 2011); scholarships for girls in Kenya, with a cost of USD $7.14/0.1\sigma$ (Kremer, Miguel, & Thornton, 2009); contract teachers and streaming in Kenya, with a cost of USD $5/0.1\sigma$ (Duflo et al., 2015; Duflo, Dupas, & Kremer, 2011); and teacher incentives in Kenya (evaluated using data from high-stakes tests), with a cost of USD $1.59/0.1\sigma$ (Glewwe et al., 2010).[35] Thus, the only program more cost effective than the ones we study here was also a teacher-incentive program. In addition, many education interventions have either zero effect or provide no cost data for cost-effectiveness calculations (Evans & Popova, 2016).

Taken together, our results suggest that reforms to teacher compensation structure that reward improving student learning can be highly cost effective relative to the status quo of education spending, that is largely input-based. Further, our results on complementarity between input and incentive policies suggest that such reforms may also improve the effectiveness of existing school resources. Since the default approach to education in most developing countries is based on providing more school inputs, the marginal returns to introducing performance-based pay for teachers may be particularly high.[36]

[33]Such an approach may be especially promising to consider because typical across-the-board teacher salary increases are unlikely to have any positive impact on the effectiveness of incumbent teachers as shown recently by de Ree, Muralidharan, Pradhan, and Rogers (2018).

[34]We abstract away from a risk-aversion premium that may need to be paid, because this will be second order for small spreads in pay and typical values of risk-aversion parameters.

[35]We use up to date numbers released in a standardized template by The Abdul Latif Jameel Poverty Action Lab at https://www.povertyactionlab.org/policy-lessons/education/increasing-test-score-performance. Note also, that we only include estimates from peer-reviewed published studies.

[36]Note that the 2x2 experimental design is only needed to *identify* complementarities by ensuring that both policies are changed exogenously. From a policy perspective, if status quo spending on inputs is

6 Conclusion

We report findings from a large randomized controlled trial conducted across a representative sample of 350 Tanzanian schools and over 120,000 students that studied the impact of three different programs to improve learning in early grades. These included unconditional school grants to alleviate school resource constraints; bonus payments to teachers based on student learning outcomes to improve teacher motivation and effort; and both of the above. Consistent with the existing evidence, we find that merely increasing school resources via school grants does little to improve learning outcomes. Also consistent with prior evidence from developing countries, the teacher incentive program led to improvements in student learning (but only on high-stakes tests). Test scores in schools that received both programs were significantly higher on both high-stakes and low-stakes tests. Moreover, we find strong evidence of complementarities between inputs and incentives with the effect of providing both being significantly greater than the sum of the individual effects.

The evidence of complementarities suggests that there may be multiple binding constraints to improving human development outcomes in developing countries. In such a setting, policies that alleviate some constraints but not others may have a limited impact on outcomes. This point is exemplified by the large and growing body of evidence on the limited impact on learning outcomes of simply providing more resources (and reinforced by our results on the Grant program). At the same time, our results highlight that these additional resources *can* significantly improve outcomes if accompanied by improved incentives to use them effectively.

Conversely, even well-motivated staff may not be able to deliver services effectively if they lack even the basic resources to do so. The positive effects of Incentives on their own (on the high-stakes tests) are consistent with schools having at least some resources to work with. But the complementarity with Grants clearly points to the fact that a lack of resources could be a binding constraint to quality improvement for motivated teachers.[37]

Our results may be relevant for the design of development interventions more generally. Cross-country evidence suggests that foreign aid (inputs) may be more effective in countries with more growth-friendly policies (a proxy for likelihood of using resources well) (Burnside & Dollar, 2000), but these results are not very robust (Easterly, Levine,

high, and on incentives is zero, the marginal return of improving the latter will be higher.

[37] Indeed, one reason for why many senior policy makers may genuinely believe that resource constraints are binding is that officials who have been promoted and risen to the top of their institutional hierarchies are more likely to have higher intrinsic motivation. It is thus more likely that the binding constraints for these officials are resources and not motivation.

& Roodman, 2004). Our results finding no impact of inputs on their own, and strong complementarities between inputs and incentives provides well-identified evidence of the (Burnside & Dollar, 2000) hypothesis in the context of a sector (education), that accounts for a sixth of developing country government spending (World Bank, 2015) and over fifteen billion dollars of aid spending annually (OECD, 2016).

Finally, we note that the default pattern of social sector spending in most countries (and also in donor led development assistance programs) is to expand school inputs. These include both physical inputs (like infrastructure and books), and large programs focused on teacher training and capacity building. Our results show that the marginal returns of introducing reforms to better reward improved teacher effort and student learning may be particularly high in settings where inputs are being expanded. Of course, implementing teacher performance-pay systems will require investments in implementation capacity, but our estimates suggest that this could be a cost-effective investment and that doing so may meaningfully expand state capacity for improved service delivery in developing countries.[38]

References

Almond, D., & Mazumder, B. (2013). Fetal origins and parental responses. *Annual Review of Economics*, *5*(1), 37-56.

Attanasio, O. P., Fernández, C., Fitzsimons, E. O. A., Grantham-McGregor, S. M., Meghir, C., & Rubio-Codina, M. (2014). Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in colombia: cluster randomized controlled trial. *BMJ*, *349*.

Baird, S., McIntosh, C., & Özler, B. (2011). Cash or condition? evidence from a cash transfer experiment. *The Quarterly Journal of Economics*, *126*(4), 1709–1753.

Bandiera, O., Burgess, R., Das, N., Gulesci, S., Rasul, I., & Sulaiman, M. (2017). Labor markets and poverty in village economies. *The Quarterly Journal of Economics*, *132*(2), 811–870.

Banerjee, A., & Duflo, E. (2005). Chapter 7 growth theory through the lens of development economics. In P. Aghion & S. N. Durlauf (Eds.), (Vol. 1, p. 473 - 552). Elsevier.

[38]Since the integrity of measurement may be compromised if implemented through the government itself, one viable option for scaling up the implementation of performance-pay programs in developing countries may be for governments to partner with committed and credible local third-party organizations (like Twaweza) to conduct the independent measurements on the basis of which performance-pay schemes can be implemented.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., ... Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348*(6236), 1260799.

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in mexican high schools. *Journal of Political Economy*, *123*(2), 325-364.

Birdsall, N., Savedoff, W. D., Mahgoub, A., & Vyborny, K. (2012). *Cash on delivery: a new approach to foreign aid*. Center for Global Development.

Bleakley, H. (2010). Health, human capital, and development. *Annual Review of Economics*, *2*(1), 283-310.

Blimpo, M. P., Evans, D. K., & Lahire, N. (2015). *Parental human capital and effective school management : Evidence from the gambia* (Policy Research Working Paper No. 7238). World Bank.

Bloom, N., Lemos, R., Sadun, R., & Van Reenen, J. (2015). Does management matter in schools? *The Economic Journal*, *125*(584), 647–674.

Burnside, C., & Dollar, D. (2000). Aid, policies, and growth. *The American Economic Review*, *90*(4), 847–868.

Calefati, J. (2016). *Dozens of california districts with worst test scores excluded from extra state help.* Retrieved 2018-05-05, from https://calmatters.org/articles/dozens-california-districts-worst-test-scores-excluded-extra-state-help/

Collier, K. (2016). *Lawmakers look at tying school funding to performance.* Retrieved 2018-05-05, from https://www.texastribune.org/2016/08/03/senators-examining-performance-based-funding-schoo/

Contreras, D., & Rau, T. (2012). Tournament incentives for teachers: Evidence from a scaled-up intervention in chile. *Economic Development and Cultural Change*, *61*(1), 219-246.

Cunha, F., & Heckman, J. (2007, May). The technology of skill formation. *American Economic Review*, *97*(2), 31-47.

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.

Deci, E., & Ryan, R. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer US.

de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. (2018). Double for nothing? experimental evidence on an unconditional teacher salary increase in indonesia. *The Quarterly Journal of Economics*, *133*(2), 993-1039.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, *101*(5), 1739-74.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*, *123*, 92–110.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, *102*(4), 1241–1278.

Easterly, W., Levine, R., & Roodman, D. (2004). Aid, policies, and growth: Comment. *The American Economic Review*, *94*(3), 774–780.

Evans, D., & Popova, A. (2016). What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews. *The World Bank Research Observer*, *31*(2), 242–270.

Fehr, E., & Falk, A. (2002). Psychological foundations of incentives. *European economic review*, *46*(4), 687–724.

Ganimian, A. J., & Murnane, R. J. (2014, July). *Improving educational outcomes in developing countries: Lessons from rigorous evaluations* (Working Paper No. 20284). National Bureau of Economic Research.

Geng, T. (2018). *The complementarity of incentive policies in education: Evidence from new york city* (Working Papers). Columbia University.

Gilligan, D. O., Karachiwalla, N., Kasirye, I., Lucas, A., & Neal, D. (2018, May). *Educator Incentives and Educational Triage in Rural Primary Schools* (IZA Discussion Papers No. 11516).

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

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? textbooks and test scores in Kenya. *American Economic Journal: Applied Economics*, *1*(1), 112–35.

Glewwe, P., & Muralidharan, K. (2016). Chapter 10 - improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In S. M. Eric A. Hanushek & L. Woessmann (Eds.), (Vol. 5, p. 653 - 743). Elsevier.

Gurkan, A., Kaiser, K., & Voorbraak, D. (2009). *Implementing public expenditure tracking surveys for results: lessons from a decade of global experience* (PREM Notes; No. 145).

Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences*, *104*(33), 13250–13255.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7*,

24–52.

Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms *. *The Quarterly Journal of Economics*, *131*(1), 157-218.

Johnson, R. C., & Jackson, C. K. (2017, June). *Reducing inequality through dynamic complementarity: Evidence from head start and public school spending* (Working Paper No. 23489). National Bureau of Economic Research.

Johnston, B. F., & Mellor, J. W. (1961). The role of agriculture in economic development. *The American Economic Review*, *51*(4), 566-593.

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*.

Kerwin, J. T., & Thornton, R. L. (2017). *Making the grade: The trade-off between efficiency and effectiveness in improving student learning* (Working Paper). University of Minnesota.

Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: Some lessons. *The American Economic Review*, *93*(2), pp. 102-106.

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

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, *91*(3), 437–456.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy*, *110*(6), 1286–1317.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review*, *99*(5), 1979-2011.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*(3), 1071–1102.

Lemos, R., Muralidharan, K., & Scur, D. (2018). *Personnel management and school productivity: Evidence from india* (Working Paper). University of California, San Diego.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, *8*(4), 183–219.

Malamud, O., Pop-Eleches, C., & Urquiola, M. (2016, March). *Interactions between family and school environments: Evidence on dynamic complementarities?* (Working Paper No. 22112). National Bureau of Economic Research.

Mbiti, I. (2016). The need for accountability in education in developing countries. *Journal of Economic Perspectives*, *30*(3), 109–32.

Mesecar, D., & Soifer, D. (2016). *How performance-based funding can improve educa-*

tion funding. Retrieved 2018-05-05, from https://www.brookings.edu/blog/brown-center-chalkboard/2016/02/24/how-performance-based-funding-can-improve-education-funding/

Mullainathan, S. (2005). Development economics through the lens of psychology. In *Annual world bank conference on development economics 2005: Lessons of experience.*

Muralidharan, K. (2012). *Long-term effects of teacher performance pay: Experimental evidence from india* (Working Paper). University of California, San Diego.

Muralidharan, K., Das, J., Holla, A., & Mohpal, A. (2017). The fiscal cost of weak governance: Evidence from teacher absence in india. *Journal of Public Economics*, *145*, 116–135.

Muralidharan, K., & Niehaus, P. (2017). Experimentation at scale. *Journal of Economic Perspectives*, *31*(4), 103–24.

Muralidharan, K., Romero, M., & Wuthrich, K. (2018). *Factorial designs, model selection, and (incorrect) inference in experiments* (Working Paper). University of California, San Diego.

Muralidharan, K., & Sundararaman, V. (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from india. *Economic Journal*, *120*, F187–F203.

Muralidharan, K., & Sundararaman, V. (2011a). Teacher opinions on performance pay: Evidence from india. *Economics of Education Review*, *30*(3), 394–403.

Muralidharan, K., & Sundararaman, V. (2011b). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*(1), 39–77.

Muralidharan, K., & Sundararaman, V. (2013, September). *Contract teachers: Experimental evidence from india* (Working Paper No. 19440). 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.

OECD. (2016). *Education-related aid data at a glance.* (data retrieved from, http://www.oecd.org/dac/financing-sustainable-development/development-finance-data/education-related-aid-data.htm and https://stats.oecd.org/Index.aspx?QueryId=58197)

Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., & Artha, R. P. (2014, April). Improving educational quality through enhancing community participation: Results from a randomized field experiment in indonesia. *American Economic Journal: Applied Economics*, *6*(2), 105-26.

Ray, D. (1998). *Development economics*. Princeton University Press.

Reinikka, R., & Smith, N. (2004). *Public expenditure tracking surveys in education*. UNESCO, International Institute for Educational Planning.

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

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485).

United Nations. (2015). Transforming our world: The 2030 agenda for sustainable development. *Resolution adopted by the General Assembly*.

Uwezo. (2013). *Are our children learning? numeracy and literacy across east africa* (Uwezo East-Africa Report). Nairobi: Uwezo. (Accessed on 05-12-2014)

Uwezo. (2017). Are our children learning? *Uwezo Tanzania Sixth Learning Assessment Report. Dar es Salaam: Twaweza East Africa*.

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

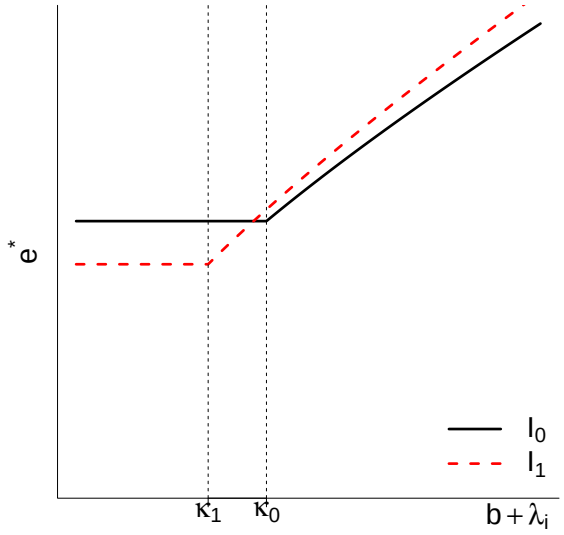
World Bank. (2012). *Tanzania service delivery indicators* (Tech. Rep.). Washington D.C.: World Bank.

World Bank. (2015). *Expenditure on primary as % of government expenditure on education (%).* (data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SE.XPD.PRIM.ZS?locations=TZ)

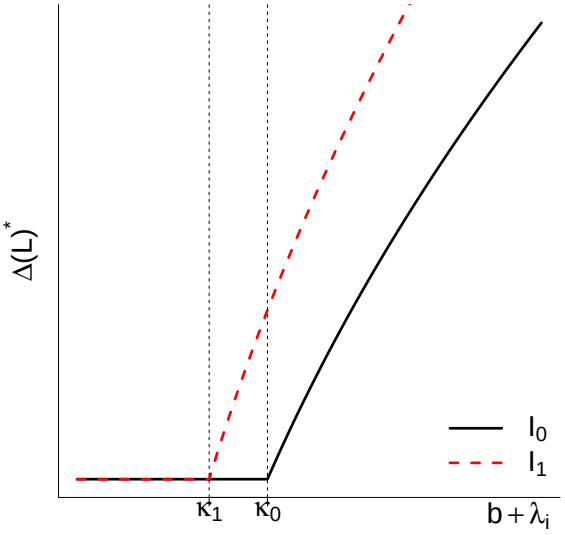
World Bank. (2017). *Education statistics (edstats).* (data retrieved from, http://datatopics.worldbank.org/education/wDashboard/dqexpenditures)

World Bank. (2018). *World development report 2018: Learning to realize education's promise.* The World Bank. Retrieved from http://www.worldbank.org/en/publication/wdr2018

Figure 1: Effort and learning as a function of motivation, at different levels of inputs



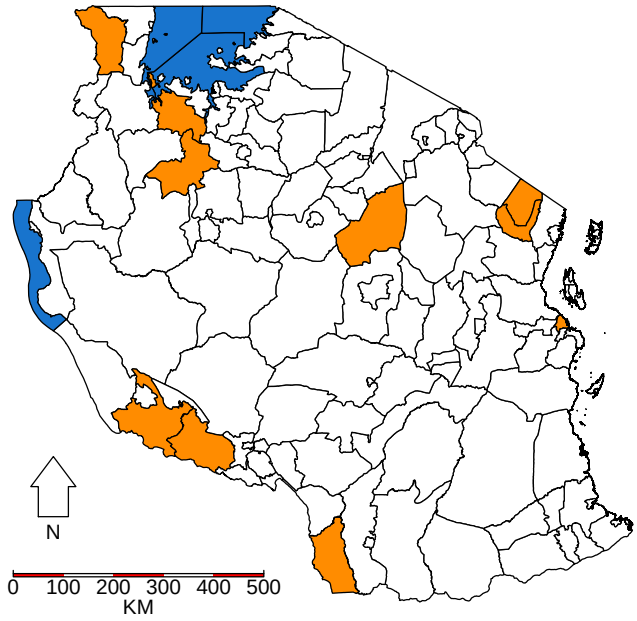
(a)



(b)

Note: Figures 1a and 1b show how teacher's chosen level of effort (e^*) and the learning that results from this level of effort (ΔL^*) vary for different values of $b + \lambda_i$, across two levels of inputs ($I_1 > I_0$). In both figures $f(e, I) = \ln(e) + \ln(I) + e \cdot I$, $c_i(e) = e^2$, $I_0 = 1$, $I_1 = 1.2$, $\underline{\Delta L} = 0$, and $b + \lambda_i \in (0, 1)$. κ_c is the threshold at which the constraint in Equation 1c is no longer binding for input level I_c, and therefore $e^*(I_c) = e^*_{mc}(I_c)$ to the right of κ_c.

Figure 2: Districts in Tanzania from which schools were selected



Note: We drew a nationally representative sample of 350 schools from a random sample of 10 districts in Tanzania.

Figure 3: Timeline

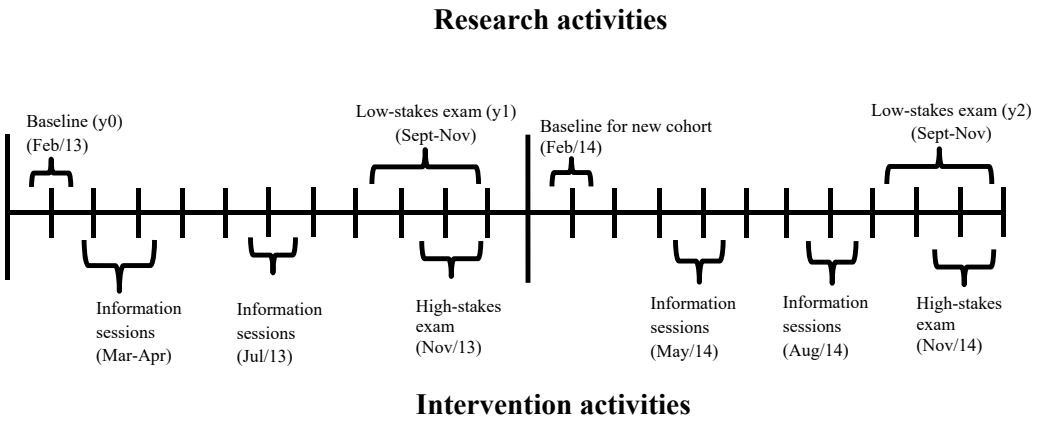
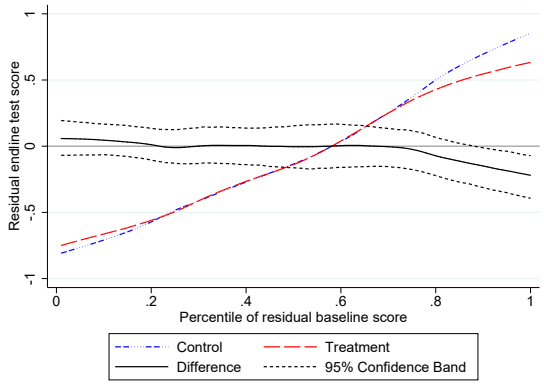
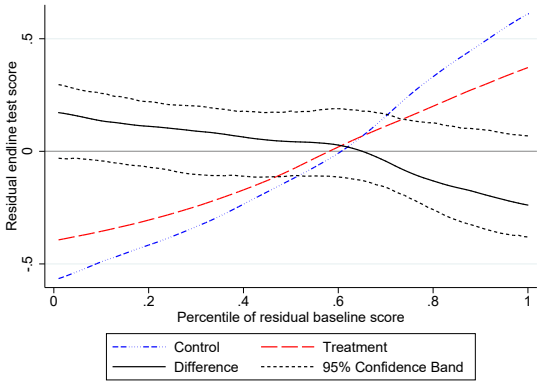


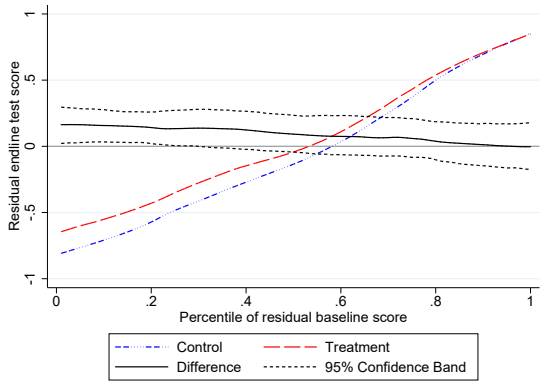
Figure 4: Non-parametric treatment effects by percentile of baseline score (low-stakes)



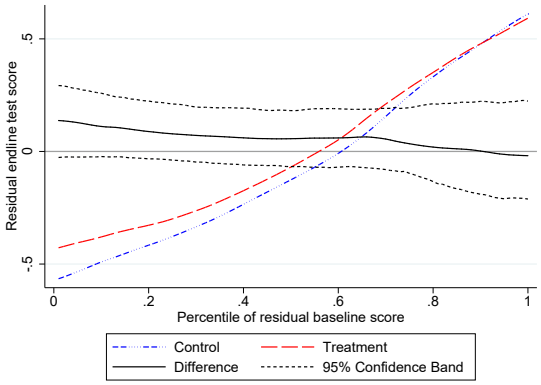
(a) Inputs - Year 1



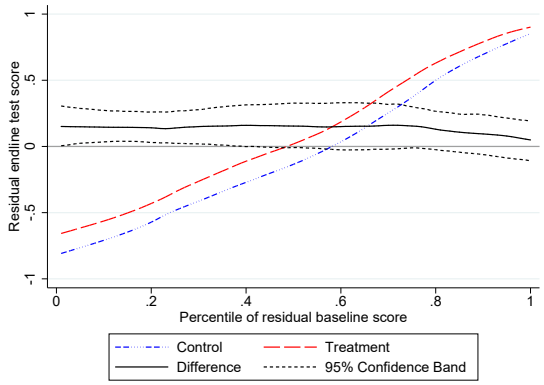
(b) Inputs - Year 2



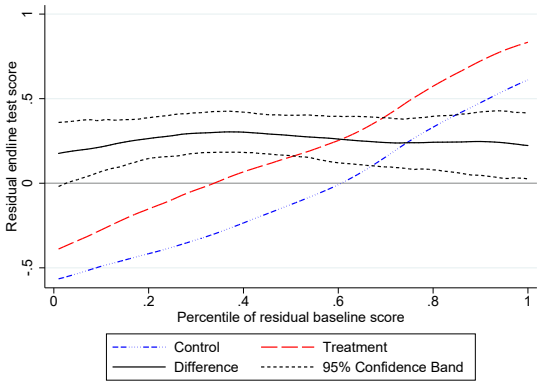
(c) Incentives - Year 1



(d) Incentives - Year 2



(e) Combination - Year 1



(f) Combination - Year 2

Note: These treatment and control lines are estimated using local linear regressions. The pointwise treatment effect is calculated as the difference. The 95% confidence intervals are estimated using bootstrapping. The x-axis is the percentile of the residual of a regression of a PCA index of the student's test score across all subjects at baseline on student and school characteristics. The y-axis is the residual of a regression of a PCA index of the student's test score across all subjects at each follow-up on student and school characteristics.

Table 1: Summary statistics across treatment groups at baseline (February 2013)

	(1) Combination	(2) Grants	(3) Incentives	(4) Control	(5) p-value all equal
Panel A: Students (N=13,996)					
Male	0.50	0.49	0.50	0.50	0.99
	(0.01)	(0.01)	(0.01)	(0.01)	
Age	8.94	8.96	8.94	8.97	0.94
	(0.05)	(0.05)	(0.05)	(0.04)	
Normalized Kiswahili test score	0.05	-0.02	0.06	0.00	0.41
	(0.07)	(0.07)	(0.08)	(0.05)	
Normalized math test score	0.06	0.01	0.06	0.00	0.59
	(0.06)	(0.06)	(0.07)	(0.05)	
Normalized English test score	-0.02	-0.02	-0.00	0.00	0.91
	(0.04)	(0.05)	(0.05)	(0.04)	
Attrited in year 1	0.13	0.13	0.11	0.13	0.21
	(0.01)	(0.01)	(0.01)	(0.01)	
Attrited in year 2	0.10	0.10	0.10	0.10	0.95
	(0.01)	(0.01)	(0.01)	(0.01)	
Panel B: Households (N=7,001)					
HH size	6.23	6.26	6.41	6.26	0.19
	(0.12)	(0.12)	(0.13)	(0.08)	
Wealth index (PCA)	0.02	0.01	0.00	-0.02	0.99
	(0.16)	(0.16)	(0.17)	(0.12)	
Pre-treatment expenditure (TZS)	34,198.67	33,423.19	34,638.63	36,217.09	0.50
	(4,086.38)	(3,799.66)	(4,216.98)	(2,978.25)	
Panel C: Schools (N=350)					
Pupil-teacher ratio	54.78	58.78	55.51	60.20	0.50
	(2.63)	(3.09)	(2.53)	(3.75)	
Single shift	0.60	0.59	0.64	0.63	0.88
	(0.06)	(0.06)	(0.06)	(0.04)	
Infrastructure index (PCA)	-0.08	0.07	-0.12	0.06	0.50
	(0.13)	(0.14)	(0.16)	(0.08)	
Urban	0.16	0.13	0.17	0.15	0.85
	(0.04)	(0.04)	(0.05)	(0.03)	
Enrolled students	739.07	747.60	748.46	712.45	0.83
	(48.39)	(51.89)	(51.66)	(30.36)	
Panel D: Teachers (Grade 1-3) (N=1,569)					
Male	0.34	0.34	0.31	0.33	0.92
	(0.04)	(0.04)	(0.04)	(0.03)	
Age (in 2013)	39.36	39.53	39.05	39.49	0.52
	(0.85)	(0.85)	(0.74)	(0.52)	
Years of experience (in 2013)	15.34	15.82	15.11	15.71	0.32
	(0.88)	(0.92)	(0.75)	(0.54)	
Teaching Certificate	0.62	0.60	0.61	0.57	0.50
	(0.04)	(0.04)	(0.04)	(0.03)	

This table presents the mean and standard error of the mean (in parenthesis) for several characteristics of students in our sample (Panel A), households (Panel B), schools (Panel C) and teachers (Panel D) across treatment groups. The student sample consists of all students tested by the research team. The sample consists of 30 students sampled in year one (10 from grade 1, 10 from grade 2, and 10 from grade 3) and 10 students sampled in year 2 (from the new grade 1 cohort). The attrition in year 1 is measured using only the original 30 students sampled per school. The attrition in year 2 is measured using the sample of 30 students that are enrolled in grades 1, 2 and 3 in that year. 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 of 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 of a Principal Component Analysis of indicator variables for: outer wall, staff room, playground, library, and kitchen. Standard errors are clustered at the school level for test of equality. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: How are schools spending the grants?

	(1)	(2)	(3)
		TZS per student	
	Year 1	Year 2	Average
Admin.	1,773.07	2,069.72	1,912.14
	(148.29)	(199.23)	(126.52)
Students	622.45	456.27	533.80
	(94.69)	(82.08)	(64.16)
Textbooks	3,858.69	1,315.83	2,585.75
	(257.56)	(172.39)	(154.05)
Teaching aids	1,761.43	2,132.32	1,947.61
	(126.53)	(190.00)	(118.45)
Teachers	0.00	3.36	1.68
	(0.00)	(3.36)	(1.68)
Construction	60.35	69.76	65.49
	(36.58)	(61.16)	(35.33)
Total Expenditure	8,075.99	6,047.26	7,046.46
	(318.42)	(352.57)	(238.98)
Unspent funds	1,924.01	3,952.74	2,953.54
	(318.42)	(352.57)	(238.98)
Total Value of CG	10,000.00	10,000.00	10,000.00
	(0.00)	(0.00)	(0.00)

Mean grant expenditure per student of school grants. *Admin:* Administrative cost (including staff wages), rent and utilities, and general maintenance and repairs. *Student:* Food, scholarships and materials (notebooks, pens, etc.). *Textbooks:* Textbooks. *Teaching aids:* Classroom furnishings, maps, charts, blackboards, chalk, practice exams, etc. *Teachers:* Salaries, bonuses and teacher training. Standard errors in parentheses. 1 USD = 1,600 TZ Shillings. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Effect of grants on school, household, and total expenditure

	(1)	(2)	(3)	(4)	(5)
			TZS per student		
	Grant exp.	Other school exp.	Total school [(1)+(2)]	Household exp.	Total exp. [(3)+(4)]
Panel A: Year 1					
Grants (α_1)	8,070.68***	-2,407.92***	5,662.75***	-1,014.96	4,647.79***
	(314.09)	(813.88)	(848.58)	(1,579.79)	(1,724.64)
N. of obs.	210	210	210	210	210
Mean control	0.00	5,959.67	5,959.67	28,821.01	34,780.68
Panel B: Year 2					
Grants (α_1)	6,033.08***	-2,317.74**	3,715.34***	-2,164.18*	1,585.75
	(336.95)	(1,096.16)	(1,122.60)	(1,201.53)	(1,548.42)
N. of obs.	209	209	209	210	209
Mean control	0.00	4,524.03	4,524.03	27,362.34	31,886.37
Panel C: Year 1 + Year 2					
Grants (α_1)	7,055.98***	-2,367.94***	4,688.04***	-1,589.57	3,133.33**
	(230.07)	(688.89)	(724.91)	(1,053.64)	(1,241.09)
N. of obs.	419	419	419	420	419
Mean control	0.00	5,241.85	5,241.85	28,091.68	33,333.53

Results from estimating Equation 2 for grant expenditure per child, other school expenditure per child, total school expenditure per child, and household reported expenditure in education. Column (1) shows grant expenditure as the dependent variable. Column (2) shows other school expenditure. Column (3) shows total school expenditure. Column (4) shows household data on expenditure in education. Column (5) shows total expenditure (total school expenditure + household expenditure). Panel C regressions include data from both follow-ups, and therefore coefficients represent the average effect over both years. 1USD = 1,600 TZ Shillings. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Effect of grants on test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Grants (α_1)	-0.05	-0.01	-0.02	-0.03	0.01	-0.00	0.02	0.01
	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439

Results from estimating Equation 3 for different subjects at both follow-ups. Control variables include student characteristics (age, gender, grade and lag test scores) and school characteristics (PTR, Infrastructure PCA index, indicator for whether the school is in an urban or rural location, a PCA index of how close is the school to different facilities, and an indicator for whether the school is single shift or not). Clustered standard errors, by school, in parentheses. See Table A.7 for a version without school and household controls. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Effect of incentives on test scores: high- and low-stakes exams

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Panel A: Z-scores, low-stakes								
Incentives (α_2)	0.06	0.05	0.06	0.06*	0.07*	0.01	0.00	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)
N. of obs.	5,496	5,496	5,496	5,496	5,653	5,653	5,653	5,653
Panel B: Z-scores, high-stakes								
Incentives (β_2)	.	.	.	.	0.17***	0.12**	0.12**	0.21***
					(0.05)	(0.05)	(0.05)	(0.07)
N. of obs.	.	.	.	.	19,256	19,256	19,256	19,256

Results from estimating Equation 3 for different subjects at both follow-ups. Control variables include student characteristics (age, gender, grade and lag test scores), school characteristics (PTR, Infrastructure PCA index, indicator for whether the school is in an urban or rural location, a PCA index of how close is the school to different facilities, and an indicator for whether the school is single shift or not), and household characteristics (household size, a PCA wealth index, and education expenditure prior to the intervention). Panel B Year 1 results are not available due to data constraints (see text for details). Clustered standard errors, by school, in parentheses. See Table A.7 for a version without school and household controls. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Effect of grants, incentives, and their interaction on test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Panel A: Z-scores, low-stakes								
Grants (α_1)	-0.05	-0.01	-0.02	-0.03	0.01	-0.00	0.02	0.01
	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
Incentives (α_2)	0.06	0.05	0.06	0.06*	0.07*	0.01	0.00	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)
Combination (α_3)	0.10**	0.10***	0.10**	0.12***	0.20***	0.21***	0.18***	0.23***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	0.10	0.06	0.07	0.09	0.12	0.20	0.16	0.18
p-value ($\alpha_4 = 0$)	0.09	0.27	0.28	0.11	0.08	0.00	0.05	0.01
$\alpha_5 := \alpha_3 - \alpha_2$	0.05	0.05	0.05	0.06	0.13	0.20	0.18	0.19
p-value ($\alpha_5 = 0$)	0.31	0.22	0.38	0.21	0.01	0.00	0.00	0.00
Panel B: Z-scores, high-stakes								
Incentives (β_2)	.	.	.	.	0.17***	0.12**	0.12**	0.21***
					(0.05)	(0.05)	(0.05)	(0.07)
Combination (β_3)	.	.	.	.	0.25***	0.23***	0.22***	0.36***
					(0.05)	(0.06)	(0.06)	(0.08)
N. of obs.	.	.	.	.	46,886	46,882	46,882	46,882
$\beta_5 := \beta_3 - \beta_2$	.	.	.	.	0.08	0.11	0.10	0.15
p-value ($\beta_5 = 0$)	.	.	.	.	0.05	0.01	0.06	0.01
Panel C: Difference								
$\beta_2 - \alpha_2$	.	.	.	.	0.09	0.11	0.12	0.17
p-value($\beta_2 - \alpha_2 = 0$)	.	.	.	.	0.14	0.05	0.07	0.02
$\beta_3 - \alpha_3$	.	.	.	.	0.03	0.01	0.03	0.12
p-value($\beta_3 - \alpha_3 = 0$)	.	.	.	.	0.53	0.81	0.63	0.08
$\beta_5 - \alpha_5$					-0.05	-0.09	-0.09	-0.05
p-value($\beta_5 - \alpha_5 = 0$)					0.35	0.05	0.17	0.42

Results from estimating Equation 3 for different subjects at both follow-ups. Control variables include student characteristics (age, gender, grade and lag test scores), school characteristics (PTR, Infrastructure PCA index, indicator for whether the school is in an urban or rural location, a PCA index of how close is the school to different facilities, and an indicator for whether the school is single shift or not), and household characteristics (household size, a PCA wealth index, and education expenditure prior to the intervention). Clustered standard errors, by school, in parentheses. Panel B Year 1 results are not available due to data constraints (see text for details). Consequently, Panel C Year 1 is also not available. See Table A.7 for a version without school and household controls. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Spillovers to other subjects and grades

	(1)	(2)	(3)	(4)	(5)	(6)
	Science		Grade 7 PSLE 2013		Grade 7 PSLE 2014	
	Year 1	Year 2	Pass	Score	Pass	Score
Grants (α_1)	0.02	-0.04	-0.02	-0.03	-0.02	-0.05
	(0.05)	(0.06)	(0.03)	(0.05)	(0.03)	(0.05)
Incentives (α_2)	0.01	-0.01	-0.01	-0.01	-0.00	-0.02
	(0.05)	(0.05)	(0.03)	(0.04)	(0.03)	(0.05)
Combination (α_3)	0.09	0.09*	0.02	0.05	0.02	0.06
	(0.05)	(0.05)	(0.03)	(0.05)	(0.03)	(0.05)
N. of obs.	9,142	9,439	26,074	26,074	23,751	23,751
Mean control group			0.52	2.60	0.58	2.70
$\alpha_4 = \alpha_3 - \alpha_2 - \alpha_1$	0.058	0.13*	0.060	0.099	0.043	0.12*
p-value ($\alpha_4 = 0$)	0.48	0.096	0.15	0.14	0.31	0.080

Columns (1) and (2) estimate Equation 3 for science Z-scores in focal grades (Grd 1 - Grd 3) using data from low-stakes tests conducted by the research team. Columns (3)-(6) use data from the national exit examination as dependent variables: pass rates and average test scores. Clustered standard errors, by school, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Effect of grants, incentives, and their interaction on teacher behavior

	(1)	(2)	(3)	(4)
		Self-reported		
	Attendance	Tests	Tutoring	Remedial
Grants (α_1)	0.03	-0.27	0.01	-0.03
	(0.03)	(0.69)	(0.02)	(0.03)
Incentives (α_2)	-0.02	1.16*	0.03	-0.06*
	(0.03)	(0.66)	(0.03)	(0.03)
Combination (α_3)	-0.00	-0.18	0.05**	0.03
	(0.02)	(0.58)	(0.02)	(0.02)
N. of obs.	2,278	2,260	2,278	2,278
Mean of dep. var.	0.79	9.21	0.090	0.84
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-0.020	-1.07	0.00053	0.12
p-value ($\alpha_4 = 0$)	0.61	0.27	0.99	0.0048***

Results from estimating treatment effects on teacher behavior. Column (1) shows teacher attendance independently measured by enumerators during a surprise visit in the middle of the school year. Column (2) shows the number of tests per period as the dependent variable. Column (3) shows a dummy variable that indicates whether the teacher provided any extra tutoring to students as the dependent variable. Column (4) shows a dummy variable that indicates whether the teacher provided remedial teaching to students as the dependent variable. All regressions include data from both follow-ups, and therefore coefficients represent the average effect over both years. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Effect of grants, incentives, and their interaction on expenditure

	(1) Grant exp.	(2) Other school exp.	(3) Total school [(1)+(2)]	(4) Household exp.	(5) Total exp. [(3)+(4)]
Panel A: Year 1					
Grants (α_1)	8,070.68***	-2,407.92***	5,662.75***	-1,014.96	4,647.79***
	(314.09)	(813.88)	(848.58)	(1,579.79)	(1,724.64)
Incentives (α_2)	-6.77	-10.05	-16.82	-977.78	-994.60
	(63.15)	(642.21)	(638.81)	(1,294.84)	(1,439.10)
Combination (α_3)	8,329.38***	-1,412.22	6,917.16***	-1,382.23	5,534.93***
	(241.13)	(932.79)	(919.07)	(1,153.27)	(1,564.93)
N. of obs.	350	350	350	350	350
Mean control	0.00	5,959.67	5,959.67	28,821.01	34,780.68
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	265.47	1,005.76	1,271.23	610.51	1,881.74
p-value ($\alpha_4 = 0$)	0.50	0.44	0.33	0.77	0.45
$\alpha_3 - \alpha_1$	258.70	995.70	1,254.41	-367.27	887.14
p-value ($\alpha_3 - \alpha_1 = 0$)	0.51	0.39	0.28	0.83	0.67
Panel B: Year 2					
Grants (α_1)	6,033.08***	-2,317.74**	3,715.34***	-2,164.18*	1,585.75
	(336.95)	(1,096.16)	(1,122.60)	(1,201.53)	(1,548.42)
Incentives (α_2)	22.70	-1,166.46	-1,143.75	235.40	-907.97
	(98.63)	(818.24)	(830.33)	(1,214.01)	(1,422.09)
Combination (α_3)	5,620.07***	-1,896.28**	3,723.79***	-75.59	3,646.85**
	(320.69)	(928.05)	(989.27)	(1,151.27)	(1,520.20)
N. of obs.	349	349	349	350	349
Mean control	0.00	4,524.03	4,524.03	27,362.34	31,886.37
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-435.71	1,587.91	1,152.20	1,853.19	2,969.07
p-value ($\alpha_4 = 0$)	0.35	0.15	0.33	0.30	0.16
$\alpha_3 - \alpha_1$	-413.01	421.46	8.45	2,088.59	2,061.10
p-value ($\alpha_3 - \alpha_1 = 0$)	0.37	0.56	0.99	0.11	0.18
Panel C: Year 1 + Year 2					
Grants (α_1)	7,055.98***	-2,367.94***	4,688.04***	-1,589.57	3,133.33**
	(230.07)	(688.89)	(724.91)	(1,053.64)	(1,241.09)
Incentives (α_2)	8.02	-588.31	-580.30	-371.19	-951.10
	(59.68)	(535.92)	(542.97)	(984.59)	(1,092.17)
Combination (α_3)	6,974.56***	-1,654.05**	5,320.51***	-728.91	4,590.24***
	(224.51)	(692.00)	(721.74)	(919.30)	(1,240.62)
N. of obs.	699	699	699	700	699
Mean control	0.00	5,241.85	5,241.85	28,091.68	33,333.53
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-89.43	1,302.20	1,212.77	1,231.85	2,408.01
p-value ($\alpha_4 = 0$)	0.78	0.13	0.19	0.42	0.18
$\alpha_3 - \alpha_1$	-81.42	713.89	632.47	860.66	1,456.91
p-value ($\alpha_3 - \alpha_1 = 0$)	0.80	0.29	0.39	0.46	0.30

Results from Estimating Equation 2 for grant expenditure per child, other school expenditure per child, total school expenditure per child, and household reported expenditure on education. Column (1) shows grant expenditure as the dependent variable. Column (2) shows other school expenditure. Column (3) shows total school expenditure. Column (4) shows household data on expenditure in education. Column (5) shows total expenditure (total school expenditure + household expenditure). Panel C regressions included data from both follow-ups, and therefore coefficients represent the average effect over both years. 1 USD =1,600 TZ Shillings. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Effect of grants, incentives, and their interaction on textbook expenditure by grade

	(1) Grades 4-7	(2) Grades 1-3	(3) Difference [(2)-(1)]
Grants (α_1)	1,743.61***	1,259.14***	-484.47***
	(224.77)	(183.70)	(159.30)
Incentives (α_2)	-131.56	-50.42	81.13
	(105.69)	(71.51)	(92.99)
Combination (α_3)	1,504.34***	1,563.35***	59.01
	(194.64)	(202.35)	(228.66)
N. of obs.	2,780	2,100	4,880
Mean control	846.26	498.74	-347.52
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-107.71	354.64	462.35
p-value ($\alpha_4 = 0$)	0.72	0.19	0.10
$\alpha_3 - \alpha_1$	-239.27	304.21	543.48
p-value ($\alpha_3 - \alpha_1$=0)	0.40	0.25	0.045

Results from estimating Equation 2 on textbook expenditure per student for grades 4-7 (Column 1), grades 1-3 (Column 2), and the difference between them (Column 3). Expenditure per student in grades 4-7 are show in Column 1, expenditure per student enrolled in grades 1-3 are shown in Column 2, and the difference in Column 3. The regression includes data from both follow-ups, and therefore coefficients represent the average effect over both years. 1USD = 1,600 TZ Shillings. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Student			Teacher			School		
	Male	Age	Lagged score	Male	Salary	Yr. Birth	Facilities	PTR	Management
Grants*Covariate	0.02	0.00	-0.06**	-0.21*	0.00	-0.00	0.08	0.00	0.07
	(0.04)	(0.01)	(0.03)	(0.11)	(0.00)	(0.01)	(0.07)	(0.00)	(0.08)
Incentives*Covariate	-0.07*	-0.00	-0.01	0.01	-0.00	-0.00	0.14**	-0.00	-0.07
	(0.04)	(0.01)	(0.02)	(0.10)	(0.00)	(0.01)	(0.07)	(0.00)	(0.06)
Combination*Covariate	-0.10**	-0.03*	-0.06**	0.07	0.00	0.00	0.09	-0.00	0.15**
	(0.04)	(0.01)	(0.03)	(0.12)	(0.00)	(0.00)	(0.07)	(0.00)	(0.06)
N. of obs.	18,581	18,581	18,581	18,581	18,581	18,581	18,581	18,581	18,206

The dependent variable is the standardized composite (PCA) test score. Each regression has a different covariate interacted with the treatment dummies. The column title indicates the covariate interacted. The first three columns have the following covariates at the student level: the standardized test score at baseline; Gender, a dummy equal to one if the student is male; and the age in years. Columns 4-6 have the following covariates as the school level: a dummy for whether the PCA index of facilities is above the median; the pupil-teacher ratio; and a dummy equal to one if the PCA index for managerial ability of the principal is above the median. Columns 7-9 have the following covariates at the teacher level: a dummy if the teacher is male; the annual salary; and the year of birth of the teacher. The teacher covariates are averaged across teachers in both years. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A Additional tables and figures

Table A.1: Effect of grants, incentives, and their interaction on the pass rate in the high-stakes exam

	(1)	(2)	(3)	(5)	(6)	(7)
		Year 1			Year 2	
	Math	Kiswahili	English	Math	Kiswahili	English
Incentives (γ_2)	5.94***	6.87*	1.28	7.70***	7.28**	2.10**
	(1.95)	(3.61)	(1.00)	(1.84)	(3.35)	(0.81)
Combination (γ_3)	8.99***	11.70***	1.58	10.30***	13.64***	3.49***
	(2.05)	(3.59)	(0.99)	(1.97)	(3.27)	(1.06)
N. of obs.	327	327	327	327	327	327
Control mean	20.06	36.76	3.73	20.99	43.97	3.01
$\gamma_3 - \gamma_2$	3	4.8*	.3	2.6	6.4**	1.4
p-value ($\gamma_3 - \gamma_2 = 0$)	.1	.071	.69	.17	.018	.17

The dependent variable is the pass rate in the high-stakes exam. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.2: Effect of grants, incentives, and their interaction on household expenditure

	(1) Total expenditure	(2) Fees	(3) Textbooks	(4) Other books	(5) Supplies	(6) Uniforms	(7) Tutoring	(8) Transport	(9) Food	(10) Other
Panel A: Year 1										
Grants (α_1)	-1,014.96	-145.37	-33.05	-27.04	363.57	-334.43	-1,061.87	-143.55	542.56	-39.38
	(1,579.79)	(632.75)	(84.42)	(44.32)	(270.40)	(663.91)	(845.69)	(150.10)	(1,140.43)	(219.47)
Incentives (α_2)	-977.78	-11.27	7.73	-3.96	180.38	-287.47	-502.75	303.21	-240.27	-144.49
	(1,294.84)	(451.70)	(101.54)	(50.20)	(229.47)	(636.92)	(840.70)	(306.75)	(1,043.16)	(248.75)
Combination (α_3)	-1,382.23	-526.39	135.08	23.41	-52.45	-240.56	-708.35	86.01	-41.01	-210.18
	(1,153.27)	(391.13)	(82.78)	(56.94)	(253.33)	(640.66)	(874.28)	(270.39)	(779.80)	(217.14)
N. of obs.	350	350	350	350	350	350	350	350	350	350
Mean control	28,821.01	3,247.03	273.35	139.44	5,004.53	11,362.63	4,760.02	235.37	4,689.80	1,549.91
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	610.51	-369.75	160.41	54.40	-596.40	381.33	856.27	-73.66	-343.30	-26.31
p-value ($\alpha_4 = 0$)	0.77	0.64	0.26	0.47	0.13	0.71	0.51	0.85	0.82	0.94
$\alpha_3 - \alpha_1$	-367.27	-381.02	168.14	50.44	-416.02	93.86	353.52	229.56	-583.57	-170.80
p-value ($\alpha_3 - \alpha_1 = 0$)	0.83	0.58	0.084	0.36	0.20	0.91	0.72	0.38	0.62	0.45
Panel B: Year 2										
Grants (α_1)	-2,164.18*	-919.53*	-210.52**	46.71	-105.93	-427.54	-439.50	-70.46	-1,341.18**	-342.89*
	(1,201.53)	(550.69)	(100.77)	(65.39)	(246.27)	(638.46)	(693.04)	(301.90)	(624.04)	(204.00)
Incentives (α_2)	235.40	-147.95	-96.95	48.26	410.99	217.61	570.57	-445.89	-1,152.35**	-73.60
	(1,214.01)	(765.96)	(121.33)	(63.20)	(261.44)	(608.93)	(799.43)	(329.30)	(584.26)	(211.05)
Combination (α_3)	-75.59	-297.84	-145.61	85.07	175.34	320.83	-647.17	-420.25	-148.02	-101.52
	(1,151.27)	(605.34)	(92.38)	(61.37)	(253.04)	(589.29)	(749.68)	(316.05)	(872.65)	(184.35)
N. of obs.	350	350	350	350	350	350	350	350	350	350
Mean control	27,362.34	2,782.55	442.72	137.02	4,178.28	14,437.64	3,252.00	468.80	3,565.93	2,003.89
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	1,853.19	769.64	161.86	-9.90	-129.72	530.76	-778.24	96.10	2,345.52	314.98
p-value ($\alpha_4 = 0$)	0.30	0.38	0.29	0.92	0.73	0.57	0.49	0.78	0.031	0.28
$\alpha_3 - \alpha_1$	2,088.59	621.69	64.91	38.37	281.27	748.37	-207.67	-349.79	1,193.17	241.38
p-value ($\alpha_3 - \alpha_1 = 0$)	0.11	0.12	0.49	0.62	0.31	0.29	0.80	0.018	0.18	0.23

Results from estimating Equation 2 for household expenditure per child disaggregated by categories. 1USD = 1,600 TZ Shillings. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.3: Number of high-stakes test takers

	(1) Test Takers
Incentives (β_2)	0.01
	(0.02)
Combination (β_3)	0.05***
	(0.02)
N. of obs.	540
Mean control group	0.78
$\alpha_3 = \alpha_2 - \alpha_1$	0.033**
p-value($\alpha_3 = 0$)	0.019

The dependent variable is the proportion of test takers (number of test takers as a proportion of the number of students enrolled) during the high-stakes exam at the end of the second year. Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: Lee bounds for high-stakes exams: Z-scores

	(1) Math	(2) Kiswahili	(3) English	(4) Combined (PCA)
Incentives (β_2)	0.17***	0.12**	0.12**	0.21***
	(0.05)	(0.05)	(0.05)	(0.07)
Combo (β_3)	0.25***	0.23***	0.22***	0.36***
	(0.05)	(0.06)	(0.06)	(0.08)
N. of obs.	46,886	46,882	46,882	46,882
$\beta_4 = \beta_3 - \beta_2$	0.081**	0.11**	0.099*	0.15**
p-value ($H_0 : \beta_4 = 0$)	0.046	0.012	0.060	0.015
Lower 95% CI (β_2)	0.068	0.011	0.013	0.066
Higher 95% CI (β_2)	0.26	0.22	0.23	0.35
Lower 95% CI (β_3)	0.14	0.12	0.093	0.21
Higher 95% CI (β_3)	0.35	0.34	0.33	0.52
Lower 95% CI (β_4)	-0.00071	0.024	-0.014	0.027
Higher 95% CI (β_4)	0.16	0.20	0.20	0.28

The dependent variable is the standardized test score for different subjects. For each subject we present Lee (2009) bounds for all the treatment estimates (i.e., trimming the left/right tail of the distribution in Incentive and Combination schools so that the proportion of test takes is the same as the number in control schools). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: Heterogeneity by difference in dates between high- and low-stakes exams

	(1) Math	(2) Kiswahili	(3) English	(4) Combined (PCA)
Panel A: Both years				
Incentives	0.108	0.045	0.031	0.071
	(0.070)	(0.074)	(0.084)	(0.073)
Combo	0.288***	0.208***	0.221***	0.274***
	(0.074)	(0.072)	(0.083)	(0.074)
Incentives*Difference(Days)	-0.001	-0.001	-0.000	-0.001
	(0.002)	(0.002)	(0.003)	(0.002)
Combination*Difference(Days)	-0.005**	-0.002	-0.004	-0.004*
	(0.002)	(0.002)	(0.003)	(0.002)
N. of obs.	9,534	9,534	9,534	9,534
Panel B: Year 1				
Incentives	0.147	0.141	0.153	0.169*
	(0.099)	(0.091)	(0.094)	(0.090)
Combo	0.296***	0.159*	0.198**	0.252***
	(0.096)	(0.095)	(0.098)	(0.094)
Incentives*Difference(Days)	-0.002	-0.002	-0.003	-0.002
	(0.003)	(0.003)	(0.003)	(0.003)
Combination*Difference(Days)	-0.005*	-0.001	-0.003	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)
N. of obs.	4,674	4,674	4,674	4,674
Panel C: Year 2				
Incentives	0.096	0.032	-0.007	0.047
	(0.121)	(0.120)	(0.135)	(0.119)
Combo	0.275**	0.235*	0.273*	0.297**
	(0.123)	(0.119)	(0.144)	(0.124)
Incentives*Difference(Days)	-0.000	-0.002	-0.001	-0.001
	(0.005)	(0.005)	(0.006)	(0.005)
Combination*Difference(Days)	-0.003	-0.002	-0.007	-0.004
	(0.006)	(0.006)	(0.006)	(0.005)
N. of obs.	4,860	4,860	4,860	4,860

The dependent variable is the standardized test score. The absolute value of the time difference (in days) between the low-stakes and the high-stakes exams is interacted with the treatment dummies. Panel A pool the data for the low-stakes exam of both years. Panel B uses data from the low-stakes exam in the first year. Panel C uses data from the low-stakes exam in the second year. The average difference in testing dates in the first year is 29.9 days. In the second year the average difference is 17 days. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: Effect of grants, incentives, and their interaction on test scores on a fix cohort of students

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Grants (α_1)	-0.02	-0.04	-0.00	-0.02	0.06	0.01	0.03	0.04
	(0.05)	(0.05)	(0.05)	(0.04)	(0.06)	(0.06)	(0.06)	(0.05)
Incentives (α_2)	0.02	0.02	0.09*	0.05	0.09*	-0.02	0.01	0.03
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Combination (α_3)	0.12**	0.10**	0.13**	0.14***	0.25***	0.21***	0.18***	0.24***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.06)	(0.04)
N. of obs.	6,043	6,043	6,043	6,043	6,343	6,343	6,343	6,343
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	0.11	0.12*	0.046	0.11	0.096	0.21***	0.14	0.17**
p-value ($\alpha_4 = 0$)	0.12	0.090	0.55	0.12	0.21	0.0081	0.12	0.026

Results from estimating Equation 3 for different subjects at both follow-ups. Sample only includes students treated over the two-year period (i.e., students in grade 1 and grade 2 at baseline 2013). Control variables include only student characteristics (age, gender, grade and lag test scores). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.7: Effect of grants, incentives, and their interaction on test scores without controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Grants (α_1)	-0.05	-0.01	-0.03	-0.03	0.01	0.00	0.03	0.02
	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.05)	(0.06)	(0.05)
Incentives (α_2)	0.06	0.06	0.06	0.07*	0.08*	0.01	0.00	0.04
	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)	(0.04)
Combination (α_3)	0.10**	0.11***	0.10**	0.12***	0.21***	0.22***	0.19***	0.24***
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	0.096	0.059	0.065	0.085	0.12	0.20***	0.16*	0.18**
p-value ($\alpha_4 = 0$)	0.12	0.32	0.33	0.16	0.10	0.0068	0.054	0.011

Results from estimating Equation 3 for different subjects at both follow-ups. Control variables only include student characteristics (age, gender, grade and lag test scores). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: Heterogeneity by distance to the passing threshold

	(1)	(2)	(3)	(4)	(5)	(6)
	Year 1			Year 2		
	Math	Kiswahili	English	Math	Kiswahili	English
Panel A: Linear distance						
Grants × Distance	0.241*	-0.041	0.132	0.151	-0.036	-0.049
	(0.104)	(0.123)	(0.100)	(0.130)	(0.131)	(0.109)
Incentives × Distance	0.127	0.091	0.008	0.106	0.138	-0.095
	(0.108)	(0.120)	(0.105)	(0.116)	(0.137)	(0.088)
Combination × Distance	0.168	0.022	-0.101	0.175	0.186	-0.068
	(0.122)	(0.119)	(0.111)	(0.109)	(0.144)	(0.093)
N. of obs.	9,142	9,142	9,142	9,439	9,439	9,439
Panel B: Quadratic distance						
Grants × Distance^2	0.212	-0.050	0.101	0.201	-0.041	-0.049
	(0.113)	(0.160)	(0.085)	(0.151)	(0.162)	(0.095)
Incentives × Distance^2	0.074	0.082	0.007	0.074	0.179	-0.079
	(0.115)	(0.157)	(0.087)	(0.135)	(0.172)	(0.080)
Combination × Distance^2	0.203	0.010	-0.112	0.144	0.248	-0.056
	(0.142)	(0.158)	(0.097)	(0.131)	(0.189)	(0.082)
N. of obs.	9,142	9,142	9,142	9,439	9,439	9,439

The dependent variable is the standardized test score. The absolute value of the difference (in percentage points) between the baseline percentile and the overall pass rate (1-pass rate to be exact) in the control schools (in the high-stakes test) is interacted with the treatment dummies. For example, the pass rate in Grade 2 in the math test in Year 2 was 17%. Hence, a student in the 83 percentile would be right at the cutoff (and at a distance of zero). A student in the 20th percentile would be at a distance of 63 percentage points. A student in the 90th percentile would be at a distance of 7 percentage points. The value of the variable distance ranges from 0 to 1. Panel A interacts the treatment dummies with the absolute value of the distance. Panel B interacts the treatment dummies with the square value of the distance. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.9: Inputs for cost-effectiveness calculations

	Direct	Implementation	Low-stakes effect	High-stakes effect
Grants	5.89	1.24	0	0
Incentives	2.52	4.58	0	0.21
Combination	8.71	4.58	0.23	0.36