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

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

January 14, 2019

Abstract

We present results from a large-scale randomized experiment across 350 schools in Tanzania that studied the impact of providing schools with (a) unconditional grants, (b) teacher incentives based on student performance, and (c) both of the above. After two years, we find (a) no impact on student test scores from providing school grants, (b) some evidence of positive effects from teacher incentives, and (c) significant positive effects 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 combining spending on school inputs (which is the default policy) with improved teacher incentives could substantially increase the cost-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 Dinkelman, 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 acknowledge the support of Bryan Plummer and the J-PAL Africa staff during the launch of the project. Erin Litzow, Jessica Mahoney, Kristi Post, and Rachel Steinacher provided excellent on-the-ground research support through Innovations for Poverty Action. We also thank Austin Dempewolff, and 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 outstanding 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

[§] ITAM: mtromero@itam.mx

[¶] 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 have found it difficult to convert increases in spending and enrollment into improvements in student learning (World Bank, 2018). One reason could be that education systems face several additional constraints beyond limited school resources.[1] Thus, simply augmenting school resources may have limited impact on learning outcomes if other binding constraints are not alleviated at the same time.

A specific constraint that may limit the effectiveness of school inputs is low teacher effort — exemplified by high rates of teacher absence documented in several developing country settings (Chaudhury, Hammer, Kremer, Muralidharan, & Rogers, 2006). Thus, while school inputs may improve learning when teacher effort is high (either due to intrinsic motivation or external incentives/monitoring), they may be less effective when teacher effort is low. Conversely, returns to teacher effort may be low in the absence of adequate school inputs. In such a setting, the impact of jointly improving school resources and teacher effort may be greater than the sum of doing both on their own.

This paper tests for such complementarities 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 of TZS 10,000 (∼US$6.25 at the time of the study) per-student, effectively doubling discretionary school resources.[2] The second program 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] Some of these include poor student health and nutrition, low student attendance, mismatch between curriculum/pedagogy and student learning levels, and low levels of teacher effort and accountability. 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] The Government's capitation grant policy aimed to provide schools with TZS 10,000/student. The program we study provided another TZS 10,000/student over and above this grant, effectively doubling the school grant. Since teacher salaries were paid directly by the government and did not pass through the schools, these grants (from the government and the program) were the main source of discretionary funding available to schools. Schools were not permitted to use grant funds to augment teacher salaries or hire new teachers (consistent with government rules for capitation grant expenditure).

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 program significantly increased per-student discretionary expenditure in treated schools. We find evidence of reduction in school and household spending in the Grant schools. Even after this reduction, there was a significant increase in *net* discretionary 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, student test-scores in Incentive schools were modestly higher than those in the control group, but these differences were not statistically significant for most subjects. However, on the high-stakes tests administered by Twaweza (that were used to calculate teacher bonus payments), we find significant positive treatment effects. 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. 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 and interpretation of the differences in results on the two sets of tests to Section 4.2.

Third, students in Combination schools, that received both school grants and teacher incentives, had significantly higher test scores 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 pass rates on the latter were 49%, 31%, and 116% higher in math, Kiswahili, and English.

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

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

were significantly greater than the sum of the gains in the 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 meaningful: point estimates of the impact of the Combination treatment are over *five times greater* than the sum of the impact of the Grant and Incentives treatments after two years.[4] Thus, school inputs may be effective when teachers have incentives to use them effectively, but not otherwise. Conversely, motivated teachers may be more effective with additional school inputs.

While we find strong evidence of complementarities between the grant and incentive programs as implemented, cost-effectiveness calculations also depend on the cost of implementing the programs, and the dose-response relationship between different values of grants and incentives and impacts on test-scores. Assuming a linear dose-response relationship, we estimate that the combination of grants and incentives would clearly be more cost effective at improving test scores compared to spending the total cost of the Combination treatment on larger school grants instead. However, we cannot rule out the possibility that it may have been just as cost-effective to spend all the money spent on the Combination program on a larger teacher incentive program instead.[5]

To help interpret our results, we develop a simple stylized model where teachers optimize effort choices given an education production function (increasing in teacher effort and school inputs), their non-monetary and monetary rewards from improving student learning, and a minimum learning constraint. The model highlights that it is only under the implicit (and usually unstated) assumption that teachers have non-financial reasons for exerting effort that we should expect extra inputs to improve test scores. Instead, if teachers act like agents in standard economic models, then the optimal response to an increase in inputs may be to reduce effort, which may attenuate impacts of additional inputs on learning. However, the introduction of financial incentives will typically raise the optimal amount of teacher effort when inputs increase, yielding policy complementarities between inputs and incentives in improving learning outcomes.

Our first and most important contribution is to experimentally establish the existence

[4] Since the number of students passing exams was greater in Combination schools than in Incentive schools, total program spending in Combination schools was 3.5% greater than the sum of spending in Input and Incentive schools. The results on complementarities are robust to accounting for this additional expenditure (see calculations in Section 4.4).

[5] This is because implementation costs were a much larger fraction of total costs in the Incentive program (though the total amount spent on the Grant and Incentive programs were very similar at USD 7.13 and 7.10 per student/year respectively). Thus, spending all the money from the Combination program on incentives would enable the value of the incentives to be 3.45 times higher than provided under the Incentives program (because the implementation cost does not change with the size of the bonus), whereas it would only yield 2.05 times the value of grants in the Grants program (see details in Section 5.3).

of complementarities across policies to improve learning outcomes. While several field experiments have employed factorial (or cross-cutting) designs that could in principle be used to test for such complementarities, these studies have usually been under-powered to detect economically meaningful complementarities (typically due to budget and sample size constraints). 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)). Overall, as noted in a recent meta-analysis of education studies, "[t]here are surprisingly few experiments (in education) with fully factorial designs that allow for strong experimental tests of [complementarities]" (McEwan, 2015).[6] [7]

Second, 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); de Ree, Muralidharan, Pradhan, and Rogers (2018)). 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.[8]

[6] A notable exception from the early childhood development literature 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. Two other noteworthy studies are Behrman, Parker, Todd, and Wolpin (2015) who study providing incentives to teachers, students, and both of them (in Mexico) and List, Livingston, and Neckermann (2012) who study providing incentives individually to teachers, students, and parents or to a combination of these (in Chicago). However, neither study tests for complementarities because the variants of the incentives implemented are different across the individual and joint treatments.

[7] There is a parallel literature on dynamic complementarities in human capital investments over time (Cunha & Heckman, 2007; Malamud, Pop-Eleches, & Urquiola, 2016; Johnson & Jackson, 2017). While conceptually similar, this literature is substantively different because it focuses on complementarities across sequential investments. Our paper is situated more in the development economics tradition, where the idea that there may be complementarities across policies implemented *contemporaneously* (due to multiple constraints binding simultaneously) has been a central theme (Ray, 1998; Banerjee & Duflo, 2005).

[8] 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 to translating additional school resources into improved learning may not bind if teachers are suitably motivated to use school resources better.

Finally, we contribute to the broader literature on teacher incentives. While global evidence on the effectiveness of teacher incentives is mixed, the patterns in the results suggest that such policies are more effective in developing countries, perhaps due to greater slack in teacher effort (Ganimian & Murnane, 2016). 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.[9]

An important policy challenge for global development is that disadvantaged places also tend to be those with weaker governance. For instance, teacher absence rates are consistently higher in countries and states with lower per-capita income (Chaudhury et al., 2006). Thus, places that are most in need of additional resources to provide basic services like education are also places that are likely to be the least efficient at converting additional spending into improved outcomes. Our results suggest that combining funds for education inputs (which is what is done under the status quo) with incentives for improved outcomes may be a promising option for addressing this challenge.

This idea 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 reforms that link parts of school financing to state-level learning outcomes (Collier, 2016; Mesecar & Soifer, 2016). Our results provide empirical support for such policy approaches.

2 Context and Interventions

2.1 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 in 2001 of school fees in public primary schools, 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 grade-2 literacy Kiswahili (the national language and medium of instruction in primary schools), or in basic numeracy. Proficiency in English

[9] 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.

(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,[10] 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 met the World Bank definition of having 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 students (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, students 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.

2.2 Interventions and Implementation

The interventions studied in this paper were implemented by Twaweza, an East African civil-society organization focusing on citizen agency and public service delivery. Through its Uwezo program, Twaweza has conducted large-scale, citizen-led independent 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 nationally-representative sample of schools across Tanzania over two years (2013 and 2014). Twaweza (with technical inputs from the research team) implemented the interventions as part of

[10] 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).

a randomized controlled trial to facilitate learning about program impacts. Twaweza also worked closely with government officials to ensure smooth implementation of the program and evaluation. The interventions are described below.

2.2.1 Capitation Grant (Grants) Program

Schools randomly selected for the capitation grants program received TZS 10,000 (∼US$6.25 at the time of the study) per student from Twaweza. This was over and above funds received under the Government's capitation grant program, which also had a stipulated value of TZS 10,000/student. Guidelines for expenditure from program funds were similar to that of the government's own capitation grant program.[11] In practice, there were three key differences in the implementation quality of the government and Twaweza grant programs. First, the per capita Twaweza grant was larger than the per-capita Government grant actually received by schools.[12] 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 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.

The grant value was sizeable. For context, GDP/capita in Tanzania in 2013 was ∼US$1,000 and the per-student grant value was ∼0.6% of GDP/capita. If schools spent all of their grants on books, the funds would be sufficient to purchase about ∼ 4-5 textbooks/student. Overall, Twaweza disbursed ∼US$350,000/year to the 70 schools in the

[11] For instance, capitation grant rules do not allow these funds to be used for teacher salaries. The Twaweza Grants program had the same guidelines.

[12] On average, schools received only around 60% of the stipulated grant value of TZS 10,000/student from the government's capitation grant program, and many received much less than that (World Bank, 2012). Reasons included inadequate budgetary allocations by the central government, diversion of funds for other uses by local governments, and delays in disbursements.

Grant program and delivered what a *well-implemented* school capitation grant program would look like. Studying the impact of the Twaweza program on learning outcomes therefore provides a likely upper bound of the impact of a scaled-up government school-grant program, since the Twaweza program was implemented better.

2.2.2 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 ($\sim$ US$3) bonus for each student who passed a simple, externally-administered, grade-appropriate test based on the national curriculum. Additionally, the head teacher was paid TZS 1,000 ($\sim$ US$0.6) for each subject test a student passed.[13]

The term used by Twaweza for the teacher-incentive program was "Cash on Delivery" to reinforce the contrast between the approaches that underlay the two programs — with the Grants program being one of unconditional school grants, and the teacher incentive program being one where payments were contingent on outcomes.[14] The communication to schools and teachers emphasized that the aim of the Incentives 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 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 measures of student and teacher value addition.

There are also important limitations to such a threshold-based design (Ho, Lewis, & MacGregor Farris, 2009). 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

[13] 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.

[14] Twaweza used the term "Cash on Delivery" as a local version of a concept developed in the context of foreign aid by Birdsall et al. (2012).

considered decision to keep the formula simple in the interest of transparency, simplicity of explaining to teachers, and ease of implementation.[15] 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.

2.2.3 Combination Arm

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

[15] 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 Research Design

3.1 Sampling and Randomization

We conducted the experiment in a *nationally-representative* sample of 350 public schools across 10 districts in mainland Tanzania.[16] 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 1). 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. Thus, adding over 10 districts, the study had a total of 70 schools in each of the 3 treatment arms (Grants, Incentives, and Combination) and 140 schools in the control group (Figure 1).

3.2 Data

Our analysis uses 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/expenditure.[17] While most categories of school expenditure are difficult to map onto 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, and 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 (and teachers, students, and parents were informed of this). From this set of 10,500 students, we randomly sampled 10 students from each school (five from each of grade 2 and 3) to conduct household surveys. These

[16] 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).

[17] Data on school expenditures were collected by reviewing receipts, accounting books, and other accounting records, following the methods of the expenditure tracking surveys developed and used by the World Bank (Reinikka & Smith, 2004; Gurkan, Kaiser, & Voorbraak, 2009). These data do not include teacher salaries since salaries are paid directly by the government and do not pass through the school.

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).[18]

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 Twaweza only recorded aggregated pass rates (needed to calculate bonus payments) in the first year. The low- and high-stakes tests covered very similar content; see Appendix C for details on the design and implementation of both the low- and the high-stakes tests.

Figure 2 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.

3.3 Summary Statistics and Validity

The randomization was successful and observable characteristics of students, households, schools, and teachers are balanced across 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 average student is 9 years old and $\sim$50% are male (Panel A). The schools are mostly rural (85%), mean enrollment is $\sim$730, and class sizes are large — with an average of over 55 students per teacher (Panel C).[19] Teachers in our sample were $\sim$2/3 female, $\sim$40 years old, had $\sim$15 years of experience, and $\sim$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

[18] Because most of the survey questions focused on educational expenditures, including those 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, a representative sample of second graders (the initial cohort of first graders) was added to the household survey.

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

years, with slightly lower attrition in the second year (last two rows of Table 1: Panel A). On the high-stakes tests, there is no differential student attendance in Incentive schools relative to the control group, but attendance in Combination schools was higher (Table A.3). We therefore present bounds of treatment effects on high-stakes tests, using the approach of Lee (2009).

3.4 Empirical Strategy

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

$$Y_{sdt} = \alpha_0 + \alpha_1 Grants_s + \alpha_2 Incentives_s + \alpha_3 Combination_s + \gamma_d + \gamma_t + X_s\alpha_4 + \varepsilon_{sdt}, \tag{1}$$

where Y_{sdt} is the outcome of interest in school s in district d at time t. $Grants_s$ is an indicator variable for a school s receiving only the capitation grant program, $Incentives_s$ indicates a school s that received only the teacher incentive program, and $Combination_s$ 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 such as self-reported effort. All standard errors are clustered at the school-level.

We use the following estimating equation to study effects on learning outcomes:

$$Z_{isdt} = \delta_0 + \delta_1 Grant_s + \delta_2 Incentives_s + \delta_3 Combination_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_g + X_i\delta_4 + X_s\delta_5 + \varepsilon_{isdt}, \tag{2}$$

where Z_{isdt} 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 academic subjects, we also create a composite summary measure of test scores, by using 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 estimate impacts on this sample (without $Grants_s$). Further, because the high-stakes exam was 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.

Following our pre-analysis plan, we prioritize results using low-stakes tests, but present results on high-stakes tests to enable comparison with the literature on teacher incentives. We jointly estimate the impacts of all interventions in a pooled regression and present estimates for all interventions together in the tables below. However, for clarity of exposition, we first discuss the impacts of each treatment (Grants, Incentives, and Combination) individually, and then test for complementarities (specifically, we test $H_0 : \delta_3 - \delta_2 - \delta_1 = 0$) and discuss those results.

4 Results

4.1 Capitation Grant Program

4.1.1 How Were Grants Spent?

Table 2 (columns 1-3) 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, in compliance with program rules.

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 were uncertain (both in terms of timing and amount), we interpret this as “precautionary saving” and/or “consumption smoothing” behavior by schools (as also seen in 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).

4.1.2 Did Grants Change other Spending?

Table 3 examines the extent to which receiving the Grant program led to changes in school and household spending. Column 1 presents total extra spending from the Twaweza grant program. Schools that received Twaweza capitation grants saw a reduction in school expenditures from other sources (Column 2). Aggregating across both years, schools receiving the Grants program saw a reduction in school spending from other sources of ∼2,400 TZS per student, which is around a third of the additional spending enabled by the Grant program (Panel C - Columns 1 and 2).[20]

Since average school spending per student (excluding teacher salaries) in the control group was ∼5,200 TZS, spending the full grant value of 10,000 TZS would have tripled this amount. After accounting for savings and reductions in school spending, there was still a significant net increase in discretionary school spending per student of ∼4,700 TZS — almost double the expenditure relative to the control group (Panel C - Column 3).[21]

Next, we examine changes in household spending (Column 4) and report total net per-student spending, accounting for both school and household spending (Column 5). Consistent with the results documented by Das et al. (2013), we see an insignificant reduction in household spending by ∼1,000 TZS per student in the first year, and a larger significant reduction of ∼2,200 TZS per student in the second year ($p = 0.07$). The main categories of spending where we see cuts are fees, textbooks, and food (Table A.2).[22] 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 (discretionary) spending of ∼3,100 TZS/year (Panel C, Column 5), a 60% increase relative to mean school-spending per student in the control group (excluding teacher salaries).

4.1.3 Did Grants Improve Learning?

Despite the significant increase in per-pupil funding seen 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; Panel A in Table 4).[23] Offsets are unlikely to be the main reason for our results, as we

[20] Our analysis of school finances suggests that these expenditure reductions are due to both reduction in receipts of regular capitation grants by schools receiving Twaweza grants, as well as increased saving of funds from the regular capitation grant by schools. Since we care most about actual increases in spending and their impact on learning, we focus on expenditure as opposed to income or savings.

[21] We focus our analysis on discretionary spending at the school level (that is mainly funded by the government's capitation grants to schools) and exclude items like teacher salaries that are outside the control of the schools. If teacher salaries are included, the net increase in school spending was 16%.

[22] Das et al. (2013) posit that the time pattern in the reduction of household spending 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. It is also possible that some of the reductions (like fees and textbooks) are driven by schools expecting parents to contribute less in the second year after receiving the Twaweza grant.

[23] Grade retention was not an outcome of interest in our pre-analysis plan because Tanzanian education policy stipulates that grade promotion is mostly automatic in the early years of school. We test and confirm that there was no difference in grade retention across the treatment groups (Table A.11)

do not see any impacts of the grant on test scores even in the first year, when the net increase in discretionary 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).

4.2 Teacher incentives

On the low-stakes tests administered by the research team, test scores in Incentive schools are modestly higher than those in the control group, but typically not significant (Table 4: 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 ($p < 0.05$), and 0.21σ higher ($p < 0.01$) on the composite measure (Table 4: Panel B).[24]

We now consider possible reasons for the difference in estimated impacts across the two sets of tests. First, the content of the tests was very similar and so the differences are unlikely to be explained by test content. Second, Twaweza employed strict security protocols for the high-stakes test (as mentioned in Section 2.2.2), 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. Third, low-stakes tests were conducted $\sim$3 weeks before high-stakes test in both years. Since schools often conduct reviews and practice exams at the end of the school year, the superior performance on high-stakes tests could reflect this additional preparation (which was likely more intense in the Incentive schools).

[24] 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.

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, the high-stakes tests implemented by Twaweza were 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 for grades 1, 2, and 3. Further, qualitative interviews suggest that teachers were more likely to have emphasized the importance of these tests 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 across tests is the greater salience of the high-stakes tests in the Incentive schools and a resulting increase in student effort on these tests. The estimated difference in the treatment effects across the two sets of tests — 0.10-0.15σ, is exactly in line with recent experimental estimates that quantify the role of testing-day student effort on measured test scores (Levitt, List, Neckermann, & Sadoff, 2016; Gneezy et al., 2017).

The confirmation that test-taking effort is a salient component of measured test scores by Levitt et al. (2016) and Gneezy et al. (2017) 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.

Following our pre-analysis plan, we focus on the low-stakes tests in this paper because these were conducted by the research team (as opposed to the implementation team) and were conducted in *all* treatment groups, which is essential to test for complementarities (high-stakes tests were not carried out in Grant schools). 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 (including authors of meta-analyses of teacher incentives) may prefer to focus on the estimates from the high-stakes tests for cost-effectiveness calculations and comparing with existing studies. We present both sets of results for completeness.

4.3 Combination of Capitation Grant and Teacher Incentives

4.3.1 Grant Expenditure and Offsets

Combination schools spent their extra grant funds in a similar manner as those receiving only the grants (Table 2: Columns 4-6) and we find no significant difference in expenditure patterns of these funds between the Grant schools and the Combination schools (Column 7). Similar to the Grant schools, we find a reduction in school and household expenditures in the Combination schools as well (Table 3: Columns 2 and 4). As in the case of the Grant schools, these responses attenuated the impact of the Twaweza grant on per-student spending, but did not fully offset it.[25] On net, Combination schools saw a significant increase in per-student discretionary spending of $\sim$4,600 TZS/year (Panel C, Column 5), a 90% increase relative to mean per-student spending in control schools.

4.3.2 Impact on Test Scores

After one year, relative to the control group, students in Combination schools scored 0.10σ higher on the low-stakes tests in all three focal subjects, and scored 0.12σ higher on the composite measure ($p < 0.05$ in all cases, Table 4: Panel A). After two years, they scored 0.20σ, 0.21σ, 0.18σ higher on math, Kiswahili, and English, and scored 0.23σ higher on the composite measure of learning ($p < 0.01$ in all cases).[26]

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, and scored 0.36σ higher on the composite measure ($p < 0.01$ in all cases, Table 4: Panel B).[27] Pass rates (which 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 10.3, 13.6, and 3.5 percentage-point increases relative to the control means of 21%, 44%, and 3%. Pass rates were also higher for 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

[25] The magnitudes of the reduction in school and household spending are lower in the Combination schools than in the Grant schools. However, the differences are not significant (Panel C - last row).

[26] 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. Appendix Table A.6 shows the results using only the students who were exposed to the interventions in both years. We find very similar results among this group.

[27] 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 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.

4.4 Complementarities Across Programs

Using the low-stakes conducted in all schools, we find strong evidence of complementarities between the grant and incentive programs. Specifically, after two years, the impact of the Combination program on test scores was significantly greater than the *sum* of the impacts of the Grant and Incentive programs on their own. This difference is significant for *every* academic subject and also for the composite measure of learning (α_4 in Table 4: Panel A). The point estimate for complementarities is also positive in 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.

Since the number of students who passed the exams (on which the bonuses were paid) was higher in the Combination schools than in the Incentive schools (Table A.1), the total amount spent per student in the Combination schools was slightly (3.5%) higher than the sum of the per-student spending in the Grant and Incentive schools.[28] We therefore test for $\alpha_3 = 1.035 * (\alpha_1 + \alpha_2)$ and strongly reject equality ($p < 0.01$). Since grant spending is the same in both Combination and Grant schools, but incentive payments were 12% higher in Combination schools than in Incentive schools, we also test for $\alpha_3 = \alpha_1 + 1.12 * \alpha_2$ and reject equality ($p < 0.02$). In short, school inputs appear to be effective when teachers have incentives to use them effectively, but not otherwise. Conversely, motivated teachers (either for non-financial reasons or through incentives) can be more effective with additional educational inputs.

While we cannot test for complementarities on the high-stakes tests (because these were not conducted in Grant schools), we see suggestive evidence of 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: Panel A), then we can interpret the significant difference on the high-stakes tests between Combination and Incentive

[28] Specifically, per student program spending in Grant, Incentive, and Combination schools was USD 5.89, 2.52, and 8.71 respectively. Thus spending in Combination schools was 3.5% higher ((8.71/(5.89+2.52)) than the sum of spending in the Grant and Incentive schools. The additional spending is small because a large fraction of the bonus payments are made to teachers based on students who would have passed anyway (as seen by the pass rate in the control group), and so the additional incentive payment in Combination schools is only 12% higher ((8.71-5.89)/2.52).

schools as evidence of complementarities (β_5 in Table 4: 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 4: Panel C), except for Kiswahili (p-value 0.05). In other words, the estimated effects of the "Grant plus complementarities" are similar across the low- and high-stakes tests. These results are consistent with the idea that the high stakes boost the "levels" of test scores in both Incentive and Combination schools, but that the magnitude of the complementarities with school inputs was similar on both sets of tests.

The experimental evidence of complementarities across school inputs and teacher incentives 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 complementarities between teacher incentives and inputs in prior work. For instance, Muralidharan and Sundararaman (2011b) and Muralidharan (2012) find greater impacts 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). Finally, Andrabi, Das, Khwaja, Ozyurt, and Singh (2018) find positive effects on learning outcomes from providing unconditional grants to private schools (in contrast with the literature finding no effects of such grants on learning outcomes in public schools), which may be explained by private schools having better incentives to use their resources effectively.

Yet this evidence is only suggestive because teacher education and training, or textbooks, or the incentives of private school managers 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 adequately powered to either detect or rule out economically meaningful complementarities (defined as a magnitude comparable to those of the main effects). This allows us to experimentally demonstrate the presence and importance of complementarities between input and incentive-based policies for improving learning outcomes.

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

4.5 Other Results

4.5.1 Multi-tasking

An important concern with teacher performance-pay schemes is the risk that such programs could 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 these programs are able to improve students' 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 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 5). Further, mirroring the patterns we see on the incentivized subjects, we find evidence of complementarities between grants and incentives in science 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).

4.5.2 Intra-school Resource Allocation

For the Grant and Combination schools, the value of the school grant was based on the total enrollment across all grades (with the same per-student value of 10,000 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 by examining spending on textbooks (an expenditure category that can be mapped to grades) across students in Grades 1 to 3 (focal grades for the study) and Grades 4-7 (non-study grades). Grant schools spent nearly 40% more on textbooks in higher grades; however, we see no such pattern in the Combination schools, where per-student textbook spending is similar across grades (Table 6). This difference may be explained by the presence of teacher incentives for learning outcomes in lower grades in the Combination schools but not in Grant schools.

Finally, we examine 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 any of the treatment arms on this metric, both in terms of average scores or pass rates (Table 5: Columns 3-6). Thus, despite textbook spending in grades 4-7 increasing to nearly triple the value in the control group, we find no impact on 7th grade test scores in either the Grant schools or the Combination schools. These results again suggest that teacher incentives were key to making effective use of the additional resources (since Combination schools only had incentives for Grades 1-3 and not for Grade 7).[30]

4.5.3 Heterogeneity

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 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.9.[31]

Next, we test for heterogeneity by student, teacher, and school characteristics using Equation 1, 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 7, 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

[30] There is some evidence of complementarities on Grade 7 test scores in the second year (p-value 0.08). However, since the Combination program had no impact per se and there is no evidence of complementarities in the first year, we see this as suggestive evidence.

[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.

Incentive schools, but are broadly consistent (columns 1-3).[32] 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 latter, 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 school inputs and incentives.

We also find suggestive evidence of greater effects of receiving school grants (significantly so in 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 main result of complementarities between school inputs and conditions where these inputs are used well. 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 Discussion

5.1 Theoretical Framework

Our results confirm the lack of impact of inputs on their own, but also show that inputs can improve learning when teachers are motivated to do so. To help interpret our results, we present a simple stylized theoretical framework in Appendix B. The model specifies a production function for test scores (that is increasing in school inputs and teacher effort), teacher utility, and a minimum learning constraint below which teachers get sanctioned. It clarifies that the impact of an education intervention on learning outcomes will depend on both the production function and behavioral responses by teachers.

The model highlights that it is only under the implicit (and usually unstated) assumption that teachers have non-monetary motivation to improve learning that increasing in-

[32] We also 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). 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 (Figure A.1).

puts should be expected to improve test scores. In contrast, if teachers behave like agents in standard economic models (with disutility of effort and limited non-monetary utility from teaching), 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. The intuition is straightforward: when inputs increase, teachers can achieve the minimum learning-level constraint with lower effort. However, providing incentives to teachers will typically raise the optimal effort when inputs are increased, giving rise to policy complementarities between providing inputs and incentives.[33]

While this model is not the only possible explanation for our results, it provides an intuitive and parsimonious framework to interpret our experiment and results, as well as existing results in the literature. In addition to the several experimental studies in developing countries cited earlier that find no impact on test scores from providing additional inputs, there is also considerable evidence that teachers in developing countries reduce effort when provided with more resources.[34] The model can explain all of these existing results and helps to clarify the importance of teacher motivation (either financial or non-financial) in translating school inputs into learning outcomes.

5.2 Mechanisms

As suggested by the model above, a likely mechanism for the results we find is increased teacher effort (due to the incentives) and increased effectiveness of this additional effort when the teacher has more educational materials to work with. However, we do not find impacts on survey-based measures of teacher attendance, and teacher self-reports (Table A.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 practice tests given, or provision of remedial teaching.

In practice, it is likely that the test-score results are driven by increased intensity of

[33] This discussion also helps to clarify two important points regarding the study of complementarities in human capital formation. First, though much of the theoretical literature focuses on *production function* complementarities, 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. Second, 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, which increase the likelihood of complementarities (consistent with our findings).

[34] 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.

teaching effort within the classroom. However, this is difficult to measure well through surveys and observations, and we do not have any direct evidence of this since we prioritized collecting data on expenditure and outcomes and 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.

We do see two pieces of suggestive evidence of increased teacher effort in Combination schools. First, the increase in net expenditure (Table 3: 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 effort in Combination schools, to encourage parents to not reduce their own education spending in response to the school grants. For example, Combination schools seem to have not offered any fee reductions in the second year, while Grant schools did (Table A.2). Second, 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 (Table 6).

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.

5.3 Cost Effectiveness

Moving from treatment effects to cost-effectiveness calculations requires a discussion of three additional issues. These include the cost of implementing the programs, discussions on scaling of the magnitude of impacts at larger value of grants and incentives, and whether we should rely on estimates from low-stakes or high-stakes tests.

The main cost of implementing the capitation grant program was for conducting the audits. The costs of implementing the teacher incentive program included those of independently testing all the students, calculating bonuses, paying them out, and communicating these details to teachers. The cost of implementing the Combination program was the same as implementing the Incentives program (because the audits were conducted during the same visit as that in which students were tested). Table A.10 provides the direct and implementation costs of all three programs (per student). These are as follows:

Grants — 5.89 and 1.24 USD (total of 7.13 USD); Incentives — 2.52 and 4.58 USD (total of 7.10 USD); Combination — 8.71 and 4.58 USD (total of 13.29 USD).

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 5.78 USD.

We next consider the issue of scaling. Specifically, what would the impacts be if we spent all the money from the Combination program on inputs or incentives? Doing this requires us to make an assumption of a linear dose-response relationship between per-student program spending and impact (which we justify below).[35] Spending the full value of the Combination program on inputs would yield a per-student input expenditure of 12.05 USD (13.29 minus implementation cost of 1.24), which would be 2.05 times greater than the value provided in the Grants treatment (5.89 USD). We therefore test $\alpha_3 = 2.05 * \alpha_1$ in Table 4 (0.23 vs 0.02), and reject equality ($p = 0.03$). Thus, it is highly unlikely that spending all the money on grants would have raised test scores by the amount seen in the Combination schools.[36]

If we spent the full amount of the Combination program on the Incentive program, the value of the Incentives would be 8.71 USD (13.29 minus the implementation cost of 4.58), which is 3.45 times greater than the bonuses provided in the Incentives treatment. Conducting a similar test, the point estimate of α_3 is greater than $3.45 * \alpha_2$ in Table 4

[35] Our experiment was designed to test for complementarities among two program as implemented: i.e. is the impact of providing two programs together greater than the sum of the impacts of providing them individually. But, this design requires a linearity assumption for cost-effectiveness calculations against a counterfactual of spending all the money in the Combination arm on one treatment. An alternative experimental design would have been to allocate the same amount of money to Grant, Incentive, and Combination schools (including implementation costs) to test for how the marginal dollar should be optimally allocated (though even here, the researcher needs to decide how to split the fixed amount of money between the two individual treatments in the Combination schools). However, while this design can do cost effectiveness calculations without any further assumptions, it requires a linearity assumption to test for complementarities since the individual treatments will now spend more on each treatment than the joint one and the treatments are technically different across each treatment arm (see List et al. (2012) for an example of such a research design in schools in Chicago). Thus, a functional form assumption is unavoidable for simultaneously testing complementarities and doing cost-effectiveness calculations in a study with three treatment arms. Our design prioritized testing for complementarities between two identical policies as implemented, which is the textbook definition of complementarities.

[36] The linearity assumption is plausible here for 3 reasons. First, the grant spending was not for infrastructure or teachers (which could be 'lumpy' and subject to non-linearities in impact) but for books and materials, which would vary more continuously. Second, we are not aware of any study that has found evidence of non-linearities in the impact of school grants. Third, we find no heterogeneity of the impact of either Grants of Combination by enrollment (Table 7: Column 8). The 5-95 percentile range of school enrollment ranged from 235 to 2,602 students, yielding a range of USD 1,300 to USD 16,000 in grant value across schools in this range. Thus, if there were meaningful economies of scale and non-linearities in the use of inputs, we would expect to see some heterogeneity by enrollment, which we do not.

(0.23 vs 0.10), but this difference is not significant ($p = 0.39$). These calculations suggest that we cannot rule out the possibility that spending all the money on incentives may have been as cost-effective as spending on a combination of inputs and incentives.[37]

This result is even stronger when we use estimates of treatment effects from the high-stakes exams (which may provide better comparability with existing studies on teacher incentives). 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. Performing the same exercise as above, we now see that the point estimate of β_3 is considerably *less* than 3.45 * β_2 in Table 4: Panel B (0.36 vs 0.72), and the difference is significant ($p = 0.09$). These results suggest that spending all the money on incentives may be as or more cost-effective than spending on a combination of inputs and incentives at the current margin (where input spending is considerable and incentive spending is zero).

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. Specifically, these could be replaced with a cost-neutral alternative that has a lower base increase but greater performance-linked pay.[38] 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.[39] 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 2.9 in Combination schools (including the input cost but not the incentive cost).[40]

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

[37] While there is less evidence to motivate a functional form for the relationship between the extent of teacher incentives and test score gains, one piece of suggestive evidence for linearity comes from Muralidharan (2012). The paper finds that individual teacher incentives strongly outperform group incentives over five years, but effects are comparable if the group incentive treatment is coded as $1/n$ as the individual incentive treatment (where n is the number of teachers in the group incentive schools). Thus, the estimated treatment effect was proportional to the value of the incentives teachers faced at the individual level — suggesting a linear dose-response relationship.

[38] 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 et al. (2018).

[39] 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.

[40] With a linear dose-response relationship between bonus size and performance, the cost effectiveness of incentives can be increased considerably by increasing the mean-preserving spread of pay (increasing the share of the bonus). If we were to spend all the money from the combination program on incentives, the cost per 0.1σ per student would fall to 0.63 USD.

(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 7.14/0.1σ (Kremer, Miguel, & Thornton, 2009); contract teachers and streaming in Kenya, with a cost of USD 5/0.1σ (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 1.59/0.1σ (Glewwe et al., 2010).[41] 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, the complementarities of teacher incentives with inputs suggest that improving teacher incentives may also improve the effectiveness of existing school inputs. Thus, our 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 high, and on incentives is zero, the marginal return of improving the latter may be higher.

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.

[41] 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. We only include estimates from peer-reviewed published studies.

The evidence of complementarities suggests that there may be multiple binding constraints to improving learning 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.[42]

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, & 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 and 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, and programs for 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.[43]

[42] 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.

[43] 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 may be for governments to partner with committed and credible local third-party organizations (like Twaweza) to conduct

References

Andrabi, T., Das, J., Khwaja, A. I., Ozyurt, S., & Singh, N. (2018). *Upping the ante: the equilibrium effects of unconditional grants to private schools* (Policy Research working paper No. 8563). World Bank.

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.

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.

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.

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.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006, March). Missing in action: Teacher and health worker absence in developing countries. *Journal of Economic Perspectives*, *20*(1), 91-116.

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.

the independent measurements on the basis of which performance-pay schemes can be implemented.

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. (2016). Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research*, *86*(3), 719–755.

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.

Gneezy, U., List, J. A., Livingston, J. A., Sadoff, S., Qin, X., & Xu, Y. (2017, November). *Measuring success in education: The role of effort on the test itself* (Working Paper No. 24004). National Bureau of Economic Research. doi: 10.3386/w24004

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).

Ho, A. D., Lewis, D. M., & MacGregor Farris, J. L. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice*, *28*(4), 15–26.

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.

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., 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.

List, J. A., Livingston, J. A., & Neckermann, S. (2012). *Harnessing complementarities in the education production function.* (mimeo)

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.

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

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

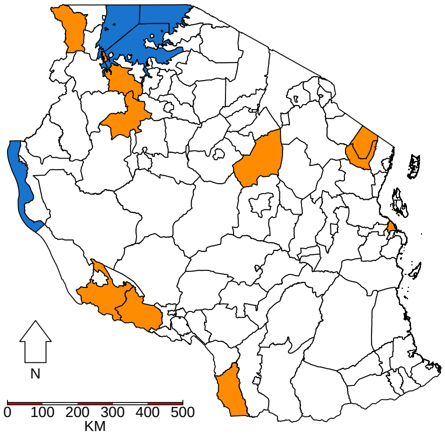
World Bank. (2012). *Tanzania service delivery indicators* (Tech. Rep.). Washington D.C.: World Bank.

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: Sampling and Experimental Design



		Incentives	
		No	*Yes*
Inputs	*No*	140	70
	Yes	70	70

Note: We drew a nationally representative sample of 350 schools from a random sample of 10 districts in Tanzania (left panel). These schools were randomly assigned to treatment and control groups as shown in the right panel.

Figure 2: Timeline

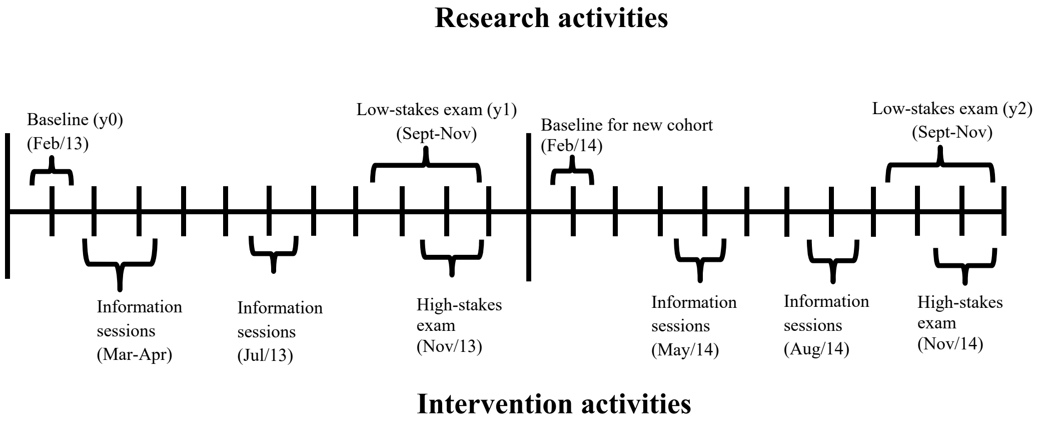

Research activities
Baseline (y0)
(Feb/13)
Low-stakes exam (y1)
(Sept-Nov)
Baseline for new cohort
(Feb/14)
Low-stakes exam (y2)
(Sept-Nov)
Information sessions (Mar-Apr)
Information sessions (Jul/13)
High-stakes exam (Nov/13)
Information sessions (May/14)
Information sessions (Aug/14)
High-stakes exam (Nov/14)
Intervention activities

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

	(1)	(2)	(3)	(4)	(5)
	Combination	Grants	Incentives	Control	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)	(4)	(5)	(6)	(7)
	Grants schools			Combination schools			Diff.
	Year 1	Year 2	Average	Year 1	Year 2	Average	(6)-(3)
Admin.	1,773.07	2,069.72	1,912.14	1,995.24	2,023.31	2,009.28	93.60
	(148.29)	(199.23)	(126.52)	(138.95)	(167.94)	(129.21)	(165.50)
Students	622.45	456.27	533.80	450.50	409.02	429.76	-110.96
	(94.69)	(82.08)	(64.16)	(82.64)	(65.03)	(49.55)	(75.38)
Textbooks	3,858.69	1,315.83	2,585.75	3,774.74	1,278.87	2,526.80	-65.52
	(257.56)	(172.39)	(154.05)	(192.57)	(192.66)	(140.58)	(181.79)
Teaching aids	1,761.43	2,132.32	1,947.61	2,029.13	1,831.09	1,930.11	-8.25
	(126.53)	(190.00)	(118.45)	(115.84)	(157.49)	(96.41)	(133.44)
Teachers	0.00	3.36	1.68	2.74	0.00	1.37	-0.29
	(0.00)	(3.36)	(1.68)	(1.97)	(0.00)	(0.98)	(1.90)
Construction	60.35	69.76	65.49	98.13	67.31	82.72	16.78
	(36.58)	(61.16)	(35.33)	(51.42)	(39.29)	(37.59)	(50.23)
Total Expenditure	8,075.99	6,047.26	7,046.46	8,350.48	5,609.62	6,980.05	-78.44
	(318.42)	(352.57)	(238.98)	(254.66)	(352.11)	(241.74)	(319.79)
Unspent funds	1,924.01	3,952.74	2,953.54	1,649.52	4,390.38	3,019.95	78.44
	(318.42)	(352.57)	(238.98)	(254.66)	(352.11)	(241.74)	(319.79)
Total Value	10,000.00	10,000.00	10,000.00	10,000.00	10,000.00	10,000.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Mean grant expenditure per student of school grants in Grants schools (in TZ Shillings). *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. Column (7) shows the difference (after taking into account the randomization design, i.e., the stratification dummies) between the average spending in Combination schools and the average spending in Grants schools. None of the differences are significant at the 10% level. 1 USD = 1,600 TZ Shillings.

Table 3: Treatment effects on expenditure

	(1)	(2)	(3)	(4)	(5)
	Grant exp.	Other school exp.	Total school	Household exp.	Total exp.
			[(1)+(2)]		[(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 1 for grant expenditure per student, other school expenditure per student, total school expenditure per student, and household reported expenditure on education. All in TZ Shillings. 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. The coefficient for Incentives schools is not exactly zero in Column 1 due to school controls. 1 USD =1,600 TZ Shillings. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Treatment effects on test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined	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.	.	.	.	.	46,883	46,879	46,879	46,879
$\beta_5 := \beta_3 - \beta_2$	.	.	.	.	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	0.10	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 2 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). For Panel A, in the first year the weights of the three subjects to the PCA index are: 0.35 for Kiswahili, 0.3 for English, and 0.35 for Math. In the second year the weights of the three subjects to the PCA index are: 0.35 for Kiswahili, 0.31 for English, and 0.34 for Math. For Panel B, in the second year the weights of the three subjects to the PCA index are: 0.38 for Kiswahili, 0.24 for English, and 0.38 for Math. 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. Sample sizes are larger in year 2 because the research team had more resources to prevent attrition. See Table A.7 for a version without school and household controls. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: 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	0	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 2 for science Z-scores in focal grades (Grd 1 - Grd 3) using data from low-stakes tests conducted by the research team. Sample sizes are larger in year 2 because the research team had more resources to prevent attrition. 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 6: Treatment effects on per student textbook expenditure by grades

	(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 1 on textbook expenditure per student for grades 4-7 (Column 1), grades 1-3 (Column 2), and the difference between them (Column 3). Expenditure in grades 4-7 are show in Column 1, expenditure 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

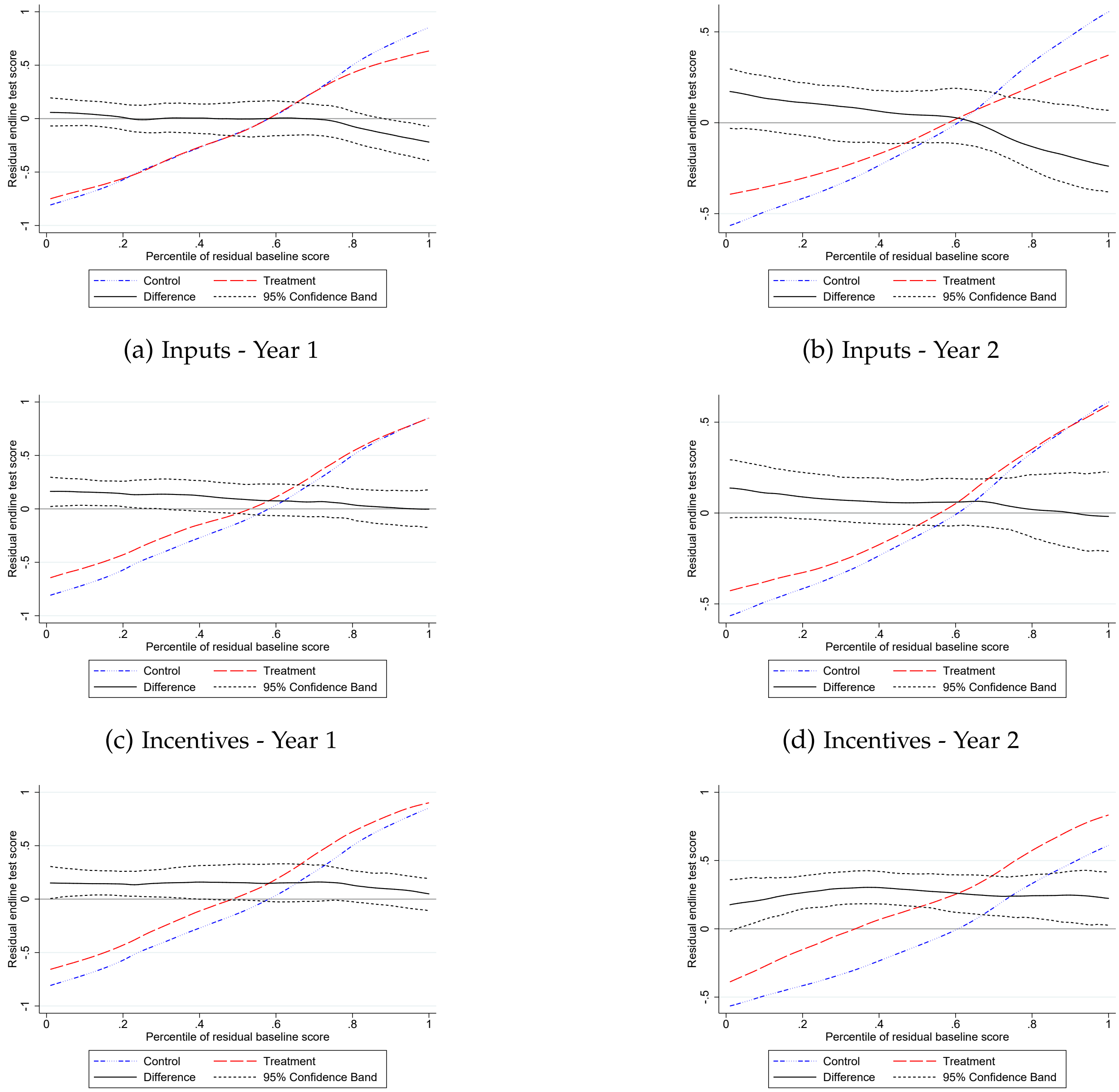
Table 7: Heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Student			Teacher			School		
	Male	Age	Lagged score	Male	Salary	Motivation	Facilities	Enrollment	Management
Grants*Covariate	0.02	0.00	-0.06**	-0.25**	0.00	0.12	0.08	0.11	0.07
	(0.04)	(0.01)	(0.03)	(0.11)	(0.00)	(0.13)	(0.07)	(0.07)	(0.08)
Incentives*Covariate	-0.07*	-0.00	-0.01	-0.01	-0.00	-0.00	0.14**	-0.04	-0.07
	(0.04)	(0.01)	(0.02)	(0.10)	(0.00)	(0.12)	(0.07)	(0.07)	(0.06)
Combination*Covariate	-0.10**	-0.03*	-0.06**	0.04	0.00	-0.05	0.09	-0.10	0.15**
	(0.04)	(0.01)	(0.03)	(0.12)	(0.00)	(0.10)	(0.07)	(0.07)	(0.06)
N. of obs.	18,581	18,581	18,581	18,581	18,581	18,209	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 at the teacher level: a dummy if the teacher is male; the annual salary; and a dummy if the teacher claims it would choose another career path if they could start over at baseline. The teacher covariates are averaged across teachers in both years. Columns 7-9 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. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A Additional tables and figures

Figure A.1: 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 A.1: Treatment effects 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: Treatment effects 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 1 for household expenditure per student 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	-0.008	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 on the low-stakes test. 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: Treatment effects on test scores on a fixed 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 2 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: Treatment effects 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 2 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: Treatment effects on teacher behavior

	(1)	(2)	(3)	(4)	(5)
				Self-reported	
	Attendance	In-classrom	Tests	Tutoring	Remedial
Grants (α_1)	0.04	0.08	-0.19	0.01	-0.04
	(0.03)	(0.06)	(0.69)	(0.02)	(0.03)
Incentives (α_2)	-0.01	-0.00	1.19*	0.03	-0.06*
	(0.03)	(0.05)	(0.66)	(0.03)	(0.03)
Combination (α_3)	0.00	0.04	-0.14	0.05**	0.03
	(0.02)	(0.05)	(0.58)	(0.02)	(0.02)
N. of obs.	2,260	1,029	2,242	2,260	2,260
Mean of dep. var.	0.79	0.50	9.22	0.091	0.84
$\alpha_4 := \alpha_3 - \alpha_2 - \alpha_1$	-0.029	-0.038	-1.15	-0.00050	0.12
p-value ($\alpha_4 = 0$)	0.46	0.62	0.24	0.99	0.0041***

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 teacher in-classroom presence independently measured by enumerators during a surprise visit in the middle of the school year. This was only measured during the second year of the experiment. Thus, there are fewer observations than in other columns. Column (3) shows the number of tests per period as the dependent variable. Column (4) shows a dummy variable that indicates whether the teacher provided any extra tutoring to students as the dependent variable. Column (5) 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 A.9: 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.10: Inputs for cost-effectiveness calculations

	Direct	Implementation	Total	Low-stakes effect	High-stakes effect
Grants	5.89	1.24	7.13	0	0
Incentives	2.52	4.58	7.1	0	0.21
Combination	8.71	4.58	13.29	0.23	0.36

Direct costs (in USD per student) reflect the cost of the program per se. For Grants, this is the cost of the Grants. In Incentive schools this is the cost of the teacher incentives. In Combination schools is the sum of the grants and the incentives. The Implementation costs (in USD per student) reflect the administrative cost of implementing the program. That is, any administrative cost of transferring the money, the cost of visiting schools for monitoring purposes (and to explain the program), testing teachers in Incentive and Combination schools, and financial audits in Grant and Combination schools. The total cost (in USD per student) is the sum of the direct and the implementation cost. The low-stakes and high-stakes treatment effects are for the PCA index of all subjects in the second year.

Table A.11: Grade Retention

	(1) Combination	(2) Grants	(3) Incentives	(4) Control	(5) p-value all equal
Lower grade than expected (Yr2)	0.09 (0.01)	0.08 (0.01)	0.09 (0.01)	0.10 (0.01)	0.49

This table presents the mean and standard error of the mean (in parentheses) for whether a student is in a lower grade than expected at the end of the second 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). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B 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 complementarities between providing inputs and incentives.

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

$$\max_{e} U_i(e) = W + \lambda_i \Delta L - c_i(e) \tag{3}$$

subject to

$$\begin{aligned} W &= S + b\Delta L && \text{(3a)} \\ \Delta L &= f(e, I) && \text{(3b)} \\ \Delta L &\geq \underline{\Delta L} \geq 0 && \text{(3c)} \end{aligned}$$

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 utility from improving student learning. This could reflect multiple factors including intrinsic motivation, and supervision from parents and officials. 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.[44]

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 B.2a 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 B.2b shows the corresponding learning gains. In the absence of incentives or intrinsic motivation (i.e., $b + \lambda_i = 0$), it is Equation 3c 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.[45] Effort does not change as b increases up to the point where the marginal benefit ($b + \lambda_i$) is equal to the marginal cost of providing effort. This corresponds to the flat region to the left of the thresholds κ_0 and κ_1 in Figure B.2a.

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.[46] Since the binding constraint for effort continues to be Equation 3c, 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 B.2b.

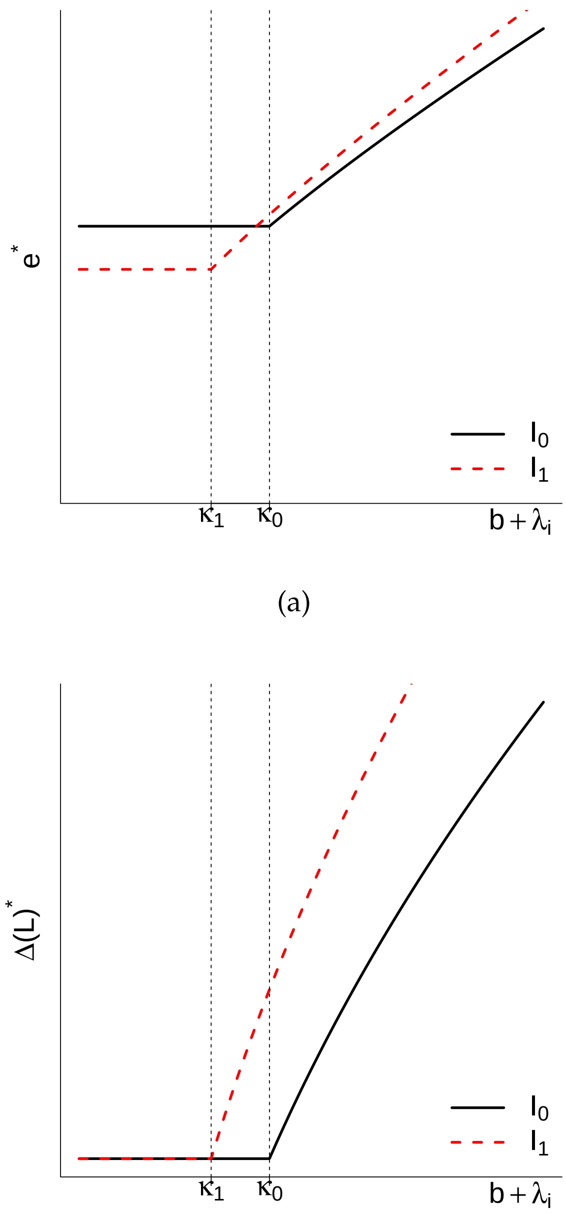
Thus, in the absence of incentives for improving learning outcomes, the relationship between extra inputs and improved test scores will depend on the *distribution* of λ_i in the population of teachers. In settings where λ_i is high for most teachers, improving

[44] $\Delta L \geq \underline{\Delta L} \geq 0$ 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.

[45] If $\lambda_i > 0$ the qualitative results do not change as long as λ_i is low enough that Equation 3c binds, leading to $e^*(I) = e_{min}(I)$.

[46] For instance, Duflo et al. (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 et al. (2017) show, using panel data from India, that reducing pupil-teacher ratios in public schools was correlated with an increase in teacher absence.

Figure B.2: Effort and learning as a function of motivation, at different levels of inputs



(a)

(b)

Note: Figures B.2a and B.2b 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 3c is no longer binding for input level I_c, and therefore $e^*(I_c) = e^*_{mc}(I_c)$ to the right of κ_c.

school inputs may improve test scores.[47] Increasing inputs lowers the threshold (from

[47] For instance, Jackson, Johnson, and Persico (2016) find positive effects of school spending on education

κ_0 to κ_1 in Figure B.2a) that $b + \lambda_i$ needs to exceed for Equation 3c 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 B.2a, 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.[48] 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 3c is not binding, the complementarity in the production function ($f_{eI} > 0$) will also yield complementarities in the policy effects.

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. 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.

outcomes in the US, but teacher utility from improving student outcomes in the US may be higher than in developing countries (partly due to more educated parents and better supervision).

[48] 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.

C Test Design

The tests used in this evaluation were developed by Tanzanian education professionals, and were similar in content to the tests used in the Uwezo annual learning assessment.[49] The consultants developed two types of tests: a low-stakes test that was used for research purposes and a high-stakes test that was used to by Twaweza to determine teacher bonuses. Both types of tests were designed to match the national curriculum.

C.1 Low-stakes test

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. Given the low levels of learning in Tanzania, we conducted one-on-one tests in which a test enumerator sits with the student and guides her/him through a large font test booklet. This improved data quality and also enabled us to capture a wide range of skills in the event the student was not literate. Students are asked to read and answer the test questions to the administrator who records the number of correctly read or answered test items. For the numeracy questions and the spelling questions students were allowed to use pencil and paper. In order to avoid ceiling and floor effects, we requested the consultants to include "easy", "medium", and "hard" items.

The baseline test collected at the beginning of the 2013 school year was based on the standard Uwezo assesment. As such, the test was anchored against the second grade curriculum in Tanzania. As all students in our sample took this test, it was not a grade specific test. However, the test was adaptive and included a wide range of numeracy and literacy skills. Students started the test at different skill levels depending on their grade, and then progressed to either harder or easier skills depending on their performance. The test featured 15 items in Kiswahili, 15 in English, and 33 in math. In Kiswahili and Swahili, the key concepts covered in the test were syllables, reading words, and reading comprehension. In math, the test covered simple counting, number recognition, inequalities of number (i.e. which is greater), addition, subtraction, multiplication, division, and "everyday mathematics" (or culturally appropriate word problems).

During both endline tests (in 2013 and 2014), we tested students based on the grade we expected them to be enrolled. Both of these tests were grade specific tests designed

[49] More information is available at https://www.twaweza.org/go/uwezo

to measure the main competencies outlined in the curriculum.[50] The content of the tests is summarized in in Table C.12. The number of items of each test varied. In the first year the Kiswahili and English tests included 27 items for grade 1, 20 items for grade 2, and 9 items for grade 3. In the second year, the number of items was reduced mainly by dropping items that required students to write (or spell). For math, there were 34 items for grade 1, 24 items for grade 2, and 24 items for grade 3. In the second year, the number of items on the grade 1 math test was reduced. However, we added a number of easier items to the grade 3 test, and left the length of the grade 2 test unchanged.

We standardize the test scores by grade so that the mean and standard deviation in the control group is equal to one. To put our results in context, the average difference between students who are at grade level and those who are not, in Controls schools, is of 1.8σ in Kiswahili and Math, and 4.8σ in English at the end of 2013. In 2014, the average difference is of 1.6σ in Kiswahili and Math, and 4.8σ in English.

C.2 High-stakes test

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 the program required them to calculate the bonuses. Since no bonuses were paid in Grants schools and high-stakes testing is expensive, Twaweza did not conduct the high-stakes tests in Grants schools. Twaweza did, however, administer them in a sample of 40 randomly selected 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.

A number of measures were introduced to enhance test security. First, to prevent test-taking by non-target grade candidates, students could only be tested if their name had been listed and their photo taken at baseline. Second, there were ten versions of the tests to prevent copying and leakage; each student was assigned a randomly generated number from a table to identify the test version, with the choice of the number based on day of the week and the first letter of the student's name. Finally, tests were handled, administered, and scored by Twaweza teams without any teacher involvement. Several checks were done ex-post by Twaweza to ensure there was not any cheating on the high-

[50] During the third round of data collection (at the beginning of the 2014 school year) we only tested grade 1 students. The test followed the format of the grade 1 test during the second round of data collection.

stakes test. Twaweza only found one instance of cheating, where passes from one school were moved to another school. This happened in 2013 and after the discovery the school tallies were corrected to reflect to original numbers of students passing. The corrected data have been used in the analysis for this paper.

C.3 Comparability of tests

The high-stakes tests were designed by the same Tanzanian education professionals as the low-stakes test using the same test-development framework. As a result, the subject order, question type, and phrasing was similar across both tests. The main difference is the high-stakes test is shorter, measuring only the skills necessary to calculate whether students passed the proficiency threshold or not. The low-stakes test covered more skills and as a result had more questions in each section (Kiswahili, English and math) to avoid bottom- and top-coding. Consequently, the high-stakes test was shorter (about 15 minutes) than the low-stakes test (40 minutes).

The content of both rounds of the high-stakes test was similar. The specific skills tested are outlined in Table C.12.

Although the content between the two types of test is similar, there are a number of important differences in the test administration across the low- and high-stakes tests. As mentioned above, the low-stakes test took longer (40 minutes) than the high-stakes test (15 minutes). The low-stakes test had more questions in each section (Kiswahili, English and math) to avoid bottom- and top-coding, and also included an "other subject" module at the end to measure potential spillover effects. In order to keep field logistics and costs manageable, the high-stakes test was designed to secure the data required for the bonus calculations while keeping per-student test time within an upper limit of 20 minutes.

Further, even though both tests were administered individually to students, the testing environment was different. Low-stakes tests were administered by taking sampled students out of their classroom during a regular school day. In contrast, the high-stakes test was more "official" as all students in Grades 1-3 were tested on a prearranged test day. On the test day, students in other grades would sometimes be sent home to avoid distractions. Extra-curricular activities were also canceled during the Twaweza test. In addition, most schools used the incentivized test as the end of year test. However, the Twaweza test team was n charge of all the test proceedings.

Table C.12: Comparison of low-Stakes and high-Stakes test content

	Low- Stakes						High-stakes		
	Year 1			Year 2			Both Years		
	Kiswahili			Kiswahili			Kiswahili		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
Syllables	+	-	-	+	+	+	+	-	-
Words	+	+	-	+	+	+	+	+	-
Sentences	+	+	-	+	+	+	+	+	-
Writing words	+	+	+	-	-	-	-	-	-
Reading one paragraph	-	+	+	-	+	+	-	+	-
Reading comprehension	-	-	+	-	-	+	-	-	+
	English			English			English		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
Letters	+	-	-	+	+	+	+	-	-
Words	+	+	-	+	+	+	+	+	-
Sentences	+	+	-	+	+	+	+	+	-
Writing words	+	+	+	-	-	-	-	-	-
Reading One paragraph	-	+	+	-	+	+	-	+	-
Reading Comprehension	-	-	+	-	-	+	-	-	+
	Math			Math			Math		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
Counting	+	-	-	+	+	+	+	-	-
Number identification	+	-	-	+	+	+	+	-	-
Inequality of numbers	+	+	-	+	+	+	+	+	-
Addition	+	+	+	+	+	+	+	+	+
Subtraction	+	+	+	+	+	+	+	+	+
Multiplication	-	+	+	-	+	+	-	+	+
Division	-	-	+	-	-	+	-	-	+

Note: The Table summarizes the test content for each subject across different grades and data collection rounds. Both high-stakes and low-stakes tests were developed using the same test-development framework as the Uwezo national assessments. The main difference between the high-stakes and low-stakes test is the high-stakes test is designed to measure proficiency so the test has a variety of stopping rules to reduce testing time.

Table C.13: Bottom- and top-coding

	% of students
Kiswahili low-stakes: Bottom-coded	3.59
Kiswahili high-stakes: Bottom-coded	6.94
Kiswahili low-stakes: Top-coded	8.46
Kiswahili high-stakes: Top-coded	14.19
English low-stakes: Bottom-coded	7.51
English high-stakes: Bottom-coded	24.17
English low-stakes: Top-coded	0.03
English high-stakes: Top-coded	1.30
Math low-stakes: Bottom-coded	0.02
Math high-stakes: Bottom-coded	1.11
Math low-stakes: Top-coded	0.53
Math high-stakes: Top-coded	4.47

Percentage of students that are either bottom-coded (score zero) or top-coded (perfect score) across subjects and tests.