INPUTS, INCENTIVES, AND COMPLEMENTARITIES IN EDUCATION: EXPERIMENTAL EVIDENCE FROM TANZANIA*

ISAAC MBITI
KARTHIK MURALIDHARAN
MAURICIO ROMERO
YOUDI SCHIPPER
CONSTANTINE MANDA
RAKESH RAJANI

We present results from a large-scale randomized experiment across 350 schools in Tanzania that studied the impact of providing schools with (i) unconditional grants, (ii) teacher incentives based on student performance, and (iii) both of the above. After two years, we find (i) no impact on student test scores from providing school grants, (ii) some evidence of positive effects from teacher incentives, and (iii) significant positive effects from providing both programs. Most important, we find strong evidence of complementarities between the programs,

*We are grateful to Joseph Mmbando who superbly oversaw the implementation team, and Twaweza staff and management for their support. We are especially grateful to Lawrence Katz (the editor) and five anonymous referees for detailed comments. We also 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 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 preanalysis plan are available at https://www.socialscienceregistry.org/trials/291. Financial support from the Asociación Mexicana de Cultura, A.C., is gratefully acknowledged by Romero.

The Quarterly Journal of Economics (2019), 1627–1673. doi:10.1093/qje/qjz010.
Advance Access publication on April 24, 2019.

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 (the default policy) with improved teacher incentives could substantially increase the cost-effectiveness of public spending on education. *JEL* Codes: C93, H52, I21, M52, O15.

I. 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 (Glewwe and Muralidharan 2016; Mbiti 2016). 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 et al. 2006). Thus, while school inputs may improve learning when teacher effort is high (due to intrinsic motivation or external incentives/monitoring), they may be less effective when teacher effort is low. Conversely, the 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 article 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 study the impact of providing these programs jointly. The first program aimed to alleviate resource constraints by providing schools with grants of 10,000 Tanzanian Shillings (TZS) (∼US$6.25 at the time of the study) per student, effectively doubling discretionary school resources.[1] The second program aimed to improve teacher

1. The government's capitation grant policy aimed to provide schools with TZS 10,000/student. The program we study provided schools with another TZS

motivation and effort by providing teachers with performance-based bonuses—based on the number of their students who passed basic tests of math, Kiswahili (the local language), and English. A teacher with average enrollment could earn up to 125% of their monthly base pay as a bonus.

We conducted the experiment in a large nationally representative sample of 350 public schools (and more than 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. All programs were implemented by Twaweza, a leading Tanzanian nonprofit 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 a 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 one and two years.

Second, we find mixed evidence on the impact of teacher performance pay on student learning. On the low-stakes tests conducted by the research team, student 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-administered tests in math, Kiswahili, and English—the outcome on which teacher bonuses were based. Overall, scores

10,000/student over and above this grant, effectively doubling the school grant. Teacher salaries were paid directly by the government and did not pass through the schools. Thus, these grants (from the government and the program) were the main source of discretionary funding available to schools. Including teacher salaries, the grants led to a 16% increase in net school spending.

on high-stakes tests were 0.21σ higher in treated schools after two years. As specified in our preanalysis plan, the analysis in this article 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 IV.B.

Third, students in Combination schools, which received school grants and teacher incentives, had significantly higher test scores in all subjects on 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 were significantly greater than the sum of the gains in the Grant and Incentives schools in the three subjects (math, Kiswahili, and English). Using a composite measure of test scores across subjects, the "interaction" effect between school inputs and teacher incentives was equal to 0.18σ ($p < .01$). 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.[2] 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.

Although 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

2. 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 IV.D).

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 from the Combination program on a larger teacher incentive program instead.[3]

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 nonmonetary 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 nonfinancial 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 contribution is to experimentally establish the existence of complementarities across policies to improve learning outcomes. Although several field experiments have employed factorial (or cross-cutting) designs that in principle could be used to test for such complementarities, these studies have usually been underpowered to detect economically meaningful complementarities. 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 2011; Gilligan et al. 2018). Overall, as noted in a recent meta-analysis of education studies, "[There are] few experiments [in

3. This is because implementation costs were a much larger fraction of total costs in the Incentive program. 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 V.C).

education] with fully factorial designs that allow for strong experimental tests of [complementarities]" (McEwan 2015, 376).[4]

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 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 et al. 2018). Our results replicate the findings on the nonimpact of providing additional school inputs, but we also show that the inputs can improve learning when combined with teacher incentives.[5]

Third, 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 and Murnane 2016). Our results are consistent with this view and with results from Lavy (2002, 2009), Glewwe, Ilias, and Kremer (2010), Muralidharan and Sundararaman (2011), Duflo, Hanna, and Ryan (2012), Contreras and Rau (2012), and Muralidharan (2012) who find that various forms of performance-linked pay for teachers in low- and middle-income countries improved student test scores.[6]

Finally, our results may be relevant to the literature on the effectiveness of development aid. 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

4. There is a parallel literature on dynamic complementarities between sequential human capital investments over time (Cunha and Heckman 2007; Malamud, Pop-Eleches, and Urquiola 2016; Johnson and Jackson 2017). Our article 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 and Duflo 2005).

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

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

resources well) (Burnside and Dollar 2000), but these results are not very robust (Easterly, Levine, and Roodman 2004). Our results finding no impact of inputs on their own and strong complementarities between inputs and incentives provide 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 $15 billion of aid spending annually (OECD 2016).

An important policy challenge for global development agencies and federal governments in large countries 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 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.[7]

II. Context and Interventions

II.A. Context

Our study is set in Tanzania, which is the sixth largest African country by population and home to more than 50 million people. Partly due to the abolishment in 2001 of school fees in public primary schools, Tanzania has made striking progress toward universal primary education with net enrollment growing from 52% in 2000 to over 94% in 2008 (Valente 2015). Yet despite this increase in enrollment, learning levels remain low. In 2012, nationwide learning assessments showed that less than one-third of grade 3 students were proficient at a grade 2 literacy level in Kiswahili (the national language and medium of instruction in primary schools) or in basic numeracy. Proficiency in English (the medium of instruction in secondary schools) was especially limited, with less than 12% of grade 3 students able to read at a grade 2 level in English (Uwezo 2013; Jones et al. 2014).

Despite considerable public spending on education, budgetary allocations to education (and actual funds received by schools)

7. All appendix tables, figures, and other supplementary materials are in the Online Appendix.

have not kept pace with the rapid increases in enrollment.[8] 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 two 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.

II.B. Interventions and Implementation

The interventions studied in this article 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

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

the research team) implemented the interventions as part of a randomized controlled trial to facilitate learning about the program's impacts. Twaweza also worked with government officials to ensure smooth implementation of the program and evaluation. The interventions are described below.

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 capitation grant program.[9] 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.[10] 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

9. For instance, capitation grant rules do not allow these funds to be used to augment teacher salaries or hire new teachers. The Twaweza grants program had the same guidelines.

10. On average, schools received only around 60% of the stipulated grant value 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.

displaying summary financial statements in a public area in the school.

The grant value was sizable. 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 four or five textbooks/student. Overall, Twaweza disbursed ∼US$350,000/year to the 70 schools in the 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.

2. Teacher Performance Pay (Incentives) Program. The teacher performance pay program provided cash bonuses to teachers based on the performance of their students on independent learning assessments conducted by Twaweza. Given Twaweza's emphasis on early grade learning, the program was limited to teachers in grades 1, 2, and 3 and focused on numeracy (mathematics) and literacy in English and Kiswahili. For each of these subjects, an eligible teacher earned a TZS 5,000 (∼US$3) bonus for each student who passed a simple, externally administered, grade-appropriate test based on the national curriculum. In addition, the head teacher was paid TZS 1,000 (∼US$0.6) for each subject test a student passed.[11]

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.[12] 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

11. 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 feature bonuses for head teachers along the lines implemented in the KiuFunza project.

12. 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).

clarity. Because 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 simpler to implement compared with more complex systems based on calculating measures of student and teacher value added.

There are important limitations to such a threshold-based design (Ho, Lewis, and 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 and Schanzenbach 2010). In addition, such a design may be unfair to teachers who serve a large fraction of students from disadvantaged backgrounds, who may be further behind the passing standard. While Twaweza was aware of these limitations, they took a considered decision to keep the formula simple in the interest of transparency, simplicity of explaining to teachers, and ease of implementation.[13] Furthermore, because 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 and subjects. 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 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 10 versions of the high-stakes tests and randomly assigned these to students within a classroom. To prevent teachers from gaming the system

13. In the United States, 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 added for similar reasons.

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.

3. Combination Arm. Schools assigned to the combination arm received both the capitation grant and teacher incentive programs with identical implementation protocols.

III. Research Design

III.A. Sampling and Randomization

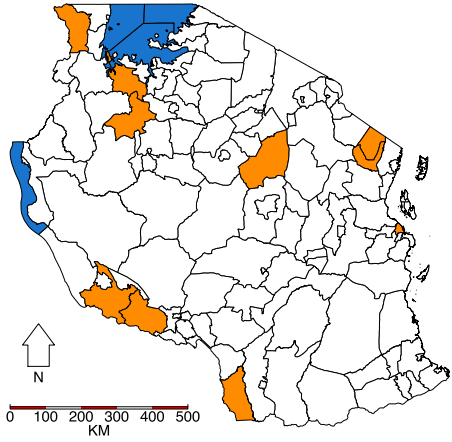
We conducted the experiment in a nationally representative sample of 350 public schools across 10 districts in mainland Tanzania.[14] We first randomly sampled 10 districts from mainland Tanzania, and then randomly sampled 35 schools within each of these districts to get a sample of 350 schools (Figure I). Within each district, seven schools were randomly assigned to receive capitation grants, seven schools to receive teacher incentives, and seven schools to receive both grants and incentives. The remaining 14 schools did not receive either program and served as our control group. Thus, over the 10 sampled districts, the study had a total of 70 schools in each of the three treatment arms (Grants, Incentives, and Combination) and 140 schools in the control group (Figure I).

III.B. 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.[15] Although most categories of school expenditure are difficult to map

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

15. 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 and Smith 2004; Gurkan, Kaiser, and Voorbraak 2009). These data do not include teacher salaries



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

FIGURE I

Sampling and Experimental Design

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.

onto specific grades, we collected data on textbook expenditures at the grade and subject level because this is a substantial expenditure item that can be easily assigned to a specific grade.

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

For data on student learning outcomes, we sampled and tested 10 students from each focal grade within each school, and followed these 30 students over the course of the study. We refer to these as low-stakes (or nonincentivized) tests because 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 from each school (5 each from grades 2 and 3) to conduct household surveys. These 3,500 household surveys were used to collect information on household characteristics,

since salaries are paid directly by the government and do not pass through the school.

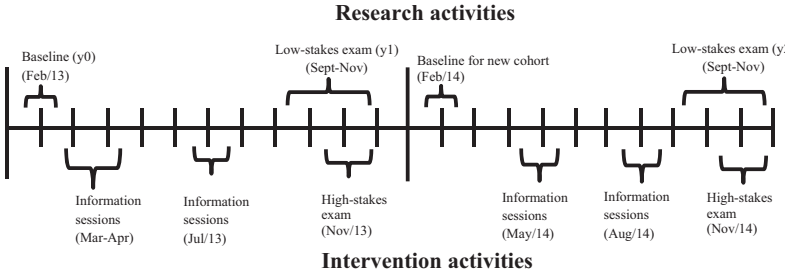


FIGURE II
Timeline.

educational expenditures, and nonfinancial educational inputs at the household level (such as helping with homework).[16]

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, which enables us to compute 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 because 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 Online Appendix C for details on the design and implementation of the low- and the high-stakes tests.

Figure II 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 preanalysis plan are available at: https://www.socialscienceregistry.org/trials/291.

16. 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 because 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.

III.C. 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 I). Table I also provides summary statistics on the (representative) study population. The average student is nine years old and ∼50% are male (Panel A). The schools are mostly rural (85%), mean enrollment is ∼730, and class sizes are large—with an average of more than 55 students per teacher (Panel C).[17] Teachers in our sample were ∼two-thirds female, ∼40 years old, and had ∼15 years of experience; ∼40% of them did not have a teaching certificate (Panel D).

Attrition on the low-stakes tests conducted by the research team is balanced across treatment arms and is low—we were able to track around 90% of students in both years, with slightly lower attrition in the second year (last two rows of Table I, 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 (Online Appendix Table A.1). We therefore present bounds of treatment effects on high-stakes tests, using the approach of Lee (2009).

III.D. Empirical Strategy

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

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

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

17. Thus, total enrollment in study schools was more than 250,000 (350 x ∼730). Total enrollment in the focal grades for the study was a little over 120,000 students.

TABLE I

SUMMARY STATISTICS ACROSS TREATMENT GROUPS AT BASELINE (FEBRUARY 2013)

	Combination (1)	Grants (2)	Incentives (3)	Control (4)	p-value all equal (5)
Panel A: Students ($N = 13{,}996$)					
Male	0.50	0.49	0.50	0.50	.99
	(0.01)	(0.01)	(0.01)	(0.01)	
Age	8.94	8.96	8.94	8.97	.94
	(0.05)	(0.05)	(0.05)	(0.04)	
Normalized Kiswahili test score	0.05	−0.02	0.06	0.00	.41
	(0.07)	(0.07)	(0.08)	(0.05)	
Normalized math test score	0.06	0.01	0.06	0.00	.59
	(0.06)	(0.06)	(0.07)	(0.05)	
Normalized English test score	−0.02	−0.02	−0.00	0.00	.91
	(0.04)	(0.05)	(0.05)	(0.04)	
Attrited in year 1	0.13	0.13	0.11	0.13	.21
	(0.01)	(0.01)	(0.01)	(0.01)	
Attrited in year 2	0.10	0.10	0.10	0.10	.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	.19
	(0.12)	(0.12)	(0.13)	(0.08)	
Wealth index (PCA)	0.02	0.01	0.00	−0.02	.99
	(0.16)	(0.16)	(0.17)	(0.12)	
Pretreatment expenditure (TZS)	34,198.67	33,423.19	34,638.63	36,217.09	.50
	(4,086.38)	(3,799.66)	(4,216.98)	(2,978.25)	

TABLE I

CONTINUED

	Combination (1)	Grants (2)	Incentives (3)	Control (4)	p-value all equal (5)
Panel C: Schools ($N = 350$)					
Pupil-teacher ratio	54.78	58.78	55.51	60.20	.50
	(2.63)	(3.09)	(2.53)	(3.75)	
Single shift	0.60	0.59	0.64	0.63	.88
	(0.06)	(0.06)	(0.06)	(0.04)	
Infrastructure index (PCA)	− 0.08	0.07	− 0.12	0.06	.50
	(0.13)	(0.14)	(0.16)	(0.08)	
Urban	0.16	0.13	0.17	0.15	.85
	(0.04)	(0.04)	(0.05)	(0.03)	
Enrolled students	739.07	747.60	748.46	712.45	.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	.92
	(0.04)	(0.04)	(0.04)	(0.03)	
Age (in 2013)	39.36	39.53	39.05	39.49	.52
	(0.85)	(0.85)	(0.74)	(0.52)	
Years of experience (in 2013)	15.34	15.82	15.11	15.71	.32
	(0.88)	(0.92)	(0.75)	(0.54)	
Teaching certificate	0.62	0.60	0.61	0.57	.50
	(0.04)	(0.04)	(0.04)	(0.03)	

Notes. This table presents the mean and standard error of the mean (in parentheses) 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 1 (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 enrolled in grades 1, 2, and 3 in that year. Column (5) 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 by school for the test of equality.

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:

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

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 (multitasking). To mitigate concerns about the potential for false positives due to multiple hypothesis testing across academic subjects, we 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.

Because 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 $Grants_s$). Furthermore, 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 preanalysis 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 individually, and then test for complementarities (specifically, we test H_0: $\delta_3 - \delta_2 - \delta_1 = 0$) and discuss those results.

IV. Results

IV.A. Capitation Grant Program

How Were Grants Spent? Table II (columns (1)–(3)) presents descriptive statistics on how Grant schools spent their extra funds. Textbooks and classroom teaching aids (like maps, charts, blackboards, chalk) were the largest category of spending, jointly accounting for ∼65% of average spending over the two years. Administrative costs, including wages of nonteaching 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). Because 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, Evans, and Marshak 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).

Did Grants Change Other Spending? Table III 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 TZS ∼2,400 per student, which is around a third of the additional spending enabled by the Grant program (Panel C, columns (1) and (2)).[18]

18. 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.

TABLE II

HOW ARE SCHOOLS SPENDING THE GRANTS?

	Grants schools			Combination schools			Diff.
	Year 1 (1)	Year 2 (2)	Average (3)	Year 1 (4)	Year 2 (5)	Average (6)	(6)−(3) (7)
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)

Notes. Mean grant expenditure per student of school grants in Grants schools (in TZS). *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. US$1 = TZS 1,600.

TABLE III

TREATMENT EFFECTS ON EXPENDITURE

	Grant exp. (1)	Other school exp. (2)	Total school [(1)+(2)] (3)	Household exp. (4)	Total exp. [(3)+(4)] (5)
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 ($\alpha_4 = 0$)	.50	.44	.33	.77	.45
$\alpha_3 - \alpha_1$	258.70	995.70	1,254.41	−367.27	887.14
p-value ($\alpha_3 - \alpha_1 = 0$)	.51	.39	.28	.83	.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 ($\alpha_4 = 0$)	.35	.15	.33	.30	.16
$\alpha_3 - \alpha_1$	−413.01	421.46	8.45	2,088.59	2,061.10
p-value ($\alpha_3 - \alpha_1 = 0$)	.37	.56	.99	.11	.18

TABLE III

CONTINUED

	Grant exp. (1)	Other school exp. (2)	Total school [(1)+(2)] (3)	Household exp. (4)	Total exp. [(3)+(4)] (5)
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 ($\alpha_4 = 0$)	.78	.13	.19	.42	.18
$\alpha_3 - \alpha_1$	−81.42	713.89	632.47	860.66	1,456.91
p-value ($\alpha_3 - \alpha_1 = 0$)	.80	.29	.39	.46	.30

Notes. 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 TZS. 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 coefficients represent the average effect over both years. The coefficient for Incentives schools is not exactly 0 in column (1) due to school controls. US$1 = TZS1,600. Clustered standard errors, by school, are in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$.

Because average school spending per student (excluding teacher salaries) in the control group was TZS ∼5,200, spending the full grant value of TZS 10,000 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 TZS ∼4,700—almost double the expenditure relative to the control group (Panel C, column (3)). We focus our analysis on discretionary spending at the school level 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%.

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 TZS ∼1,000 per student in the first year, and a larger significant reduction of TZS ∼2,200 per student in the second year ($p = .07$). The main categories of spending where we see cuts are fees, textbooks, and food (Online Appendix Table A.2).[19] 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 TZS ∼3,100/year (Panel C, column 5), a 60% increase relative to mean school spending per student in the control group (excluding teacher salaries).

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 the Grant program's impact on a composite measure of test scores were -0.03σ after one year and 0.01σ after two years (both insignificant; Table IV, Panel A).[20] Offsets

19. 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.

20. Grade retention was not an outcome of interest in our preanalysis 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

TABLE IV
TREATMENT EFFECTS ON TEST SCORES

	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
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 ($\alpha_4 = 0$)	.09	.27	.28	.11	.08	.00	.05	.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 ($\alpha_5 = 0$)	.31	.22	.38	.21	.01	.00	.00	.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 ($\beta_5 = 0$)					.05	.01	.06	.01

TABLE IV

CONTINUED

	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel C: Difference								
$\beta_2 - \alpha_2$					0.09	0.10	0.12	0.17
p-value ($\beta_2 - \alpha_2 = 0$)					.14	.05	.07	.02
$\beta_3 - \alpha_3$					0.03	0.01	0.03	0.12
p-value ($\beta_3 - \alpha_3 = 0$)					.53	.81	.63	.08
$\beta_5 - \alpha_5$					−0.05	−0.09	−0.09	−0.05
p-value ($\beta_5 - \alpha_5 = 0$)					.35	.05	.17	.42

Notes. Results from estimating equation (2) for different subjects at both follow-ups. Control variables include student characteristics (age, gender, grade, and lagged 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 the school is to different facilities, and an indicator for whether the school is single shift), 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.5 for a version without school and household controls. Clustered standard errors, by school, are in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$.

are unlikely to be the main reason for our results, as we do not see any impacts of the grant on test scores even in the first year, when the net increase in discretionary spending per student in Grant schools was three times greater than in the second year (Table III, 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, Kremer, and Moulin 2009 in Kenya, Blimpo, Evans, and Lahire 2015 in Gambia, Das et al. 2013 in India, Pradhan et al. 2014 in Indonesia, and Sabarwal, Evans, and Marshak 2014 in Sierra Leone).

IV.B. 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 IV, Panel A). The composite treatment effect at the end of the first year was 0.06σ ($p = .09$), 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 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 (Online Appendix Table A.4). 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 ($p < .01$), Kiswahili, and English ($p < .05$ for both), and 0.21σ higher ($p < .01$) on the composite measure (Table IV, Panel B).[21]

We now consider possible reasons for the difference in estimated impacts across the two sets of tests. First, the content of

difference in grade retention across the treatment groups (Online Appendix Table A.3)

21. 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.

the tests was very similar and so the differences are unlikely to be explained by test content (see Online Appendix C). Second, we verify that the differences are not driven by changes in sample composition by restricting the analysis of treatment effects on the low-stakes tests to the same 40 control schools that had the high-stakes tests (Online Appendix Table A.6). Third, Twaweza employed strict security protocols for the high-stakes test (as mentioned in Section II.B), including having 10 different versions of the test that were randomized across students in the same class and having independent proctors present for every test. The likelihood of cheating was minimized. Fourth, low-stakes tests were conducted about three weeks before high-stakes tests 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). However, the performance on the low-stakes test does not vary as a function of the number of days between the two tests (Online Appendix Table A.7).[22]

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 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. Qualitative interviews suggest that teachers were more likely to have emphasized the importance of these tests to students (since bonus payments depended on test performance). Hence, students and teachers were likely to have been more motivated by the Twaweza exams.

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

22. Results are unchanged if we include week fixed effects, which is unsurprising since we balanced the timing of the exam across treatment arms (Online Appendix Table A.8). They are also unchanged if we restrict the analysis to schools where the low-stakes tests took place before the high-stakes tests (Online Appendix Table A.9).

in student effort on these tests. The estimated difference in the treatment effects across the two sets of tests (of 0.09–0.17σ) is exactly in line with recent experimental estimates that quantify the role of testing-day student effort on measured test scores (Levitt et al. 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 the 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 preanalysis plan, we focus on the low-stakes tests in this article 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.

IV.C. Combination of Capitation Grant and Teacher Incentives

1. Grant Expenditure and Offsets. Combination schools spent their extra grant funds in a similar manner as those receiving only the grants (Table II, 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 (Table III, 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.[23] On net, Combination schools saw a significant increase in per student discretionary spending of TZS ∼4,600/year (Table III, Panel C column (5)), a 90% increase relative to mean per student spending in control schools.

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 ($p < .05$ in all cases, Table IV, Panel A). After two years, they scored 0.20σ, 0.21σ, and 0.18σ higher on math, Kiswahili, and English and scored 0.23σ higher on the composite measure of learning ($p < .01$ in all cases).[24]

Turning to the high-stakes test scores, at the end of the second year, students in Combination schools scored 0.25σ, 0.23σ, and 0.22σ higher on math, Kiswahili, and English, and scored 0.36σ higher on the composite measure ($p < .01$ in all cases, Table IV, Panel B).[25] 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 ($p < .01$ in all cases, Online Appendix Table A.4). 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 the low-stakes tests (conducted by the research team), students in schools that received both

23. 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).

24. 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. Online Appendix Table A.10 shows the results using only the students who were exposed to the interventions in both years. We find very similar results among this group.

25. Due to the differential attendance rates between Combination and control schools on the high-stakes tests (Online Appendix Table A.1), 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 (Online Appendix Table A.11).

programs had significantly higher test scores than those in control schools.

IV.D. Complementarities across Programs

Using the low-stakes tests 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 IV, 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.

Because 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 (Online Appendix Table A.4), 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.[26] We therefore test for $\alpha_3 = 1.035 * (\alpha_1 + \alpha_2)$ and reject equality ($p < .01$). 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 ($p < .02$). In short, school inputs appear to be effective when teachers have incentives to use them effectively, but not otherwise. Conversely, motivated teachers (either for nonfinancial reasons or through incentives) can be more effective with additional educational inputs.

26. Specifically, per student program spending in Grant, Incentive, and Combination schools was US$5.89, 2.52, and 8.71, respectively. Thus spending in Combination schools was 3.5% higher $\left(\frac{8.71}{5.89+2.52}\right)$ 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), so the additional incentive payment in Combination schools is only 12% higher $\left(\frac{8.71-5.89}{2.52}\right)$.

Although 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 0 (based on Table IV, Panel A), we can interpret the significant difference on the high-stakes tests between Combination and Incentive schools as evidence of complementarities (β_5 in Table IV, Panel B).[27] A second approach is to compare the difference between Combination and Incentive schools (which reflects the impact of the Grant and the complementarities) on the high-stakes and low-stakes tests. We cannot reject that this difference is 0 (β_5 - α_5 in last row of Table IV, Panel C), except for Kiswahili (p-value .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 Incentive and Combination schools, but 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 (2011) and Muralidharan (2012) find greater effects 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) conducted a randomized evaluation of a teacher performance pay program in Uganda and find no impact on learning in schools that had no textbooks but a significant positive impact in schools with textbooks (consistent with our findings in neighboring Tanzania). Finally, Andrabi et al. (2018) find positive effects on learning outcomes from providing unconditional grants to private schools (in contrast with the literature finding no effects

27. This difference is significant even after Lee-bounds based adjustment of confidence intervals for differential attrition (β_4 in Online Appendix Table A.11.)

of such grants on learning outcomes in public schools), which may be explained by private schools having better incentives to use their resources effectively.

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.

IV.E. Other Results

1. Multitasking. An important concern with teacher performance-pay schemes is that such programs could encourage teachers to focus on incentivized subjects at the cost of other subjects or activities; a classic case of the multitasking problem (Holmström and Milgrom 1991). On the other hand, if these programs can improve students' literacy and numeracy skills, they may promote student learning even in other nonincentivized subjects. Thus, the impact of performance-pay on nonincentivized outcomes will depend on the extent to which the effort needed to improve incentivized and nonincentivized outcomes are complements or substitutes (see Muralidharan and Sundararaman 2011 for a more detailed discussion).

We test for these possibilities by looking at impacts on science, a nonincentivized 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 V). Further, mirroring the patterns we see in 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 in math and language in this setting did not hurt learning in other subjects and may have helped it when the gains in math and language were significant (as was the case in Combination schools).

TABLE V

SPILLOVERS TO OTHER SUBJECTS AND GRADES

	Science		Grade 7 PSLE 2013		Grade 7 PSLE 2014	
	Year 1 (1)	Year 2 (2)	Pass (3)	Score (4)	Pass (5)	Score (6)
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 ($\alpha_4 = 0$)	.48	.096	.15	.14	.31	.080

Notes. Columns (1) and (2) estimate equation (2) for science Z-scores in focal grades (1–3) 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. * $p < .10$.

TABLE VI
TREATMENT EFFECTS ON PER STUDENT TEXTBOOK EXPENDITURE BY GRADES

	Grades 4–7 (1)	Grades 1–3 (2)	Difference [(2) − (1)] (3)
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 ($\alpha_4 = 0$)	.72	.19	.10
$\alpha_3 - \alpha_1$	−239.27	304.21	543.48
p-value ($\alpha_3 - \alpha_1 = 0$)	.40	.25	.045

Notes. 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)). The regression includes data from both follow-ups, and therefore coefficients represent the average effect over both years. US$1 = TZS 1,600. Clustered standard errors, by school, are in parentheses. *** $p < .01$.

2. Intraschool 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 TZS 10,000). 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 to 7 (nonstudy 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 VI). 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 V, 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 seventh-grade test scores in 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).[28]

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 United States). 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 Online Appendix Table A.12.[29]

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 VII, with columns (1)–(3), (4)–(6), and (7)–(9) focusing on heterogeneity by student, teacher, and school characteristics, respectively.

28. There is some evidence of complementarities on grade 7 test scores in the second year (p-value .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.

29. This is a robust result. Because this was a dimension on which we expected to find some heterogeneity (as seen in our preanalysis 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. We also examine heterogeneity nonparametrically as a function of baseline test scores and find no evidence of meaningful heterogeneity (see Online Appendix Figure A.1).

TABLE VII

HETEROGENEITY

	Student			Teacher			School		
	Male (1)	Age (2)	Lagged score (3)	Male (4)	Salary (5)	Motivation (6)	Facilities (7)	Enrollment (8)	Management (9)
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

Notes. 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 1 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 they 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 1 if the PCA index for managerial ability of the principal is above the median. Clustered standard errors, by school, are in parentheses. * $p < .10$, ** $p < .05$.

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

V. Discussion

V.A. Theoretical Framework

Our results confirm the lack of impact of inputs on their own and 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 Online Appendix B. The model specifies a production function for test scores (which 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 only under the implicit (and usually unstated) assumption that teachers have nonmonetary

motivation to improve learning should increasing inputs be expected to improve test scores. In contrast, if teachers behave like agents in standard economic models (with disutility of effort and limited nonmonetary 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. Providing incentives to teachers will typically raise the optimal effort when inputs are increased, giving rise to policy complementarities between providing inputs and incentives.

Although this model is not the only possible explanation for our results, it provides an intuitive and parsimonious framework to interpret our experimental results and existing results in the literature. In addition to the 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.[30] The model can explain all of these existing results and helps clarify the importance of teacher motivation (either financial or nonfinancial) in translating school inputs into learning outcomes.

V.B. Mechanisms

As suggested by the model, 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 effects on survey-based measures of teacher attendance, and teacher self-reports (Online Appendix Table A.13). Teacher absence rates are unchanged (consistent with Muralidharan and Sundararaman 2011), and we find little systematic evidence of impact on self-reported data on the number of practice tests given or provision of remedial teaching.

30. For instance, Duflo, Dupas, and Kremer (2015) find that providing primary schools in Kenya with an extra contract teacher led to an increase in absence rates of teachers. Muralidharan and Sundararaman (2013) find the same result 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.

In practice, it is likely that the test score results are driven by increased intensity of 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 informed by prior work showing considerable Hawthorne effects in measuring teacher classroom behavior (Muralidharan and 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 III, 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 ($p = .11$; last row of Table III, Panel B, column (4)). This is consistent with increases in (unobservable) teacher effort in Combination schools, to encourage parents not to 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 (Online Appendix Table A.2). Second, Combination schools spent significantly more per student (TZS 543) on textbooks in incentivized grades (relative to nonincentivized grades) compared to schools that only received the Grants (Table VI).

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.

V.C. 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). Online Appendix Table A.14 provides the direct and implementation costs of all three programs (per student). These are as follows: Grants, US$5.89 and US$1.24 (total of US$7.13); Incentives, US$2.52 and US$4.58 (total of US$7.10); Combination, US$8.71 and US$4.58 (total of US$13.29).

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 US$5.78.

We next consider the issue of scaling. Specifically, what would the effects 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). Spending the full value of the Combination program on inputs would yield a per student input expenditure of US$12.05 (13.29 minus implementation cost of 1.24), which would be 2.05 times greater than the value provided in the Grants treatment (US$5.89). We therefore test $\alpha_3 = 2.05 * \alpha_1$ in Table IV (0.23 versus 0.02), and reject equality ($p = .03$). 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.[31]

31. The linearity assumption is plausible here for three reasons. First, the grant spending was not for infrastructure or teachers (which could be lumpy and subject to nonlinearities 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 nonlinearities in the impact of school grants. Third, we find no heterogeneity of the impact of either Grants or Combination by enrollment (Table VII, column (8)). The 5–95 percentile range of school enrollment ranged from 235 to 2,602 students, yielding a range of US$1,300 to US$16,000 in grant value across schools in this range. Thus, if there were meaningful economies of scale and nonlinearities in the use of inputs, we would expect to see some heterogeneity by enrollment, which we do not.

If we spent the full amount of the Combination program on the Incentive program, the value of the Incentives would be US$8.71 (13.29 minus the implementation cost of 4.58), which is 3.45 times greater than the bonuses provided in the Incentives treatment. Conducting a similar test, the point estimate of α_3 is greater than 3.45 * α_2 in Table IV (0.23 versus 0.10), but this difference is not significant ($p = .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.[32]

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 US$3.38 in Incentive schools and US$3.69 in Combination schools. Performing the same exercise as above, we now see that the point estimate of β_3 is considerably less than 3.45 * β_2 in Table IV, Panel B (0.36 versus 0.72), and the difference is significant ($p = .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 0).

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, across-the-board pay increases could be replaced with a cost-neutral alternative that has a lower base increase but greater performance-linked pay.[33] In such a scenario, the main long-term cost of a teacher incentive program is the

32. Although 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 $\frac{1}{n}$ 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.

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

administrative cost of implementing the program (including costs of independent measurement and recording of student learning) and not the cost of the bonus itself.[34] Using the administrative costs in this study, the cost of increasing test scores by 0.1σ per student would be US$2.18 in Incentive schools and US$2.9 in Combination schools (including the input cost but not the incentive cost).[35]

Overall, these estimates compare well with the estimated cost-effectiveness of several other interventions to improve education in Africa. For instance, some of the interventions with positive impacts on learning reviewed by Kremer, Brannen, and Glennerster (2013) include, a conditional cash transfer in Malawi, with a cost of US$100 per 0.1σ gain per student (Baird, McIntosh, and Özler 2011); scholarships for girls in Kenya, with a cost of US$7.14/$0.1\sigma$ (Kremer, Miguel, and Thornton 2009); contract teachers and streaming in Kenya, with a cost of US$5/$0.1\sigma$ (Duflo, Dupas, and Kremer, 2011, 2015; and teacher incentives in Kenya (evaluated using data from high-stakes tests), with a cost of US$1.59/$0.1\sigma$ (Glewwe, Ilias, and Kremer 2010).[36] 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 and Popova 2016).

Taken together, our results suggest that reforms to teacher compensation structure that reward improved student learning can be highly cost-effective relative to the status quo of education spending, which is largely input-based. Furthermore, 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

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

35. 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 US$0.63.

36. 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.

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 there is no spending on incentives, the marginal return of improving the latter may be higher.

VI. Conclusion

The evidence of complementarities reported in this article 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 effect 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 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 on quality improvement for motivated teachers.

The default pattern of social sector spending in most countries (and in donor-led development assistance programs) is to expand inputs. These include physical inputs and programs for training and capacity building. Our results show that the marginal returns of introducing reforms to better reward improved effort of frontline service providers may be particularly high in settings where inputs are being expanded.

One important caveat in translating these results into policy is that the evidence of positive effects of teacher-incentive programs in developing countries has usually come from studies where implementing the incentive program has been carried out well by a motivated nonprofit organization.[37] However, these are also typically settings of weak state capacity where governments have a difficult time even ensuring adequate teacher

37. These include Muralidharan and Sundararaman (2011); Duflo, Hanna, and Ryan (2012), and this article.

attendance. Thus, implementing teacher performance-pay systems will require considerable investments in implementation capacity. Our results and calculations suggest that this could be a cost-effective investment and that doing so may meaningfully expand state capacity for improved service delivery in developing countries.[38]

University of Virginia, Abdul Latif Jameel Poverty Action Lab, and Institute of Labor Economics
University of California, San Diego; National Bureau of Economic Research; and Abdul Latif Jameel Poverty Action Lab
Instituto Tecnológico Autónomo de México
Twaweza
Yale University
Twaweza

Supplementary Material

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online. Data and code replicating tables and figures in this article can be found in Mbiti et al. (2019), in the Harvard Dataverse, doi:10.7910/DVN/AKVGQ7.

References

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, Selcuk Ozyurt, and Niharika Singh, "Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools," World Bank Policy Research Working Paper 8563, 2018.
Baird, Sarah, Craig McIntosh, and Berk Özler, "Cash or Condition? Evidence from a Cash Transfer Experiment," *Quarterly Journal of Economics*, 126 (2011), 1709–1753.
Banerjee, Abhijit, and Esther Duflo, "Growth Theory through the Lens of Development Economics," *Handbook of Economic Growth*, vol. 1, Philippe Aghion and Steven Durlauf, eds. (Amsterdam: Elsevier, 2005), 473–552.
Birdsall, Nancy, William D. Savedoff, Ayah Mahgoub, and Katherine Vyborny, *Cash on Delivery: A New Approach to Foreign Aid* (Washington, DC: Center for Global Development, 2012).
Blimpo, Moussa P., David K. Evans, and Nathalie Lahire, "Parental Human Capital and Effective School Management: Evidence from The Gambia," World Bank Policy Research Working Paper 7238, 2015.
Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen, "Does Management Matter in Schools?," *Economic Journal*, 125 (2015), 647–674.

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

Burnside, Craig, and David Dollar, "Aid, Policies, and Growth," *American Economic Review*, 90 (2000), 847–868.
Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers, "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *Journal of Economic Perspectives*, 20 (2006), 91–116.
Contreras, Dante, and Tomás Rau, "Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile," *Economic Development and Cultural Change*, 61 (2012), 219–246.
Cunha, Flavio, and James Heckman, "The Technology of Skill Formation," *American Economic Review*, 97 (2007), 31–47.
Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman, "School Inputs, Household Substitution, and Test Scores," *American Economic Journal: Applied Economics*, 5 (2013), 29–57.
de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia," *Quarterly Journal of Economics*, 133 (2018), 993–1039.
Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101 (2011), 1739–1774.
———, "School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools," *Journal of Public Economics*, 123 (2015), 92–110.
Duflo, Esther, Rema Hanna, and Stephen P. Ryan, "Incentives Work: Getting Teachers to Come to School," *American Economic Review*, 102 (2012), 1241–1278.
Easterly, William, Ross Levine, and David Roodman, "Aid, Policies, and Growth: Comment," *American Economic Review*, 94 (2004), 774–780.
Evans, David, and Anna Popova, "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews," *World Bank Research Observer*, 31 (2016), 242–270.
Ganimian, Alejandro J., and Richard J. Murnane, "Improving Education in Developing Countries: Lessons from Rigorous Impact Evaluations," *Review of Educational Research*, 86 (2016), 719–755.
Gilligan, Daniel O., Naureen Karachiwalla, Ibrahim Kasirye, Adrienne Lucas, and Derek Neal, "Educator Incentives and Educational Triage in Rural Primary Schools," NBER Working Paper no. 24911, 2018.
Glewwe, P., and K. Muralidharan, "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications," *Handbook of the Economics of Education*, vol. 5, Eric A. Hanushek, Stephen Machin, Ludger Woessmann, eds. (Amsterdam: Elsevier, 2016), 653–743.
Glewwe, Paul, Nauman Ilias, and Michael Kremer, "Teacher Incentives," *American Economic Journal: Applied Economics* 2 (2010), 205–227.
Glewwe, Paul, Michael Kremer, and Sylvie Moulin, "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, 1 (2009), 112–135.
Gneezy, Uri, John A. List, Jeffrey A. Livingston, Sally Sadoff, Xiangdong Qin, and Yang Xu, "Measuring Success in Education: The Role of Effort on the Test Itself," NBER Working Paper no. 24004, 2017.
Gurkan, Asli, Kai Kaiser, and Doris Voorbraak, "Implementing Public Expenditure Tracking Surveys for Results: Lessons from a Decade of Global Experience," *PREM Notes*, 145 (2009).
Ho, Andrew D., Daniel M. Lewis, and Jason L. MacGregor Farris, "The Dependence of Growth-Model Results on Proficiency Cut Scores," *Educational Measurement: Issues and Practice*, 28 (2009), 15–26.
Holmström, Bengt, and Paul Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, & Organization*, 7 (1991), 24–52.

Johnson, Rucker C., and C. Kirabo Jackson, "Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending," NBER Working Paper no. 23489, 2017.
Jones, Sam, Youdi Schipper, Sara Ruto, and Rakesh Rajani, "Can Your Child Read and Count? Measuring Learning Outcomes in East Africa," *Journal of African Economies*, 23 (2014), 643–672.
Kerwin, Jason Theodore, and Rebecca L. Thornton, "Making the Grade: The Trade-off between Efficiency and Effectiveness in Improving Student Learning," Working Paper, University of Minnesota, 2017.
Kremer, Michael, Conner Brannen, and Rachel Glennerster, "The Challenge of Education and Learning in the Developing World," *Science*, 340 (2013), 297–300.
Kremer, Michael, Edward Miguel, and Rebecca Thornton, "Incentives to Learn," *Review of Economics and Statistics*, 91 (2009), 437–456.
Lavy, V., "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement," *Journal of Political Economy*, 110 (2002), 1286–1317.
———, "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics," *American Economic Review*, 99 (2009), 1979–2011.
Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76 (2009), 1071–1102.
Lemos, Renata, Karthik Muralidharan, and Daniela Scur, "Personnel Management and School Productivity: Evidence from India," Working Paper, University of California, San Diego, 2018.
Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff, "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," *American Economic Journal: Economic Policy*, 8 (2016), 183–219.
Malamud, Ofer, Cristian Pop-Eleches, and Miguel Urquiola, "Interactions between Family and School Environments: Evidence on Dynamic Complementarities?," NBER Working Paper no. 22112, 2016.
Mbiti, Isaac, "The Need for Accountability in Education in Developing Countries," *Journal of Economic Perspectives*, 30 (2016), 109–132.
Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani, "Replication Data for: 'Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania'." Harvard Dataverse, 2019, doi:10.7910/DVN/AKVGQ7.
McEwan, Patrick J., "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments," *Review of Educational Research*, 85 (2015), 353–394.
Muralidharan, Karthik, "Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India," Working Paper, University of California, San Diego, 2012.
Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal, "The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India," *Journal of Public Economics*, 145 (2017), 116–135.
Muralidharan, Karthik, and Paul Niehaus, "Experimentation at Scale," *Journal of Economic Perspectives*, 31 (2017), 103–124.
Muralidharan, Karthik, and Venkatesh Sundararaman, "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India," *Economic Journal*, 120 (2010), F187–F203.
———, "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, 119 (2011), 39–77.
———, "Contract Teachers: Experimental Evidence from India," NBER Working Paper no. 19440, 2013.
Neal, Derek, and Diane Whitmore Schanzenbach, "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics*, 92 (2010), 263–283.

OECD, "Education-Related Aid Data at a Glance," 2016 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, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha, "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia," *American Economic Journal: Applied Economics*, 6 (2014), 105–126.
Ray, D., *Development Economics* (Princeton, NJ: Princeton University Press, 1998).
Reinikka, Ritva, and Nathanael Smith, *Public Expenditure Tracking Surveys in Education* (Paris: UNESCO, International Institute for Educational Planning, 2004).
Sabarwal, Shwetlena, David K. Evans, and Anastasia Marshak, "The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone," World Bank Policy Research Working Paper Series 7021, 2014.
United Nations, "Transforming Our World: The 2030 Agenda for Sustainable Development," Resolution adopted by the General Assembly, 2015.
Uwezo, "Are Our Children Learning? Numeracy and Literacy across East Africa," Uwezo East Africa report, 2013.
———, "Are Our Children Learning?," Uwezo Tanzania Sixth Learning Assessment Report, 2017.
Valente, Christine, "Primary Education Expansion and Quality of Schooling: Evidence from Tanzania," IZA Technical Report, 2015.
World Bank, "Tanzania Service Delivery Indicators," Technical Report, 2012.
———, "Expenditure on Primary as % of Government Expenditure on Education (%)," 2015. Data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SE.XPD.PRIM.ZS?locations=TZ.
———, "Education Statistics (EdStats)," 2017. Data retrieved from, http://datatopics.worldbank.org/education/wDashboard/dqexpenditures.
———, "World Development Report 2018: Learning to Realize Education's Promise," 2018.