FOR ONLINE PUBLICATION

A Randomization Details

This study builds on the sample of 350 schools that participated in the 2013 to 2014 KiuFunza study (see Mbiti et al. (in press) for more details). In the 2013-14 study the 350 schools in the sample were randomly placed into one of four treatment groups: 70 schools received school grants, 70 schools received teacher incentives (using a single threshold design), 70 schools received both grants and incentives, and 140 schools were in the control group. In order to determine teacher awards, incentivized tests were conducted in schools assigned to the incentives treatment or the combination treatment (a total of 140 schools). To faciliate the computation of treatment effects on incentivized tests, we also conducted these tests in 40 control schools.

We take the set of 180 schools where endline "incentivized" tests had been conducted in 2014. Specifically, 70 schools from the incentive arm (labeled C1), 70 schools from the combination arm (C2), and 40 schools from the control arm (C3). We use these tests as the baseline data to implement the teacher incentive schemes in this study. This baseline data is especially important for the Pay for Percentile incentive scheme as we have to split students into groups, and properly seed each contest.

In each district, there were seven schools in C1 (teacher incentives), seven in C2 (combination), and four in C3 (the control group). We randomly assign schools from the previous treatment groups into two new treatments groups (Levels or Pay for Percentile) and a control group. We stratify this randomization by district. However, in order to study the long-term impacts of teacher incentives, we assign a higher proportion of schools in C1 (which involved threshold teacher incentives) to Levels. Similarly, we assign a higher proportion of schools in the control group from the previous experiment (C3) to the control group of this experiment.

For this experiment, we stratify the random treatment assignment by district, previous treatment, and an index of the overall learning level of students in each school.[34] Table A.1 summarizes the number of schools randomly allocated to each treatment arm based on their assignment in the previous experiment. Each district has 18 schools, such that there are six schools in each of the new treatment groups (Levels, Pay for Percentile, and control). Because the study was carried out in 10 districts, overall there are 60 schools in each new treatment group: 30 above the median in baseline learning and 30 below.

[34] We created an overall measure of student learning and categorized schools as above or below the median.

All regressions account for all three levels of stratification: district, previous treatment, and an index of the overall learning level of students in each school.

Table A.1: Treatment allocation

		KiuFunza II			
		Levels	P4Pctile	Control	Total
KiuFunza I	C1	40	20	10	70
	C2	10	30	30	70
	C3	10	10	20	40
	Total	60	60	60	180

B Theoretical Framework

We present a set of simple models to clarify the potential behavioral responses of teachers and schools in our interventions. We first characterize equilibrium effort levels of teachers in both incentive systems, and then impose some additional assumptions and use numerical methods to obtain a set of qualitative predictions about the distribution of teacher effort across students of varying baseline learning levels.

B.1 Basic Setup

In our simple setup, there are different types of students (indexed by l). Students may vary by initial level of learning or by socio-demographic characteristics. Further, each classroom of students is taught by a single teacher, indexed by j. We assume student learning levels (or test scores) at endline is determined by the following process:

$$a_j^l = a_{j(t-1)}^l + \gamma^l e_j^l + v_j^l$$

where a_j^l is the learning level of a student of type l taught by teacher j, and $a_{j(t-1)}^l$ is the student's baseline level of learning.[35] γ^l captures the productivity of teacher effort (e_j^l) and is assumed to be constant across teachers. In other words, we assume teachers are equally capable.[36] v_j^l is an idiosyncratic random shock to student learning. We assume that effort is costly, and that the cost function, $c_l(e_j^l)$, is twice differentiable and convex such that $c_l'(\cdot) > 0$, and $c_l''(\cdot) > 0$.

A social planner would choose teacher effort to maximize the total expected value of student learning, net of the total costs of teacher effort as follows:

$$\sum_j \sum_l \mathbb{E}(a_{j(t-1)}^l + \gamma^l e_j^l + v_j^l) - c_l(e_j^l)$$

The first order conditions for this problem are:

$$\gamma^l = c_l'(e_j^l) \tag{2}$$

for all l and all j. To keep the model simple, we assume teachers are risk-neutral and abstract from multi-tasking concerns. To keep notation simple, we assume all teach-

[35] We assume $a_{j(t-1)}^l$ is an adequate summary statistic for all previous inputs, including past teacher effort.

[36] Barlevy and Neal (2012) also impose this assumption in their basic setup.

ers have identical ability (or productivity); however, this can easily be relaxed without altering the results presented below.

B.1.1 Pay for Percentile

In the Pay for Percentile design there are L rank-order tournaments based on student performance, where L is the number of student types or the number of groupings, such that students in the same group are similar to each other. Under this incentive scheme, teachers maximize their expected payoffs, net of costs, from each rank-order tournament. The teacher's maximization problem becomes:

$$\sum_l \left(\sum_{k \neq j} \left(\pi P(a_j^l > a_k^l) \right) - c_l(e_j^l) \right),$$

where π is the payoff per percentile. The first order conditions for the teacher's problem are:

$$\sum_{k \neq j} \pi \gamma^l f^l (\gamma^l (e_j^l - e_k^l)) = c_l'(e_j^l)$$

for all l, where f^l is the density function of $\varepsilon_{j,k}^l = v_j^l - v_k^l$.

In a symmetric equilibrium, then

$$(N-1)\pi \gamma^l f^l(0) = c_l'(e^l) \tag{3}$$

where N is the number of teachers. Without loss of generality, if the cost function is the same across groups (i.e., $c_l'(x) = c'(x)$), but the productivity of effort varies (γ^l), then the teacher will exert higher effort where he or she is more productive (since the cost function is convex). Pay for percentile can lead to an efficient outcome, as shown by Barlevy and Neal (2012), if the social planner's objective is to maximize total learning and the payoff is $\pi = \frac{1}{(N-1)f^l(0)}$.

B.1.2 Levels

In our Levels incentive scheme, teachers earn bonuses whenever a student's test score is above a pre-specified learning threshold. As each subject has multiple thresholds t, we can specify teacher j's maximization problem as:

$$\sum_l \left(\sum_t \left(C_j^l P(a_j^l > T_t) \frac{\Pi_t}{\sum_l \sum_n C_n^l P\left(a_n^l > T_t\right)} \right) - c_l(e_j^l) \right)$$

where T_t is the learning needed to unlock threshold t payment, Π_t is the total amount of money available for threshold t, and C_n^l is the number of students of type l in teacher n's class.

Assuming the number of teachers (N) is large, then the effect each teacher has on the overall pass rates is negligible. In particular, we assume it is zero (i.e., teacher's ignore the effect of their effort on the overall pass rate). Thus, the first order conditions for the teacher's maximization problem become:

$$\sum_t C_j^l \gamma^l h^l (T_t - a^l_{j(t-1)} - \gamma^l e^l_j) \frac{\Pi_t}{\sum_l \sum_n C_n^l P\left(v_n^l > T_t - a^l_{n(t-1)} - \gamma^l e^l_n\right)} = c_l'(e_j^l) \tag{4}$$

for all l, where h^l is the density function of v_j^l. Although we assume that each individual teacher's effort does not affect the overall pass rate, we cannot ignore this effect in equilibrium. Thus, we can characterize our symmetric equilibrium as:

$$\sum_t C_j^l \gamma^l h^l (T_t - a^l_{j(t-1)} - \gamma^l e^l) \frac{\Pi_t}{\sum_l N C_n^l P\left(v^l > T_t - a^l_{(t-1)} - \gamma^l e^l\right)} = c_l'(e^l) \tag{5}$$

for all l.

B.1.3 Numerical Simulation Set-up

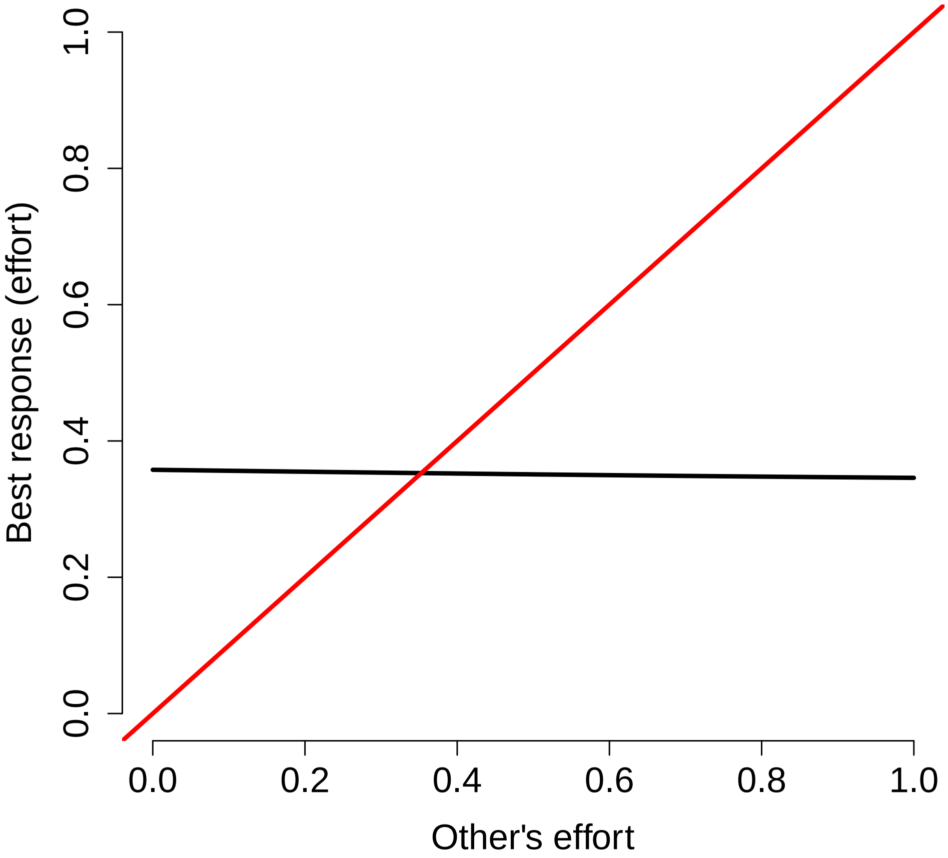
We simulate the equilibrium responses by teachers to both types of incentives in order to better understand teacher behavioral responses to the two treatments in our study. We assume that the teacher's cost function is quadratic (i.e., $c(e) = e^2$), and the shock to student learning follows a standard normal distribution (i.e., $v_i \sim N(0,1)$). We further assume that there are 1,000 teachers, each with their own classroom. Within each class, we assume that student baseline learning levels are uniformly distributed from -4 to 4, in 0.5 intervals. As a result each classroom has 17 students with one student at each (discrete) baseline learning level.[37] We set the reward per student in both schemes at \$1. Therefore, in the Pay for Percentile scheme the reward per contest won is $\$\frac{2}{99}$ (see Section B.1.1) and in the Levels the total reward is \$1 per student. In the multiple threshold scenario the reward is held constant and split evenly across all thresholds. For simplicity, we assume that there are three proficiency thresholds. We first compute the optimal teacher response assuming a single proficiency threshold and then vary the threshold value from -1 to 1. We then compute the multiple threshold case.

[37] In Appendix B.2 we show that our qualitative results are robust to a normal distribution of student baseline learning levels.

B.1.4 Levels Equilibrium

We first simulate equilibrium behavior under the Levels scheme in Figure B.1 below. Using the parameter values and functional forms discussed above, we simulate an individual teacher's best response curve and plot it against the best response of all other teachers using a wide range of initial parameter values. In our simulations we do not observe any non-quasi-concave objective functions for any given ability level. Further, since the curves are smooth, there is no indication that they would violate Brouwer's fixed point theorem. As Figure B.1 shows, in the context of our of simulations, there is only one (rational expectations) equilibrium characterized by Equation 5.

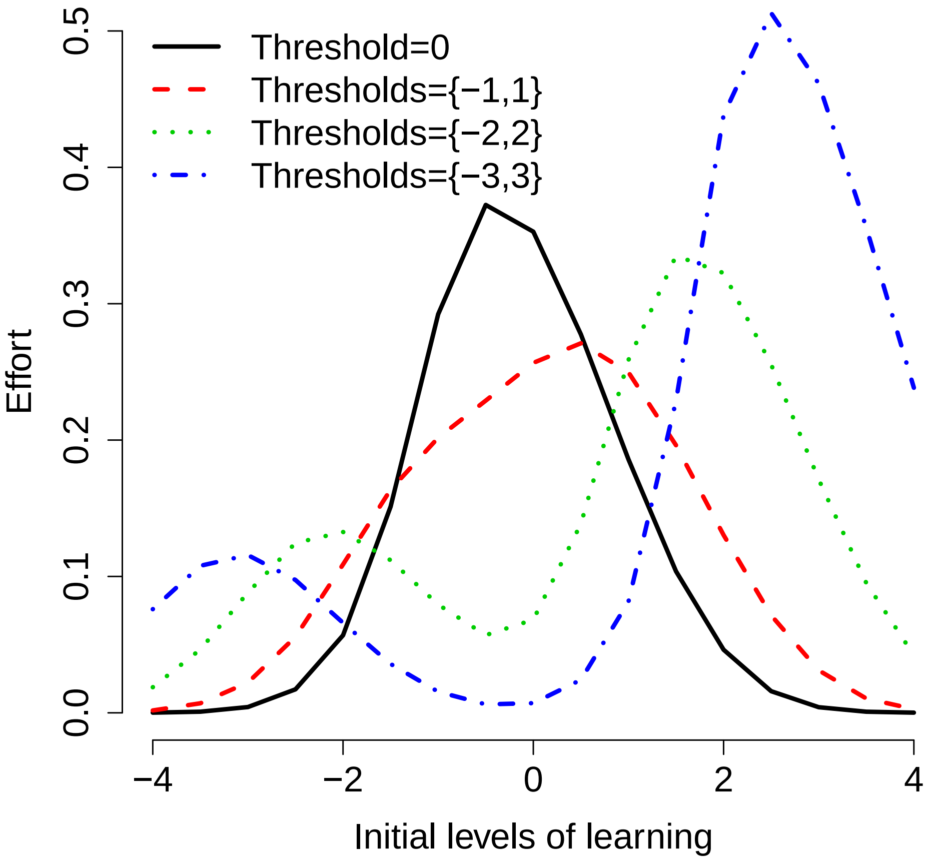
Figure B.1: Teacher $i's$ Best Response curve to other teacher's effort level



Note: An example of a set of best response curves for a given initial parameter values. We assume all teachers are giving the same value of effort for all thresholds except one (but the effort may be different across thresholds). In the x-axis we show the level of effort exerted by all except i in the threshold of interest. In the y-axis we plot teachers i effort level in that thresholds. The black line shows the best response of teacher i to the effort level of other teachers. Therefore, we have a symmetric equilibrium when the black line crosses the red line.

Our simulations also show that the choice of proficiency thresholds is important design decision. If the thresholds are too far apart then teachers may not exert any effort on students who are in between thresholds. This concern can be ameliorated by setting thresholds sufficiently close together as shown below in Figure B.2.

Figure B.2: Threshold Distance and Teacher effort



Note: Assuming a two threshold design, this figure shows the effect of increasing the distance between two thresholds on teacher effort. The distance varies from 0, to 2 (thresholds at -1 and 1), 4 (thresholds at -2 and 2), and 6 (thresholds at -3 and 3).

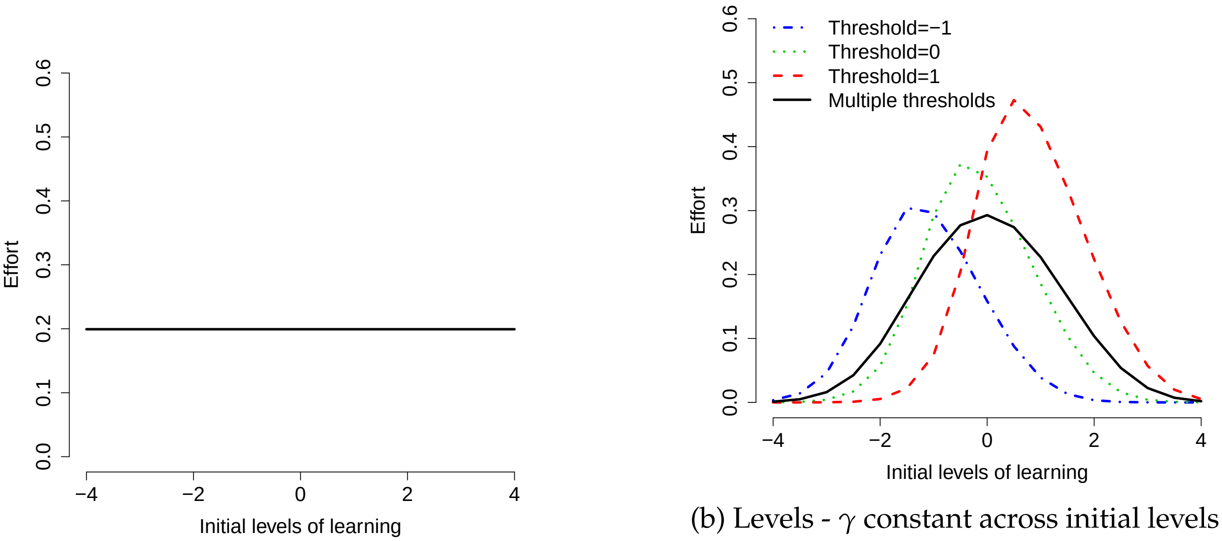
As the equilibrium behavior for teachers under Pay for Percentile was described in detail in Barlevy and Neal (2012), we refer our readers to consult their findings for additional insights.

B.1.5 A Comparison of Optimal Teacher Effort

We compute equilibrium teacher responses under two different stylized scenarios (or assumptions about the productivity of teacher effort in the production function) to illustrate how changes in these assumptions can alter equilibrium responses. The goal of this exercise is to highlight the impact of the production function specification on the distribution of learning gains in both our treatments.

Our numerical approach allows us to explore how teachers focus their efforts on students of different learning levels under both types of systems. Following the baseline model described in Barlevy and Neal (2012), we first assume that the productivity of teacher effort (γ) is constant and equal to one, regardless of a student's initial learning level. We then solve the model numerically. Figures B.3a and B.3b show the optimal teacher responses for different levels of student initial learning. Under the Pay for Percentile scheme, the optimal response would result in teachers exerting equal levels of effort with all of their students, regardless of their initial learning level. In contrast, the multiple threshold levels scheme would result in a bell-shaped effort curve, where teachers would focus on students near the threshold and exert minimal effort with students in the tails (see solid line graph in B.3b). Thus, our numerical exercise suggests that if teacher productivity is invariant to the initial level of student learning, then the Pay for Percentile scheme will better serve students at the tails of the distribution.

Figure B.3: Incentive design and optimal effort with constant productivity of teacher effort



(a) Pay for Percentile - γ constant across initial levels of learning. The total effort exerted by teachers is 3.39.

(b) Levels - γ constant across initial levels of learning. The total effort exerted by teachers is 1.55 under the -1 threshold, 1.88 under the 0 threshold, 2.37 under the 1 threshold, and 1.97 under the mutiple threshold.

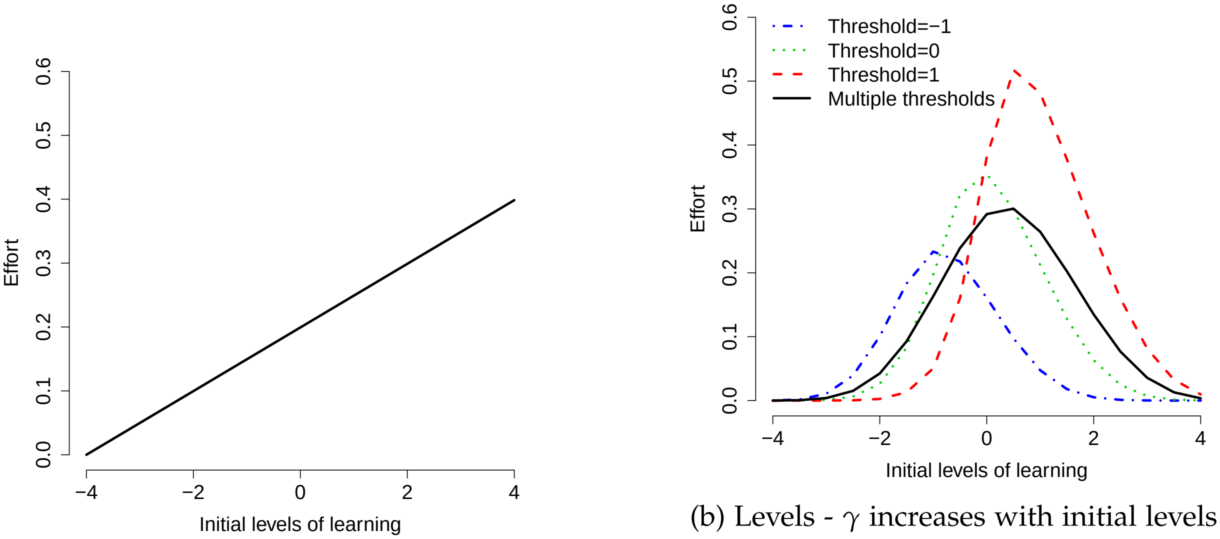
We relax the assumption of constant productivity of teacher effort and allow it to vary with initial learning levels of students. For simplicity, we specify a linear relationship between teacher productivity (γ^l) and student learning levels (a^l) such that $\gamma^l = 1 + 0.25a^l_{(t-1)}$.[38] Figures B.4a and B.4b show the numerical solutions of optimal teacher effort for different initial levels of student learning. In the Pay for Percentile system, focusing on better prepared students increases the likelihood of winning the rank-order contest (among that group of students), while the marginal unit of effort applied to the least prepared students will have a relatively smaller effect on the likelihood of winning the rank-order tournament among that group of students. Thus, in equilibrium, teachers will focus more on better prepared students and will not have an incentive to deviate from this strategy, given the structure and payoffs of the tournament. In contrast, the Levels scheme would yield a similar but slightly skewed bell-shaped curve compared to the baseline constant productivity case.

Our numerical exercise suggests that testing for equality of treatment effects across

[38] Given the uniform distribution of students across initial levels of learning, $\gamma^l = 1 + 0.25a^l_{(t-1)}$ yields the same average cost as assuming γ^l is constant and equal to 1.

the distribution of student baseline test scores in the Pay for Percentile arm allows us to better understand the specification of teacher effort in the education production function.

Figure B.4: Incentive design and optimal effort when the productivity of teacher effort is correlated with the initial level of student learning



(a) Pay for Percentile - γ increases with initial levels of learning. The total effort exerted by teachers is 3.39.

(b) Levels - γ increases with initial levels of learning. The total effort exerted by teachers is 1.12 under the -1 threshold, 1.73 under the 0 threshold, 2.53 under the 1 threshold, and 1.88 under the mutiple threshold.

B.2 Robustness of Simulation Results

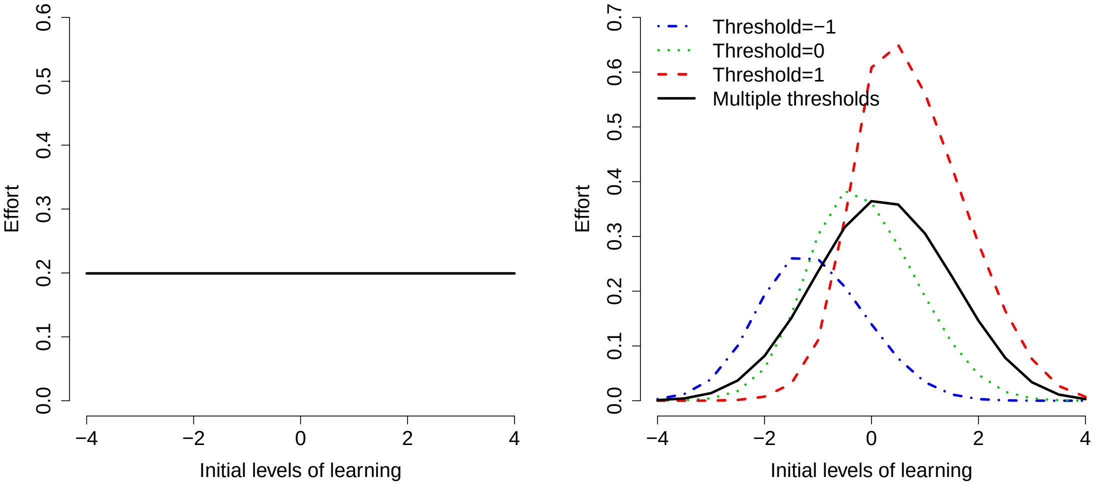
In this section we vary one of the central assumptions in our numerical simulations of the effort exerted by teachers in equilibrium discussed in Section B.1.5. In particular, we change the assumption that students are uniformly distributed across baseline test scores (recall that we had assumed student baseline learning levels to be uniformly distributed from -4 to 4, in 0.5 intervals). Instead, we assume that student baseline learning levels are roughly distributed normally around zero, such that most students are near zero and almost no students are in the tails.[39] Figures B.5 and B.6 show the optimal effort of teachers across both incentive schemes.

As can be seen in the figures below, teacher responses are equal in the pay for percentile scheme (P4Pctile) regardless of the distribution of baseline student learning. This

[39] In reality, we assume a binomial distribution centered around zero.

result is unsurprising given the equilibrium condition in Equation 3. On the other hand, for the proficiency scheme (Levels) the optimal teacher effort changes when the distribution of baseline test scores changes (see Equation 5). However, qualitatively the result is the same as with a uniform distribution of baseline test scores.

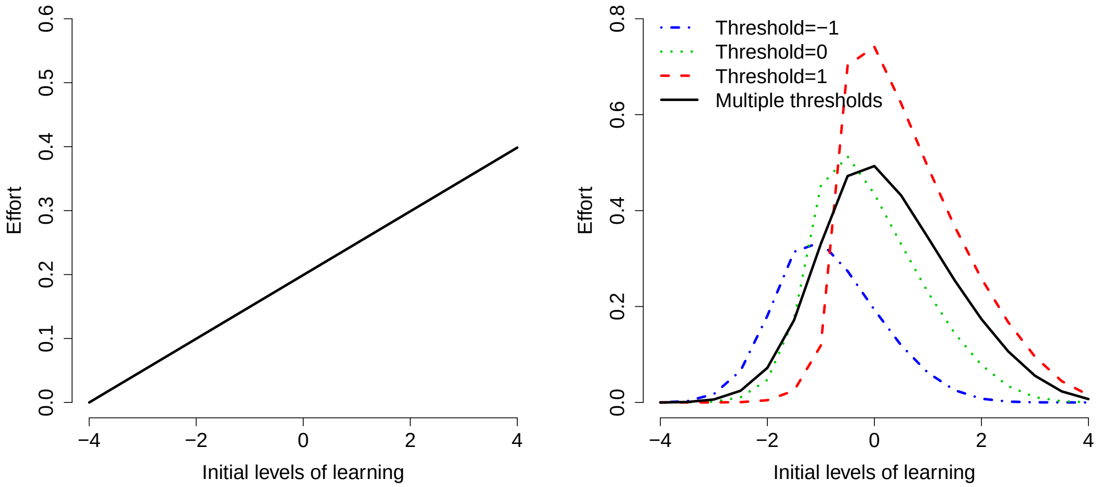
Figure B.5: Incentive design and optimal effort with constant productivity of teacher effort



(a) P4Pctile - γ constant across initial levels of learning

(b) Levels - γ constant across initial levels of learning

Figure B.6: Incentive design and optimal effort when the productivity of teacher effort is correlated with the initial level of student learning



(a) P4Pctile - γ increases with initial levels of learning

(b) Levels - γ increases with initial levels of learning

C Test Design

The tests used in this evaluation were developed by Tanzanian education professionals. The tests were based on the Tanzanian curriculum and followed a similar test development process as the Uwezo annual learning assessment — a nationwide learning assessment used to measure learning in Tanzania.[40] Two types of tests were developed by the test-developers: a non-incentivized (or low-stakes) test that was used for research purposes and an incentivized (or high-stakes) test that was used to by Twaweza to determine teacher bonuses. Both tests followed the testing procedures and protocols established in Mbiti et al. (in press).

C.1 Non-Incentivized test

The non-incentivized (or low-stakes) test was administered on sample of 30 students in each school (10 students each from Grades 1 through 3). To test for spillovers an additional 10 students from Grade 4 were also tested. Sampled students are then followed over the course of the two-year study, except Grade 4 students who were not followed into Grade 5. These non-incentivized tests were only used for research purposes. In order to prevent confusion in schools, these non-incentivized tests were conducted by a separate team to prevent confusion with the intervention team (or the incentivized tests). Given the low levels of learning in Tanzania, we conducted one-on-one tests in which a test enumerator sits with the student and guides her/him through a large font test booklet. This improved data quality and also enabled us to capture a wide range of skills in the event the student was not literate. Students are asked to read and answer the test questions to the administrator who records the number of correctly read or answered test items. For the numeracy questions and the spelling questions students were allowed to use pencil and paper. In order to avoid ceiling and floor effects, we requested the test-developers to include "easy", "medium", and "hard" items.

Since this study was built on the RCT by Mbiti et al. (in press), we used the endline tests that were administered in 2014 for that study as the baseline for this study. The material covered by our tests in Kiswahili and English included reading syllables, reading words, and a reading comprehension. In math, the tests covered simple counting, number recognition, inequalities of number (i.e., which is greater), addition, subtraction, multiplication, and division.

During both endline tests (in 2015 and 2016), we tested students based on the grade

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

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

We standardize test scores using the mean and standard deviation of the control group to compute Z-scores. We also scale the test scores using Item Response Theory (IRT) methods so that all students are on the same scale. The IRT scaling allows us to convert the estimated treatment effects (measured in SDs) to equivalent years of schooling.

C.2 Incentivized test

The incentivized (or high-stakes) tests were used to determine teacher bonuses. These tests were taken by all students in grades 1, 2, and 3. Although there are no bonuses in the control schools, we administer the same type of "incentivized tests" in control schools so that we could compute treatment effects using the incentivized test data. A number of measures were introduced to enhance test security. First, to prevent test-taking by non-target grade candidates, students could only be tested if their name had been listed and their photo taken at baseline. Second, there were ten versions of the tests to prevent copying and leakage; each student was assigned a randomly generated number from a table to identify the test version, with the choice of the number based on day of the week and the first letter of the student's name. Finally, tests were handled, administered, and scored by Twaweza without any teacher involvement. Several checks were done ex-post by Twaweza to ensure there was not any cheating on the high-stakes test.

C.3 Comparability of tests

Both types of tests followed the same test-development framework. As a result, the subject order, question type, and phrasing was similar across both tests. The main difference is the incentivized test is shorter (about 15 mins per student) and uses a variety of stopping rules to reduce testing time. The non-incentivized test took about 40 minutes and covered more skills. It also included more questions to avoid bottom- and top-coding. The specific skills tested are outlined in Table C.1.

Although the content between the two types of test is similar, there are a number of important differences in the administration of the tests. The non-incentivized tests included an "other subject" module to measure potential spillover effects. Non-incentivized tests were administered by taking sampled students out of their classroom during a regular school day. In contrast, the incentivized tests were more "official" as all students in Grades 1-3 were tested on a prearranged test day. On the test day, students in other grades would sometimes be sent home to avoid distractions. Extra-curricular activities were also canceled during the Twaweza test. In addition, most schools used the incentivized test as the end of year test. This also likely encouraged students in the control group to exert effort on the test.

Table C.1: Comparison of low-Stakes and high-Stakes test content

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

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

D Additional Tables

D.1 Properly seeded contests

Table D.1: Effect on test scores (without grade 1)

	(1)	(2)	(3)	(4)	(5)	(6)
	Year 1			Year 2		
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: Non-incentivized						
Levels (α_1)	.061	.04	.058	.11**	.13**	.14***
	(.047)	(.055)	(.051)	(.05)	(.054)	(.05)
P4Pctile (α_2)	.0013	-.051	-.029	.1**	.088*	.11**
	(.045)	(.051)	(.047)	(.045)	(.052)	(.048)
N. of obs.	3,120	3,120	3,120	3,163	3,163	3,163
$\alpha_3 = \alpha_2 - \alpha_1$	-.06	-.091*	-.087*	-.0089	-.039	-.034
p-value ($H_0 : \alpha_3 = 0$)	.18	.084	.065	.87	.46	.51
Panel B: Incentivized						
Levels (β_1)	.13***	.12**	.18***	.17***	.14**	.22***
	(.05)	(.054)	(.068)	(.051)	(.055)	(.069)
P4Pctile (β_2)	.079*	.034	.08	.09**	.063	.11*
	(.045)	(.048)	(.06)	(.045)	(.045)	(.059)
N. of obs.	30,206	30,206	30,206	32,956	32,956	32,956
$\beta_3 = \beta_2 - \beta_1$	-.054	-.09	-.1	-.083*	-.073	-.11
p-value ($H_0 : \beta_3 = 0$)	0.26	0.10	0.11	0.097	0.19	0.11
Panel C: Incentivized – Non-incentivized						
$\beta_1 - \alpha_1$	.06	.07	.11	.055	.0048	.066
p-value($\beta_1 - \alpha_1 = 0$)	.23	.15	.067	.26	.93	.28
$\beta_2 - \alpha_2$	.074	.078	.1	-.01	-.024	.0004
p-value($\beta_2 - \alpha_2 = 0$)	.15	.11	.089	.83	.6	.99
$\beta_3 - \alpha_3$	.014	.0078	-.0057	-.065	-.029	-.066
p-value($\beta_3 - \alpha_3 = 0$)	.79	.88	.92	.19	.64	.31

Results from estimating Equation 1 for different subjects at both follow-ups. Panel A uses data from the non-incentivized test taken by a sample of students. Control variables include student characteristics (age, gender, grade and lag test scores) and school characteristics (PTR, Infrastructure PCA index, a PCA index of how close the school is to different facilities, and an indicator for whether the school is single shift or not). Panel B uses data from the incentivized test taken by all students. Control variables include student characteristics (gender and grade) and school characteristics (PTR, Infrastructure PCA index, a PCA index of how close the school is to different facilities, and an indicator for whether the school is single shift or not). Panel C tests the difference between the treatment estimates in panels A and B. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.2 Results for English

Table D.2: Effect on English test scores

	(1)	(2)
	Year 1	Year 2
	English	English
Panel A: Non-incentivized		
	English	English
Levels (α_1)	.019	.11
	(.087)	(.085)
P4Pctile (α_2)	-.03	.19**
	(.077)	(.081)
N. of obs.	1,532	1,533
$\alpha_3 = \alpha_2 - \alpha_1$	-.048	.078
p-value ($H_0 : \alpha_3 = 0$)	.53	.31
Panel B: Incentivized		
Levels (β_1)	.28***	.28***
	(.066)	(.069)
P4Pctile (β_2)	.16***	.23***
	(.057)	(.055)
N. of obs.	46,018	15,458
$\alpha_3 = \alpha_2 - \alpha_1$	-.12*	-.047
p-value ($H_0 : \alpha_3 = 0$)	.079	.53
Panel C: Incentivized – Non-incentivized		
$\beta_1 - \alpha_1$	.14	.15
p-value($\beta_1 - \alpha_1 = 0$)	.15	.14
$\beta_2 - \alpha_2$	.18	.043
p-value($\beta_2 - \alpha_2 = 0$)	.031	.63
$\beta_3 - \alpha_3$	.043	-.11
p-value($\beta_3 - \alpha_3 = 0$)	.62	.29

Results from estimating Equation 1 for different subjects at both follow-ups. Panel A uses data from the non-incentivized test taken by a sample of students. Control variables include student characteristics (age, gender, grade and lag test scores) and school characteristics (PTR, Infrastructure PCA index, a PCA index of how close the school is to different facilities, and an indicator for whether the school is single shift or not). Panel B uses data from the incentivized test taken by all students. Control variables include student characteristics (gender and grade) and school characteristics (PTR, Infrastructure PCA index, a PCA index of how close the school is to different facilities, and an indicator for whether the school is single shift or not). Panel C tests the difference between the treatment estimates in panels A and B. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.3 Balance in Teacher Turnover

Table D.3: Teacher turnover

	(1)	(2)
	Still teaching incentivized grades/subjects	
	Yr 1	Yr 2
Levels (α_1)	.066	.065
	(.043)	(.04)
P4Pctile (α_2)	.054	.088**
	(.036)	(.034)
N. of obs.	882	882
Mean control	.73	.59
$\alpha_3 = \alpha_2 - \alpha_1$	-.013	.022
p-value ($H_0 : \alpha_3 = 0$)	.75	.56

Proportion of teachers of math, English or Kiswahili in grades 1, 2, and 3 who were teaching at the beginning of 2015 and still teaching those subjects (in the same school) at the end of 2015 (Column 1) and 2016 (Column 2). Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.4 Pass Rates

Table D.4: Pass rates across all skill levels

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	English	Math	Kiswahili	English
Levels (β_1)	.0358**	.0582***	.0359***	.0366***	.0682***	.0149**
	(.015)	(.02)	(.0092)	(.013)	(.016)	(.006)
P4Pctile (β_2)	.0224*	.00739	.0169**	.0331***	.0227	.0132**
	(.012)	(.018)	(.0075)	(.012)	(.017)	(.0056)
N. of obs.	210,358	129,676	129,676	248,250	181,288	30,986
Control mean	.58	.5	.041	.58	.5	.041
$\beta_3 = \beta_2 - \beta_1$	-.013	-.051**	-.019**	-.0035	-.046***	-.0018
p-value ($H_0 : \beta_3 = 0$)	.36	.014	.043	.77	.0051	.8

The independent variable is whether a student acquired a given skills as evidenced by performance on the incentivized test. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.5: Pass rates using levels thresholds in Kiswahili

	Syllables	Words	Sentences	Paragraph	Story	Reading Comprehension
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Year 1						
Levels (β_1)	.064**	.059**	.071***	.075***	.038	.024
	(.026)	(.024)	(.023)	(.022)	(.024)	(.026)
P4Pctile (β_2)	-.0057	.015	.011	.026	-.0099	-.0034
	(.025)	(.022)	(.021)	(.02)	(.021)	(.022)
N. of obs.	17,886	33,440	33,440	15,554	14,678	14,678
Control mean	.4	.59	.5	.37	.52	.56
$\beta_3 = \beta_2 - \beta_1$	-.069***	-.044*	-.06**	-.049**	-.048**	-.027
p-value ($H_0 : \beta_3 = 0$)	.0086	.081	.011	.017	.045	.27
Panel B: Year 2						
Levels (β_1)	.09***	.085***	.08***	.046**	.0032	.053**
	(.021)	(.02)	(.018)	(.019)	(.026)	(.021)
P4Pctile (β_2)	.047**	.036*	.032*	-.0089	-.027	.012
	(.023)	(.02)	(.019)	(.02)	(.022)	(.019)
N. of obs.	26,746	44,262	44,262	17,516	15,493	33,009
Control mean	.3	.6	.48	.43	.61	.56
$\beta_3 = \beta_2 - \beta_1$	-.044**	-.049***	-.048***	-.055***	-.03	-.041*
p-value ($H_0 : \beta_3 = 0$)	.027	.0082	.0058	.0042	.22	.053

The independent variable is whether a student acquired a given skills as evidenced by performance on the incentivized test. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.6: Pass rates using levels thresholds in math

	Counting	Numbers	Inequalities	Addition	Subtraction	Multiplication	Division
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Year 1							
Levels (β_1)	.0034	.014	.03**	.05**	.043**	.038**	.035*
	(.0091)	(.021)	(.014)	(.021)	(.02)	(.017)	(.018)
P4Pctile (β_2)	.031***	.031*	.033***	.018	.016	.023	.0095
	(.0078)	(.018)	(.012)	(.018)	(.016)	(.016)	(.018)
N. of obs.	17,886	17,886	33,440	48,118	48,118	30,232	14,678
Control mean	.93	.64	.74	.59	.5	.23	.22
$\beta_3 = \beta_2 - \beta_1$	.028***	.017	.0027	-.033	-.027	-.015	-.026
p-value ($H_0 : \beta_3 = 0$)	.0012	.4	.85	.12	.16	.37	.17
Panel B: Year 2							
Levels (β_1)	.000686	.0411**	.0265**	.0442**	.0462**	.0514***	.0395**
	(.0078)	(.019)	(.011)	(.019)	(.019)	(.014)	(.017)
P4Pctile (β_2)	.0108	.0595***	.0388***	.0394**	.026	.0254**	.0223
	(.0071)	(.017)	(.01)	(.017)	(.017)	(.013)	(.017)
N. of obs.	26,746	26,746	44,262	59,755	59,755	15,493	15,493
Control mean	.94	.68	.79	.6	.56	.11	.18
$\beta_3 = \beta_2 - \beta_1$	.01	.018	.012	-.0049	-.02	-.026	-.017
p-value ($H_0 : \beta_3 = 0$)	.12	.31	.23	.78	.24	.11	.34

The independent variable is whether a student acquired a given skills as evidenced by performance on the incentivized test. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.7: Pass rates using levels thresholds in English

	Syllables	Words	Sentences	Paragraph	Story	Reading Comprehension
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Year 1						
Levels (β_1)	.095***	.05***	.023***	.015**	.0079*	.013*
	(.021)	(.013)	(.0087)	(.0065)	(.0046)	(.0078)
P4Pctile (β_2)	.036**	.028**	.0041	.0073	.0079*	.019***
	(.016)	(.011)	(.007)	(.0055)	(.0046)	(.0064)
N. of obs.	17,886	33,440	33,440	15,554	14,678	14,678
Control mean	.087	.075	.023	.007	.021	.036
$\beta_3 = \beta_2 - \beta_1$	-.059***	-.022*	-.019**	-.0073	-.00001	.0057
p-value ($H_0 : \beta_3 = 0$)	.0034	.074	.043	.29	1	.44
Panel B: Year 2						
Levels (β_1)					.0074	.022**
					(.0061)	(.0086)
P4Pctile (β_2)					.012*	.02**
					(.0068)	(.0079)
N. of obs.	0	0	0	0	10,735	10,735
Control mean	.	.	.	.	.017	.025
$\beta_3 = \beta_2 - \beta_1$					.0048	-.0016
p-value ($H_0 : \beta_3 = 0$)					.5	.88

The independent variable is whether a student acquired a given skills as evidenced by performance on the incentivized test. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.5 Effects on Test Takers and Lee Bounds on the Incentivized Test

Table D.8: Number of test takers, incentivized test

	(1) Year 1	(2) Year 2
Levels (α_1)	0.02	0.05***
	(0.02)	(0.01)
P4Pctile (α_2)	-0.00	0.03**
	(0.02)	(0.01)
N. of obs.	540	540
Mean control group	0.78	0.83
$\alpha_3 = \alpha_2 - \alpha_1$	-0.02	-0.03**
p-value($\alpha_3 = 0$)	0.20	0.04

The independent variable is the proportion of test takers (number of test takers divided by the enrollment in each grade) of the incentivized exam. The unit of observation is the school-grade level. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.9: Lee bounds for the incentivized test

	(1)	(2)	(3)	(4)
	Year 1		Year 2	
	Math	Kiswahili	Math	Kiswahili
Levels (α_1)	0.11^{**}	0.13^{***}	0.14^{***}	0.18^{***}
	(0.05)	(0.05)	(0.04)	(0.05)
P4Pctile (α_2)	0.07^{*}	0.02	0.09^{**}	0.09^{*}
	(0.04)	(0.04)	(0.04)	(0.05)
N. of obs.	48,077	48,077	59,680	59,680
$\alpha_3 = \alpha_2 - \alpha_1$	-0.047	-0.11^{**}	-0.044	-0.093^{**}
p-value($\alpha_3 = 0$)	0.30	0.026	0.31	0.045
Lower 95% CI (α_1)	0.00066	0.021	-0.023	0.027
Higher 95% CI (α_1)	0.23	0.25	0.32	0.35
Lower 95% CI (α_2)	-0.012	-0.070	0.014	-0.0032
Higher 95% CI (α_2)	0.14	0.10	0.17	0.17
Lower 95% CI (α_3)	-0.16	-0.24	-0.22	-0.27
Higher 95% CI (α_3)	0.063	0.00099	0.11	0.057

The independent variable is the standardized test score for different subjects. For each subject we present Lee (2009) bounds for all the treatment estimates (i.e., trimming the left/right tail of the distribution in Levels and P4Pctile schools so that the proportion of test takers is the same as the number in control schools). Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.6 National Assessments

We test the effect of both interventions on the Primary School Leaving Examination (PSLE) taken by students in grade 7. We retrieved records for all schools in Tanzania from the National Examinations Council of Tanzania (NECTA) website (`https://necta.go.tz/psle_results`) and then merged them with out data using a fuzzy merge based on the school name, region, and district. We were able to match over 80% of schools in our data.

The PSLE is a high-stakes test for students: their progression to secondary school is related to the results of this test. Recent reforms publicized the rankings of schools based on the results of these tests. Overall, we do not find any impact of our treatment on PSLE test scores, pass rates, or the number of test takers (see Table D.10).[41]

[41] We do find that test scores decrease on the SNFA examination in 2015. However, this is not consistent

Table D.10: Effect on national assessments (Grade 7 - PSLE)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Grade 7 PSLE 2015			Grade 7 PSLE 2016			Grade 7 PSLE 2017		
	Pass	Score	Test takers	Pass	Score	Test takers	Pass	Score	Test takers
Levels (α_1)	-0.02	-0.07	6.99	0.00	-0.05	4.02	0.03	0.10	7.00
	(0.04)	(0.08)	(6.99)	(0.03)	(0.07)	(7.56)	(0.03)	(0.06)	(8.76)
P4Pctile (α_2)	-0.04	-0.07	-4.00	-0.02	-0.03	-2.29	-0.00	0.02	0.59
	(0.03)	(0.08)	(6.48)	(0.03)	(0.06)	(5.75)	(0.03)	(0.06)	(7.08)
N. of obs.	11,616	11,616	165	10,031	10,031	155	12,070	12,070	155
N. of schools	167	167	165	158	158	155	158	158	155
Mean control group	0.71	2.98	55.3	0.67	2.83	52.4	0.69	2.86	61.9
$\alpha_3 = \alpha_2 - \alpha_1$	-0.020	-0.0043	-11.0	-0.029	0.016	-6.32	-0.032	-0.074	-6.41
p-value ($H_0 : \alpha_3 = 0$)	0.63	0.96	0.10	0.42	0.84	0.39	0.30	0.23	0.47

Standard errors, clustered at the school level, are in parentheses.

with our higher-quality data on grade 4 students (see Table 5). We find an increase in test takers in 2016 (insignificant) and 2017 (significant) in the Levels treatment, which could be viewed as a positive effect of the treatment. Results available upon request.

D.7 Classroom observations

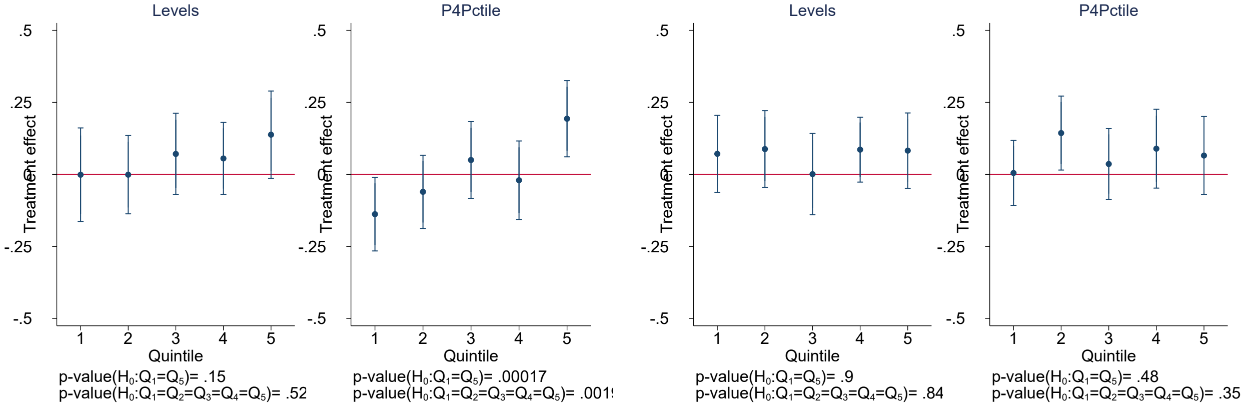
Table D.11: Classroom observations

	(1) Classroom Environment	(2) Teaching	(3) Sleeping
Levels (α_1)	-0.030	0.077	0.0013
	(0.14)	(0.14)	(0.044)
P4Pctile (α_2)	0.12	-0.064	-0.041
	(0.12)	(0.14)	(0.034)
N. of obs.	2,080	1,481	772
$\alpha_3 = \alpha_2 - \alpha_1$	.005	-.012	.13
p-value($\alpha_3 = 0$)	.25	.36	.27

The outcome here are index created taking the first component from a PCA analysis of different items measured during classroom observations. The outcome in Column 1 is an index that measures whether the classroom 's environment is conductive to learning. It is composed of the following measures: whether student's work is display on the walls, whether there are charts on the walls, and the number of charts in the wall. The outcome in Column 2 is an index that measures teacher's behavior during class time. It is composed of the following measures: whether the teacher threatens students, and whether the teacher hits students. Finally, the outcome in Column 3 shows whether any students were sleeping during class time. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.8 Additional Heterogeneity in Treatment Effects

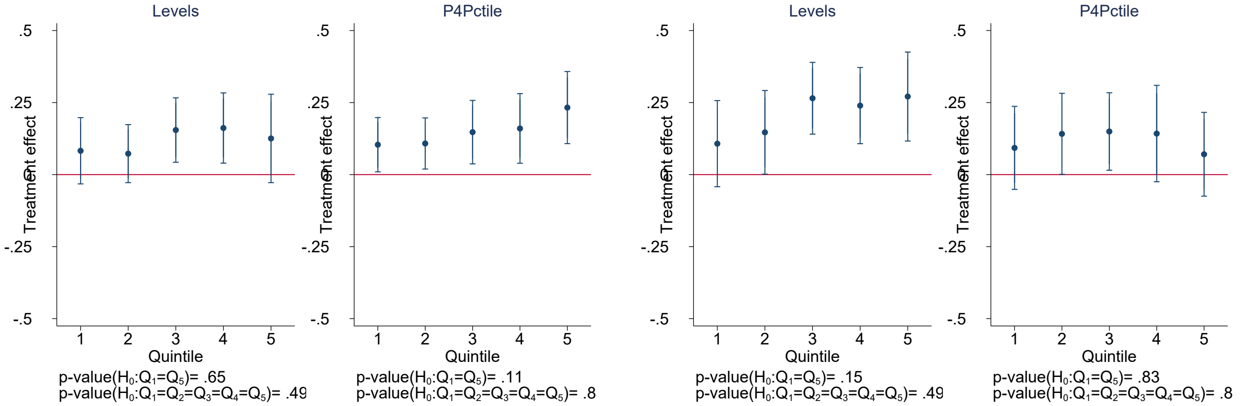
Figure D.1: Math — non-incentivized



(a) Year 1

(b) Year 2

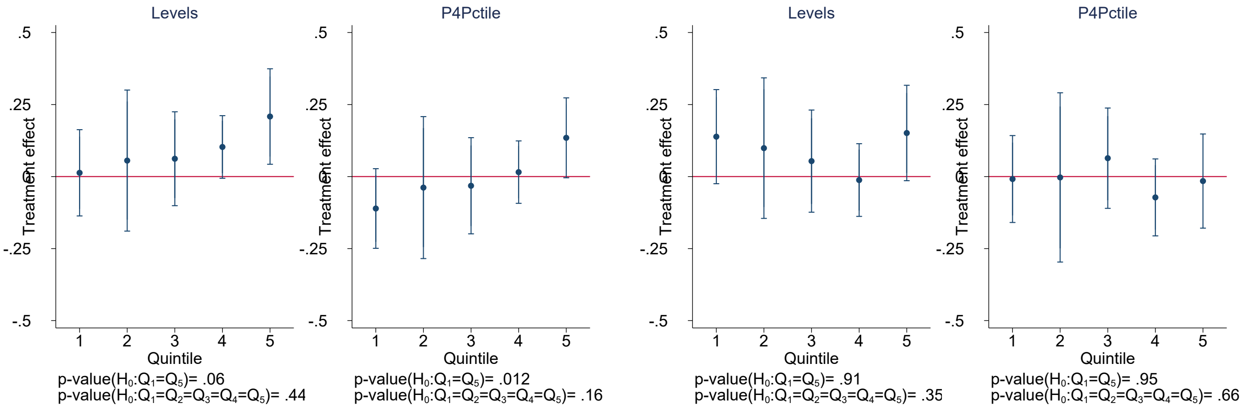
Figure D.2: Math — incentivized



(a) Year 1

(b) Year 2

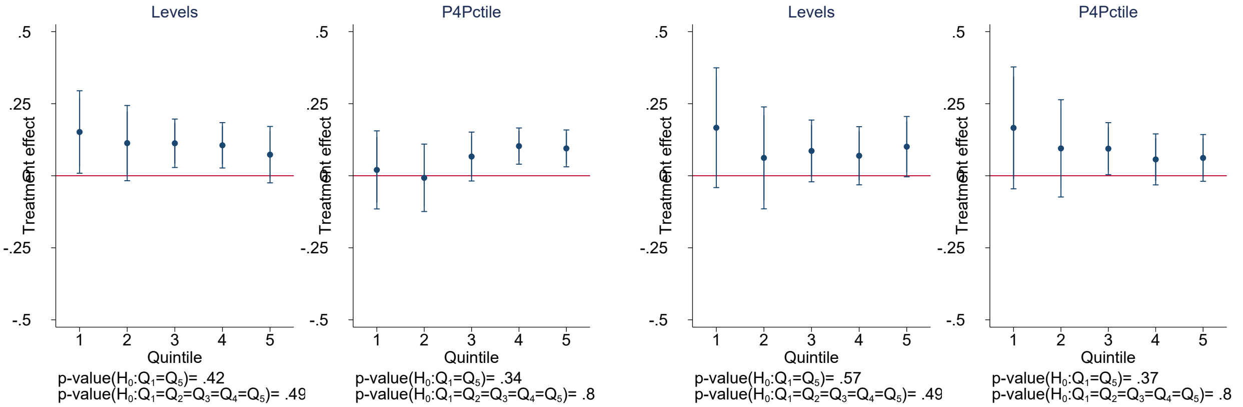
Figure D.3: Kiswahili — non-incentivized



(a) Year 1

(b) Year 2

Figure D.4: Kiswahili — incentivized



(a) Year 1

(b) Year 2

Table D.12: Heterogeneity by student characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
		Math			Swahili	
	Male	Age	Test(Yr0)	Male	Age	Test(Yr0)
Levels*Covariate (α_2)	-0.022	0.014	0.034	0.017	-0.031*	0.015
	(0.037)	(0.014)	(0.032)	(0.051)	(0.018)	(0.029)
P4Pctile*Covariate (α_1)	0.020	0.0076	0.068***	-0.024	0.0066	0.030
	(0.041)	(0.015)	(0.026)	(0.051)	(0.019)	(0.030)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650
$\alpha_3 = \alpha_2 - \alpha_1$	.042	-.006	.035	-.041	.038*	.015
p-value ($H_0 : \alpha_3 = 0$)	.3	.69	.23	.42	.05	.62

Each column interacts the treatment effect with different student characteristics: sex (columns 1, 4, and 7), age (columns 2, 5, and 8), and baseline test scores (columns 3, 6, and 9). Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.13: Heterogeneity by school characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
		Math			Swahili	
	Facilities	PTR	Fraction Weak	Facilities	PTR	Fraction Weak
Levels*Covariate (α_2)	0.035	-0.00031	-0.22	-0.023	-0.0010	-0.13
	(0.022)	(0.0015)	(0.17)	(0.026)	(0.0014)	(0.17)
P4Pctile*Covariate (α_1)	-0.022	-0.0026**	-0.24	-0.028	-0.0017	-0.28*
	(0.026)	(0.0011)	(0.15)	(0.030)	(0.0014)	(0.17)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650
$\alpha_3 = \alpha_2 - \alpha_1$	-.057**	-.0023	-.025	-.0048	-.00069	-.16
p-value ($H_0 : \alpha_3 = 0$)	.018	.18	.87	.87	.7	.37

Each column interacts the treatment effect with different school characteristics: a facilities index (columns 1, 4, and 7), the pupil-teacher ratio (columns 2, 5, and 8), and the fraction of students that are below the median student in the country (columns 3, 6, and 9). Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.9 Teacher Understanding

Since there is no comparable test for control group teachers, we cannot interact the treatment variable with teacher understanding. Instead, we split each treatment group into a high (above average) understanding group and a low (below average) understanding group, and estimate the treatment effects for these sub-treatment groups relative to the entire control group (i.e., the control group is the omitted category). Within each treatment arm, we test for differences between the high-understanding and low-understanding groups to determine if better understanding leads to better student test scores. As some teachers were not present when we conducted the teacher comprehension tests, we created an additional group for teachers with no test in both treatments.

Table D.14: Heterogeneity by teacher's understanding

	(1)	(2)	(3)
	Math	Swahili	English
Levels (high-understanding)	0.032	0.075*	0.052
	(0.044)	(0.042)	(0.060)
Levels (low-understanding)	0.073*	0.083**	0.074
	(0.042)	(0.037)	(0.049)
P4Pctile (high-understanding)	0.0093	0.029	0.12**
	(0.035)	(0.036)	(0.051)
P4Pctile (low-understanding)	0.052	-0.0059	0.032
	(0.043)	(0.041)	(0.052)
N. of obs.	9,650	9,650	6,314
Levels:High-Low	-.042	-.0073	-.022
p-value (Levels:High-Low=0)	.28	.84	.73
P4Pctile:High-Low	-.042	.035	.089
p-value (P4Pctile:High-Low=0)	.31	.41	.15
P4Pctile:High-Levels:High	-.022	-.047	.069
p-value (P4Pctile:High-Levels:High=0)	.63	.28	.3
P4Pctile:Low-Levels:Low	-.022	-.088	-.042
p-value (P4Pctile:Low-Levels:Low=0)	.67	.058	.5

The outcome variables are student test scores in math (Column 1), Kiswahili (Column 2), and English (Column 3). Each regression pools the data for both follow-ups. Teachers are classified as above or below the median in each follow-up in treatment schools. Since we do not have "understanding" questions for teachers in control schools, all teachers in the control group are compared for teachers above and below the median in treatment schools. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$