The Challenge of Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania*

Isaac Mbiti† Mauricio Romero‡ Youdi Schipper§

August 1, 2018

Abstract

Teacher performance pay systems can improve student learning outcomes, particularly in settings where existing mechanisms for teacher accountability are weak. We use a field experiment in Tanzanian public primary schools to directly compare the effectiveness on early grade learning of two different teacher performance pay systems: a pay for percentile system and a system that rewards teachers based on student proficiency levels. Pay for percentile is more complex but can (under certain conditions) induce optimal effort among teachers. However, a threshold system is easier to implement and it may be easier for teachers to figure out how to react optimally to the incentive scheme. Both systems improve student test scores. However, despite the potential to induce optimal effort of the pay for percentile system, the proficiency system is at least as effective in boosting student learning. Moreover, we find suggestive evidence that the pay for percentile system favors students from the top of the performance distribution, highlighting the challenge of designing incentives that can deliver optimal and equitable learning gains for all students.

JEL Classification: C93, H52, I21, M52, O15
Keywords: teacher performance pay, pay for percentile, incentive design, Tanzania

*We are especially grateful to Karthik Muralidharan for his collaboration in the early stages of this project and subsequent discussions. We would also like to thank John Friedman, Delia Furtado, Guthrie Gray-Lobe, Ronak Jain, Joseph Mmbando, Molly Lipscomb Johnson, Michael Kremer, Ömer Özak, Bobby Pakzad-Hurson, Jay Shimshack, Bryce Millet Steinberg, Tavneet Suri and seminar participants at UC San Diego, Universidad del Rosario, and the RISE conference for their comments. Erin Litzow and Jessica Mahoney provided excellent research assistance through Innovations for Poverty Action. We are also grateful to EDI Tanzania for their thorough data collection and implementation efforts. The EDI team included Respichius Mitti, Phil Itanisia, Timo Kyessey, Julius Josephat, Nate Sivewright and Celine Guimas. The evaluation was partially funded by a grant from the REACH Trust Fund at the World Bank. We are grateful to Peter Holland, Jessica Lee, Arun Joshi, Owen Ozier, Salman Asim and Cornelia Jesse for their support through the REACH initiative.

†University of Virginia, J-PAL, IZA; imbiti@virginia.edu
‡University of California - San Diego; mtromero@ucsd.edu
§Twaweza; yschipper@twaweza.org

1 Introduction

Over the past two decades, global education priorities have shifted from increasing primary school enrollment to promoting policies that improve learning. This shift has been driven in part by the evidence revealing poor and stagnant levels of learning among students in developing countries, despite significant investments in education (World Bank, 2018b). Given the central role of teachers in the education production function (Hanushek & Rivkin, 2012; Chetty, Friedman, & Rockoff, 2014b, 2014a), as well as the substantial share of the education budget devoted to their compensation, there is increasing interest in policies that can improve teacher effectivness (Bruns, Filmer, & Patrinos, 2011). By strengthening the links between teacher remuneration and learning, teacher performance pay programs are seen as a promising pathway to improve education quality. Consequently, the adoption rate of such programs has increased significantly over the past two decades across the OECD, as well as in less developed countries such as Brazil, Chile, and Pakistan (Alger, 2014; Ferraz & Bruns, 2012; Barrera-Osorio & Raju, in press; Contreras & Rau, 2012).[1]

Previous studies have shown that the effectiveness of teacher performance pay systems depends on key design elements such as the mapping between the performance metric and the reward (Neal & Schanzenbach, 2010; Neal, 2011; Mbiti et al., 2017; Muralidharan & Sundararaman, 2011; Bruns & Luque, 2015; Imberman & Lovenheim, 2015; Loyalka, Sylvia, Liu, Chu, & Shi, in press).[2] Yet despite the well-documented shortcomings of incentive schemes based on proficiency thresholds, such designs are commonly used (for example, in many US states under No Child Left Behind), as they are clearer and easier to comprehend and implement relative to schemes based on more complex alternatives, such as value-added measures. The administrative simplicity of such designs may be particularly appealing for developing countries with weaker state capacity. However, proficiency-based incentive systems may encourage teachers to focus on students who are close to the proficiency threshold (Neal & Schanzenbach, 2010; Neal, 2011). In contrast, more complex incentive systems, such as those based on rank-order tournaments (e.g., pay for percentile designs), may induce greater and potentially socially optimal levels of effort among teachers compared to simpler schemes (Barlevy & Neal, 2012; Neal,

[1] For instance, the share of US school districts with teacher performance pay programs has increased by more than 40% from 2004 to 2012 (Imberman, 2015). Globally, 13 of 34 countries that provided information about their teacher policies under the World Bank's Systems Approach for Better Education Results (SABER) initiative provided monetary bonuses to high performing teachers (World Bank, 2018a).

[2] Other important design elements include whose performance is incentivized (individual or group); the incentive metric (test scores, or input measures like attendance or preparation); the size of the expected incentive payment; and the simultaneous implemtation of complementary policies.

2011; Loyalka et al., in press). However, such systems are harder to implement and may be difficult for teachers to fully comprehend, which can undermine their effectiveness if teachers cannot figure out how to react optimally to the incentive scheme (Goodman & Turner, 2013; Fryer, 2013; Neal, 2011).[3]

We compare the effectiveness of the rank-order, tournament-style pay for percentile scheme proposed by Barlevy and Neal (2012) to a simpler design with multiple proficiency thresholds ("levels") using a randomized experiment in a set of 180 Tanzanian public primary schools. Both types of incentive programs rewarded teachers for performance of their students in math, Kiswahili, and English in first, second, and third grade. In addition, the per-student bonus budget was equalized (ex-ante) across grades, subjects, and treatment arms. In both incentive designs, we determined individual teacher reward payments based on actual student performance on externally administered tests. The mean teacher bonus paid in the second year of the evaluation was 3.5% of the annual net salary (roughly half a month's pay).[4] Following Neal (2013), we evaluate the incentive programs using data from both the "incentivized" test administered to all students to determine teacher bonuses and a "non-incentivized" test administered to a sample of students for research purposes. Both types of tests were collected in control schools, although the results of the incentivized test did not trigger any payments in these schools.

In the 60 schools assigned to the pay for percentile arm, students were first tested and then placed in one of several groups based on their initial level of learning. At the end of the school year, students were re-tested and ranked within their assigned group. Teachers were rewarded in proportion to their students' rankings within each group. By handicapping the differences in initial student performance across teachers, the pay for percentile system does not penalize teachers who serve disadvantaged students. In addition, since the reward schedule (or the mapping of student rankings to teacher bonuses) is exactly the same across all groups, it could encourage teachers to focus on all students, rather than solely on those who are marginal or close to a proficiency threshold. Barlevy and Neal (2012) show that pay for percentile can induce socially optimal levels of effort among teachers. However, the system can be difficult for teachers to comprehend, which can undermine its effectiveness Charness and Kuhn (2011). In addition, it is challenging to implement as school systems need to create and manage databases that

[3] In their review article Charness and Kuhn (2011) argue that rank-order tournaments may be less robust as incentive schemes because they require participants to "think strategically about their co-workers' efforts to find a Nash equilibrium" (p. 257).

[4] Similar incentive sizes were used in (Fryer, 2013; Muralidharan & Sundararaman, 2011; Lavy, 2002; Ladd, 1999; Vigdor, 2008). See Leigh (2012) for additional details.

track student learning over time.[5]

In the 60 schools assigned to receive incentives based on proficiency targets, teachers earned bonuses based on their students' mastery of several grade-specific skills outlined in the national curriculum. Because incentive programs that use single-proficiency thresholds encourage teachers to focus on students close to the passing threshold, we included several thresholds. The skill thresholds ranged from very basic to more complex skills, allowing teachers to earn rewards from a wide range of students. As reward payments for each skill were inversely proportional to the number of students that attain the skill, harder-to-pass skills are rewarded more. Given the clarity in the reward structure is easier for teacher to understand how their teaching practice is related to the incentivized metric. In addition, this system is simpler to implement: It only requires that school systems administer one test at the end of the year (and there is no need to link student test score data across time). However, since the system uses proficiency thresholds, it is arguably less likely to induce optimal effort among teachers. Moreover, as rewards are based on absolute learning levels, systems using proficiency targets may disadvantage teachers who serve students from poorer backgrounds. Finally, the effectiveness of the proficiency system depends on the thresholds, while a rank-order tournament does not have any "tuning" parameters.

Despite the theoretical advantages of the pay for percentile system, we find that the simpler proficiency levels incentive system is as effective, and sometimes more effective, than the pay for percentile system in this setting. After two years, using data from the sample of students who took the non-incentivized test, test scores in math increased by about 0.07σ under both systems (statistically significant at the 10 percent level). Kiswahili scores increased by $.11\sigma$ (p-value < 0.01) under the proficiency levels system compared to a $.058\sigma$ (p-value .1) increase under the pay for percentile system. Test scores in English increased by $.11\sigma$ (p-value .19) in proficiency levels schools and $.19\sigma$ (p-value .019) in pay for percentile schools. However, the treatment estimates are not statistically distinguishable from each other. The results using the incentivized test data show similar patterns, although the point estimates are generally larger, which is likely due to increased student effort (Levitt, List, Neckermann, & Sadoff, 2016; Gneezy et al., 2017; Mbiti et al., 2017). At the end of two years, math test scores in the incentivized test increased by $.14\sigma$ (p-value < 0.01) under the proficiency levels system compared to a $.093\sigma$ (p-value .02) increase

[5]In the long run, one additional advantage of a pay for percentile system is that it can mitigate concerns about teaching to the test since it relies on rank order measures. This property allows an education system to use different tests and test formats over time, since the design does not require the use of psychometrically equivalent tests. An examination of this long-run property is outside the scope of our paper.

under the pay for percentile system. Kiswahili scores increased by .18σ (p-value < 0.01) under the proficiency levels system compared to a .085σ (p-value .061) increase under the pay for percentile system. English test scores increased by .28σ (p-value < 0.01) in proficiency levels schools and .23σ (p-value < 0.01) in pay for percentile schools. While the treatment estimates for math and English test scores are statistically indistinguishable across treatments, the levels treatment estimate for Kiswahili is statistically larger than the pay for percentile estimate at the 5% level.

These results contrast with the findings of Loyalka et al. (in press), who find that student math test scores increased the most under a pay for percentile system compared to other systems (using the incentivized test data).[6] In addition, while Loyalka et al. (in press) find that pay for percentile teacher incentives led to equitable learning gains across the distribution of students, we find suggestive evidence that better prepared students benefited more under the pay for percentile treatment. In contrast, under the levels system learning gains were more equitably distributed, except for English in the second year of the evaluation. Formal statistical tests reject the equality of treatment effects across the distribution of student baseline test scores for pay for percentile schools in both years for Kiswahili, and in the first year for math.

In order to interpret our findings, we build a stylized model where we compare the distributional effects of our incentive systems on student learning under two different specifications of the education production function: a specification where all students benefit equally from teacher effort, and one in which students with higher baseline test scores benefit more from teacher effort.[7] Following the predictions of our model, our empirical results suggest that the productivity of teacher effort is higher for students with better baseline test scores in the education production function.[8] Even though Barlevy and Neal (2012) show that pay for percentile systems can still induce socially optimal levels of effort in the presence of this type of heterogeneity (or production function specification), teacher performance pay programs need to be carefully designed to account for equity concerns.[9]

[6] Loyalka et al. (in press) focus on sixth-grade math teachers in rural China. Math teachers in China are typically subject specialists and there is some evidence, based on small samples, that they understand mathematical concepts better than US teachers (Ma, 1999). In contrast, data from a recent World Bank study showed that only about 27 percent of Tanzania fourth-grade teachers could correctly interpret data on a graph and only half could correctly compare fractions (World Bank, 2014).

[7] The basic model examined by Barlevy and Neal (2012) also assumed that all students benefit equally from teacher effort.

[8] This is consistent with Banerjee and Duflo (2011) and Glewwe, Kremer, and Moulin (2009) who argue that education systems in developing countries tend to favor students from the top quantiles.

[9] The effect of incentives on inequality has also been explored in the context of a firm. Bandiera, Barankay, and Rasul (2007) find that performance pay for managers increased earnings inequality among

We use our comprehensive set of survey data collected from school administrators, teachers, and students to shed light on theoretically relevant mechanisms. Given the well-documented concerns about teacher misunderstanding of incentive design (Goodman & Turner, 2013; Fryer, 2013), we show that teacher comprehension was high under both systems, which allows us to rule out a relative lack of understanding as a major driving factor.[10] However, it is still possible that even if teachers understand how payments are made, they are unable to figure out (ex-ante) how their effort will be rewarded. For example, We find that teachers in the pay for percentile design expected to earn a bonus 18 percent lower, on average, than their levels counterparts. This reduction in expectations could be driven by the increased uncertainty and complexity of the pay for percentile system relative to the levels system and could dampen teachers' responsiveness to the incentives. In addition, teachers in the levels system were better able to articulate clear and specific targets for their students on the incentivized tests, perhaps due to the clearer reward structure. Overall, the results highlight the importance of clarity in the design for improving teacher response to the incentive.

Our study contributes to a growing literature on the potential of teacher incentives to improve learning outcomes in developing countries (Duflo, Hanna, & Ryan, 2012; Glewwe, Ilias, & Kremer, 2010; Muralidharan & Sundararaman, 2011; Mbiti et al., 2017; Loyalka et al., in press). By comparing the pay for percentile system to a simpler, cost-equivalent proficiency system, we provide some of the first empirical evidence of the effectiveness of pay for percentile in sub-Saharan Africa, benchmarked against a simpler system.[11] This comparison informs the trade-offs faced by education authorities who have to consider the effectiveness, feasibility of implementation, and equity of different incentive designs with limited information about the education production function. Given resource and accountability constraints and low levels of teacher knowledge (relative to global scales) in sub-Saharan African countries, evidence that can provide policy makers with insights on the relative cost-effectiveness of programs aimed at improving education quality is especially important (World Bank, 2014, 2018b).

Furthermore, because designing effective incentive systems presents a significant chal-

workers because managers focused their efforts on high ability workers.

[10]The treatment estimates are generally higher in the second year of the program (compared to the first year) across both treatment arms. While this may reflect teachers learning how to optimally react to the incentive scheme, it may also reflect teachers reacting different after they see we deliver on our promise to pay for performance in the first year.

[11]Gilligan, Karachiwalla, Kasirye, Lucas, and Neal (2018) evaluate a pay for percentile teacher incentive program in a set of rural schools in Uganda. They find that pay for percentile incentives have no impact on student learning, except in schools with textbooks. They do not compare their pay for percentile system to other incentive designs.

lenge, empirical evidence that can shed light on the potential trade-offs between complex and simpler teacher incentive systems is especially useful for policy makers from countries with weak state capacity.

As our results show, under certain education production functions, complex and theoretically optimal designs can favor better students. However, simpler systems, though potentially less optimal, may be more robust from an equity perspective to different production functions. This is consistent with a large body of literature on contract theory, such as Carroll (2015) and Carroll and Meng (2016), showing that simpler incentive schemes are often more robust mechanisms for resolving principal agent problems when there is uncertainty about the specification of the production function.[12]

2 Experimental Design

2.1 Context

Tanzania allocates about one-fifth of overall government spending (roughly 3.5 percent of GDP) to education (World Bank, 2017). Much of this spending has been devoted to promoting educational access. As a consequence, net enrollment rates in primary school increased from 53 percent in 2000 to 80 percent in 2014 (World Bank, 2017). Despite these gains in educational access, educational quality remains a major concern. Resources and materials are scarce. For example, in 2017 only 14 percent of schools had access to electricity and just over 40 percent had access to potable water (World Bank, 2017). Nation wide, there are approximately 43 pupils per teacher (World Bank, 2017), although early grades will often have much larger class sizes. In 2013, approximately five pupils shared a single mathematics textbook, while 2.5 pupils shared a reading textbook (World Bank, 2017). Student learning levels are quite low. In 2012, data from nationwide assessments showed that only 38 percent of children aged 9-13 are able to read and do arithmetic at Grade 2 level, suggesting that educational quality is a pressing policy problem (Uwezo, 2013).

The poor quality of education is driven in part by limited accountability in the education system. Quality assurance systems (e.g., school inspectors) typically focus on superficial issues such as the state of the school garden, rather than on issues that may affect learning (Mbiti, 2016). The lack of accountability is further reflected in teacher ab-

[12]Much of the contract theory literature focuses on the robustness of linear incentive schemes. Although our levels design is not linear, our results are in line with the general spirit of this literature which highlights the robustness of simpler schemes

sence rates. Data from unannounced spot checks show that almost a quarter of teachers were absent from school, and only half of the teachers who were at school were in the classroom during further spot checks (World Bank, 2011). As a result, almost 60 percent of planned instructional time is lost (World Bank, 2011).

Despite these high absence rates, teachers' unions continue to lobby for better pay as a way to address quality concerns in the education system, even though studies have found that the correlation between teacher compensation and student learning is extremely low (Kane, Rockoff, & Staiger, 2008; Bettinger & Long, 2010; Woessmann, 2011). Teachers earn approximately 500,000 TZS per month ($\sim$ US$300) — roughly 4.5 times GDP per capita (World Bank, 2017).[13] In addition, approximately 60 percent of the education budget is devoted to teacher compensation. Despite the relatively lucrative wages of Tanzanian teachers, the teachers' union called a strike in 2012 to demand a 100 percent increase in pay (Reuters, 2012; PRI, 2013).[14]

2.2 Interventions and Implementation

We compare the effectiveness of the pay for percentile scheme proposed by Barlevy and Neal (2012) to a simple proficiency threshold design, where the budgets are equalized to facilitate cost-effectiveness comparisons. The interventions were formulated and managed by Twaweza, an East African civil society organization that focuses on citizen agency and public service delivery, and were part of a series of projects launched under a broader program umbrella known as KiuFunza ('Thirst for learning' in Kiswahili).[15] A budget of $150,000 per year for teacher and head teacher incentives was split between two treatment arms in proportion to the number of students enrolled. As a result, the total reward in each treatment arm was approximately $3 per student. All interventions were implemented by Twaweza in partnership with EDI (a Tanzanian research firm) and a set of local district partners. In addition, head teachers were offered a bonus of 20 percent of the combined bonus of all incentivized teachers in his or her school. [16]

Within each intervention arm, Twaweza distributed information describing the program to schools and their respective communities via public meetings in early 2015 and

[13]The average teacher in a sub-Saharan African country earns almost four times GDP per capita, compared to OECD teachers who earn 1.3 times GDP per capita (OECD, 2017; World Bank, 2017).

[14]In recent years, other teacher strikes to demand pay increases have occurred in South Africa, Kenya, Guinea, Malawi, Swaziland, Uganda, Benin and Ghana.

[15]The first set of interventions under this program were launched in 2013 and evaluated by Mbiti et al. (2017).

[16]Twaweza included head teachers in the incentive design to ensure that they would be stakeholders in improving learning outcomes. It is also likely that any scaled-up teacher incentive program would feature bonuses for head teachers, as implemented in the KiuFunza project.

2016. Since the first public meeting in 2015, Twaweza informed teachers the program would last two years. The implementation teams also conducted mid-year school visits to re-familiarize teachers with the program, gauge teacher understanding of the bonus payment mechanisms, and answer any remaining questions. At the end of the school year, all students in grades 1, 2, and 3 in every school, including control schools, were tested in Kiswahili, English, and, math. Because this test was used to determine teacher incentive payments, it was considered "high-stakes" (from the teachers' perspective). Our non-incentivized research test was conducted on a different day around the same time. Both sets of tests were based on the Tanzanian curriculum and were developed by Tanzanian education professionals using the Uwezo learning assessment test development framework.[17]

2.2.1 Proficiency thresholds (levels) design

Proficiency based systems are easier for teachers to understand and provide more actionable targets than rank-order or value-added tournaments. Consequently, such systems are likely to increase motivation among teachers and head teachers; however, they have well-known limitations. For example, they are unable to adequately account for differences in the initial distribution of student preparation across schools and classrooms. Moreover, this type of system can encourage teachers to focus on students close to the proficiency threshold, at the expense of students who are sufficiently above or below the threshold (Neal & Schanzenbach, 2010). To mitigate this concern, our levels design features multiple thresholds ranging from very basic skills to more advanced skills in the curriculum. This design allows teachers to earn bonuses for helping a broader set of students, including students with lower and higher baseline test scores.

The levels treatment pays teachers in proportion to the number of skills students in grades 1-3 are able to master in mathematics, Kiswahili, and English. The bonus budget for each subject is split evenly among skills, while the per pass bonus paid ex-post equals the skill budget divided by the number of students passing the skill. Consequently, harder-to-pass skills have a higher per pass bonus. The total bonus for a teacher consists of the per skill rewards aggregated over all skills and all students who pass a particular skill. Teachers can earn larger bonuses if they have more students and if their students are proficient in a larger number of skills, especially harder-to-pass skills.[18]

[17] Uwezo learning assessments have been routinely conducted in Kenya, Tanzania, and Uganda since 2010.

[18] Enrollment at each school is on average 1.6% of total enrollment across schools in the levels system. Hence, we can rule out teachers strategically choosing how many students to push over a threshold to maximize earnings because the total number of students passing the threshold has a negligible effect on

Table 1 shows the skills (i.e., the thresholds) tested in each grade-subject combination. The total budget is split across grades in proportion to the number of students enrolled in each grade. The budget is then divided equally among subjects and skills within each subject. At the end of the year teachers are paid according to the following formula:

$$P_j^s = \frac{X_s}{\sum_{i \in N_L} 1_{a_i > T_s}} \sum_{k \in J} 1_{a_k > T_s} \tag{1}$$

where P_j^s is the payment to teacher j for skill s, J is the set of students of teacher j, a_k is the test score of student k, T_s is the passing threshold for skill s (i.e., the level at which we consider the student masters ability s), X_s is the total amount of money available for skill s, and N_L is the set of all students in schools across Tanzania in the "levels" treatment.

Table 1: Skills tested in the levels design

Kiswahili	**English**	**Math**
	Grade 1	
Letters	Letters	Counting
Words	Words	Numbers
Sentences	Sentences	Inequalities
		Addition
		Subtraction
	Grade 2	
Words	Words	Inequalities
Sentences	Sentences	Addition
Paragraphs	Paragraphs	Subtraction
		Multiplication
	Grade 3	
		Addition
Story	Story	Subtraction
Comprehension	Comprehension	Multiplication
		Division

2.2.2 Pay for percentile design

The pay for percentile design is based on the work of Barlevy and Neal (2012), who show that this incentive structure can, under certain conditions, induce teachers to exert

the overall pass rate across schools.

socially optimal levels of effort. For each subject-grade combination we created student groups with similar initial learning levels based on test score data from the previous school year. Students without test scores in second and third grade were grouped together in an "unknown" ability group.[19] Since none of the first grade students had incoming test scores, we created broad groups based on the historical average test scores for the school. Thus, all first-grade students within a school were assigned to the same group. We then compensated teachers proportionally to the rank of their students at the end of the school year relative to all other students (across all schools in the pay for percentile scheme) with a similar baseline level of knowledge.[20]

More formally, let a_i^{t-1} be the score of student i at the end of the previous school year. Students are divided into k groups according to a_i^{t-1}. We divided the total pot of money allocated to a subject-grade combination X^g into k groups, in proportion to the number of students in the group. That is, $X_k^g = \frac{X^g * n_k}{N_g}$, where N_g is the total number of students in grade g, n_k is the number of students in group k, and X_k^g is the amount of money allocated to group k in grade g. At the end of the year, we ranked students (in 100 rankings) within each group according to their endline test score a_i^t, and within each group we assigned teachers points proportional to the rank of their students. A teacher received 99 points for a student in the top 1% of group k, and zero points for a student in the bottom 1% of the group. Thus, within each group we have:

$$X_k^g = \frac{X^g * n_k}{N} = \sum_{i=1}^{100} b(i-1) * \frac{n_k}{100}$$

where $b(i-1)$ is the amount of money paid for each student in rank i. Therefore, $b = \frac{X^g}{N_g}\frac{2}{99}$. The total money X^g allocated to a subject-grade was proportional to the number of students in each grade and was divided equally among the three subjects. In other words, $X^g = \frac{X^T * N_g}{3\sum_{g=1}^{3} N_g}$, where X^T is the total amount of money available for the pay for percentile design. The total amount of money paid per rank is the same across all groups, in all subjects, and in all grades, and is equal to $b = \frac{X^T}{3\sum_{g=1}^{3} N_g}\frac{2}{99}$. For example, in the first year in pay for percentile schools the total bonus amount was US$70,820 and total enrollment was 22,296. Teachers earned US$1.77 for each student in the top 1% and US$0.89 for each student in the top 50%.

Although this design can deliver socially optimal levels of effort, it may be challenging

[19]Roughly 20% of students are grouped into the "unknown" ability group. This includes newly enrolled students, as well as students who were enrolled but for some reason do not have test score data.

[20]Our results are robust to excluding grade 1 students from the sample. See Table A.2 in Appendix A.2.1.

to implement at scale, particularly in settings with weak administrative capacity such as Tanzania. For instance, maintaining child-level panel databases is a non-trivial administrative challenge. Moreover, the pay for percentile system may prove difficult to grasp for the individual teacher. It presents each teacher with a series of tournaments and therefore the bonus payoff is relatively hard to predict, even if the design guarantees a fair system. The uncertainty introduced by being pitched against students from schools across the whole country may dilute the incentive.

2.3 A Note on English Language Teaching

As Kiswahili is the official language of instruction in Tanzania, English is taught as a second language in primary schools. However, English is rarely spoken outside of the classroom, so English language skills are quite low in Tanzania. For instance, roughly 12 percent of grade 3 students could pass a grade 2 level in English (Uwezo, 2012). Moreover, there is suggestive evidence that only the best students would be close to the proficiency threshold used in the Uwezo assesment (Mbiti et al., 2017). Given the challenges of teaching English in Tanzania, the subject was removed from the national curriculum in grade 1 and 2 in 2015 to allow teachers to focus on numeracy and literacy in Kiswahili in those grades. English was still taught in grade 3, under a revised curriculum. However, the Education Ministry provided little guidance on how to transition to the new curriculum and as a result, there was substantial variation in its implementation. Some schools stopped teaching English in 2015, while others continued until 2016. In addition, there was no official guidance on whether to use grade 1 English materials in grade 3, as no new books were issued to reflect the curriculum changes. As a result, Twaweza dropped English from the incentives in grade 1 and 2 in 2016, but included grade 3 English teachers. To avoid confusion, we also communicated that our end-of-year English test in 2016 would still use the pre-reform grade 3 curriculum. Given these issues in the implementation of the curriculum reform, it became unclear how to best interpret the results for English in both years in grades 1 and 2. Therefore, in order to facilitate a clearer analysis, we only evaluate the effects of our treatments on English test scores in grade 3.

3 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.

3.1 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.[21] 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.[22] 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 teachers have identical ability (or productivity); however, this can easily be relaxed without altering the results presented below.

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

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

3.2 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 $\pi = \frac{1}{(N-1)f^l(0)}$.

3.3 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 each teacher's maximization problem as:

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

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's

n 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:

$$\sum_t \gamma^l h^l (T_t - a^l_{j(t-1)} - \gamma^l e^l_j) \frac{\Pi_t}{\sum_l \sum_k C^l_n P\left(v^l_k > T_t - a^l_{k(t-1)} - \gamma^l e^l_k\right)} = c'_l(e^l_j) \tag{4}$$

for all l, where h^l is the density function of v^l_j. 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:

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

for all l.

3.4 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.

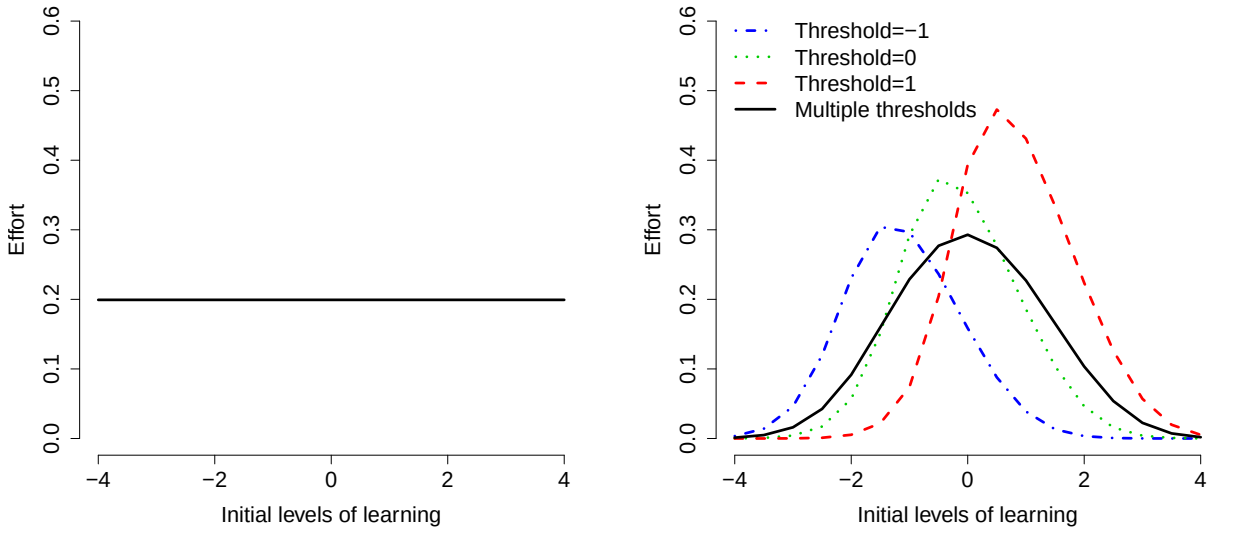
We assume that the teacher's cost function is quadratic (i.e., $c(e) = e^2$), and the shock to student learning follows a standard normal distribution (i.e., $v_i \sim N(0,1)$). We further assume that there are 1,000 teachers, each with their own classroom. Within each class, we allow student baseline learning levels to be 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.[23] 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 3.2) 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

[23] In Appendix A.2.10 we show that our qualitative results are robust to a normal distribution of student baseline learning levels.

response assuming a single proficiency threshold and then vary the threshold value from -1 to 1. We then compute the multiple threshold case.

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 1a and 1b 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 1b). 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 1: Incentive design and optimal effort with constant productivity of teacher effort



(a) Pay for Percentile - γ constant across initial levels of learning

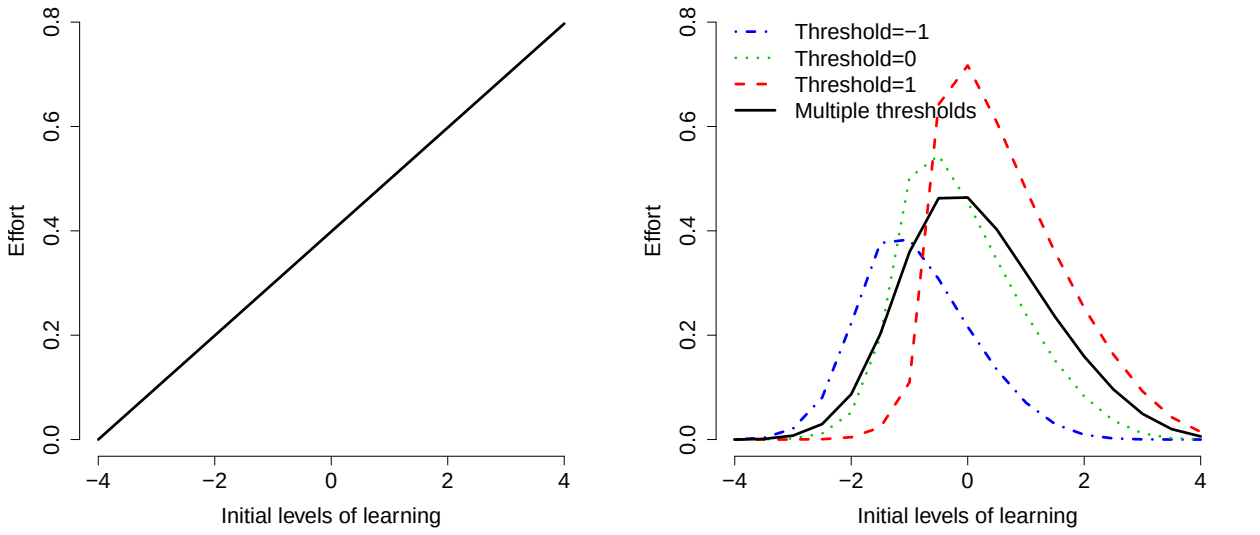
(b) Levels - γ constant across initial levels of learning

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 $\gamma^l = 2 + 0.5a^l_{(t-1)}$. Figures 2a and 2b 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 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. Moreover, the exercise suggests the teacher effort response curve is less sensitive to changes in the production function under the levels system.

Figure 2: 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

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

4 Data and Empirical Specification

4.1 Sample Selection

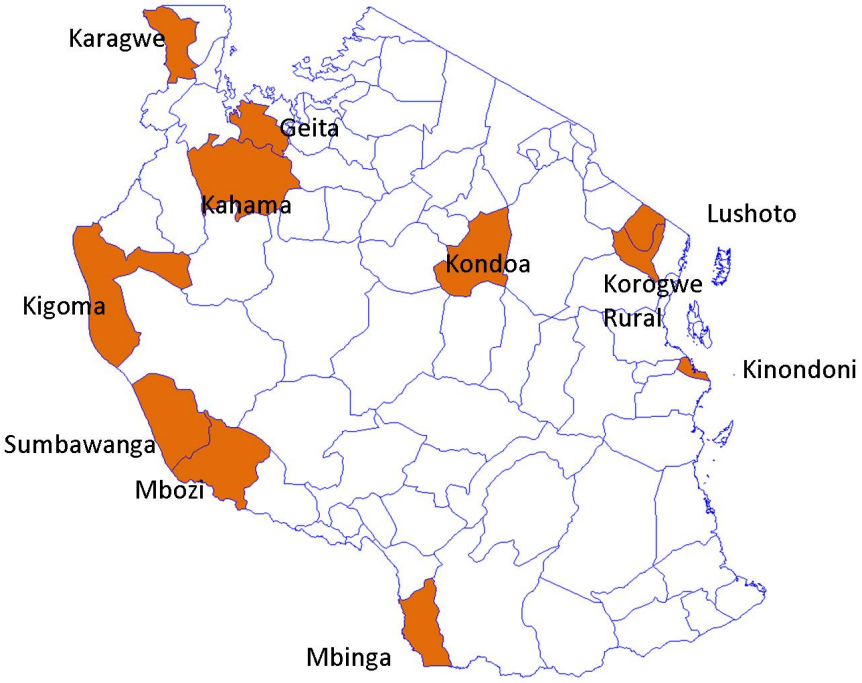
The teacher incentive programs were evaluated using a randomized design. First, 10 districts were randomly selected (see Figure 3).[24] The study sample of 180 schools was taken from a previous field experiment — studied by Mbiti et al. (2017) — where all students in grades 1, 2, and 3 had been tested at the end of 2014. These tests provided the baseline student-level test score information required to implement the pay for percentile treatment. Because pay for percentile will deliver optimal levels of effort only if teachers believe they are competing in fair contests, having reliable information about student initial learning levels was key.[25]

Within each district, we randomly allocated schools to one of our three experimental groups. Thus, in each district six schools were assigned to the levels treatment, six schools to the pay for percentile treatment, and 6 schools served as controls. In total, there were 60 schools in each group. The sample was also stratified by treatment of the previous RCT and by an index of the overall learning level of students in each school. Further details are included in Appendix A.1.

[24] The program was implemented in 11 districts, as one district was included non-randomly by Twaweza for piloting and training. We did not survey schools in the pilot district.

[25] We do not have data on whether teachers believe they are competing in a fair contest. However, before receiving any payment over 90% of teachers agree or strongly agree that the amount paid by Twaweza will be a fair, suggesting teachers think the contests are fair.

Figure 3: Districts in Tanzania from which schools are selected



Note: We drew a nationally representative sample of 180 schools from a random sample of 10 districts in Tanzania (shaded).

4.2 Data and Balance

Over the two-year evaluation, our survey teams visited each school at the beginning and end of the school year. We gathered detailed information about each school from the head teacher, including: facilities, management practices, and head teacher characteristics. We also conducted individual surveys with the teachers in our evaluation to determine personal characteristics, including education and experience, and effort measures, such as teaching practices. In addition, we conducted classroom observations, where we recorded teacher-student interactions and other measures of teacher effort such as teacher absence.

Within each school we surveyed and tested a random sample of 40 students (10 students from grades 1, 2, 3, and 4). Grade 4 students were included in our research sample in order to measure potential spillovers to other grades. Students in grades 1, 2, and 3 who were sampled in the first year of the program were tracked over the two-year evaluation period. Students in grade 4 in the first year were not tracked into grade 5 due to budget constraints. In the second year of the program we sampled an additional 10 incoming Grade 1 students. We collected a variety of data from our student sample including test scores, individual characteristics such as age and gender, and perceptions

of the school environment. Crucially, the test scores collected on the sample of students are "low-stakes" for teachers and students. We supplement the results from this set of non-incentivized student tests with the results from the incentivized tests which are used to determine teacher bonus payments and are conducted in all schools, including control schools.[26]

Although the content (subject order, question type, phrasing, difficulty level) is consistent across the incentivized and non-incentivized tests, there are a number of important differences in the test administration. The non-incentivized test took longer (40 minutes) than the incentivized test (15 minutes). The non-incentivized test had more questions in each section (Kiswahili, English and math) to avoid bottom- and top-coding, and also included an "other subject" module at the end to test spillover effects. Further, even though both tests were administered individually to students, the testing environment was different. Non-incentivized tests were administered during a regular school day. In contrast, the incentivized test was more "official" as all students in grades 1-3 were tested on a specified day. In addition, most schools used the incentivized test as the end-of-year test. 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. Finally, tests were handled, administered, and electronically scored by Twaweza teams without any teacher involvement.

Most student, school, teacher, and household characteristics are balanced across treatment arms (See Table 2, Column 4). The average student in our sample is 8.9 years old in 2013, goes to a school with 679 students, and is taught by a teacher who is 38 years old. We are able to track 88% of students in our sample at the end of the second year, with no differential attrition. Teacher turnover rates over the two-year study period were generally balanced across treatments (see Table A.3 in the appendix).

[26]Most articles studying teacher performance pay use incentivized tests to measure the overall treatment effects. However, is unclear whether incentivized or non-incentivized tests are better for measuring treatment effects. As argued by Mbiti et al. (2017) "The demonstration that test-taking effort is a salient component of measured test scores by Levitt et al. (2016) presents a conundrum for education researchers as to what the appropriate measure of human capital should be for assessing the impact of education interventions. On one hand, low-stakes tests may provide a better estimate of a true measure of human capital that does not depend on external stimuli for performance. On the other hand, test-taking effort is costly, and students may not demonstrate their true potential under low-stakes testing, in which case, an 'incentivized' testing procedure may be a better measure of true human capital.".

Table 2: Summary statistics across treatment groups at baseline (February 2015)

	(1) Control	(2) P4Pctile	(3) Levels	(4) p-value (all equal)
Panel A: Students				
Poverty index (PCA)	0.01	-0.08	0.01	0.42
	(1.99)	(1.94)	(1.98)	
Age	8.88	8.94	8.89	0.35
	(1.60)	(1.67)	(1.60)	
Male	0.50	0.48	0.51	0.05*
	(0.50)	(0.50)	(0.50)	
Kiswahili test score	-0.00	0.01	0.01	0.14
	(1.00)	(0.99)	(0.98)	
English test score	0.00	0.04	-0.02	0.71
	(1.00)	(1.03)	(1.04)	
Math test score	-0.00	-0.01	-0.01	0.56
	(1.00)	(1.04)	(1.00)	
Tested in yr0	0.91	0.89	0.90	0.41
	(0.29)	(0.31)	(0.30)	
Tested in yr1	0.87	0.87	0.88	0.20
	(0.33)	(0.34)	(0.32)	
Tested in yr2	0.88	0.88	0.89	0.56
	(0.33)	(0.32)	(0.32)	
Panel B: Schools				
Total enrollment	643.42	656.35	738.37	0.67
	(331.22)	(437.74)	(553.33)	
Facilities index (PCA)	0.18	-0.11	-0.24	0.07*
	(1.23)	(0.97)	(1.01)	
Urban	0.15	0.13	0.17	0.92
	(0.36)	(0.34)	(0.38)	
Single shift	0.63	0.62	0.62	0.95
	(0.49)	(0.49)	(0.49)	
Panel C: Teachers (Grade 1-3)				
Male	0.42	0.38	0.35	0.19
	(0.49)	(0.49)	(0.48)	
Age (Yrs)	37.89	37.02	37.70	0.18
	(11.35)	(11.23)	(11.02)	
Tertiary education	0.87	0.88	0.87	0.74
	(0.33)	(0.32)	(0.33)	

This tables presents the mean and standard error of the mean (in parentheses) for several characteristics of students (Panel A), schools (Panel B), and teachers (Panel C) across treatment groups. Column 4 shows the p-value from testing whether the mean is equal across all treatment groups (H_0 := mean is equal across groups). The p-value is for a test of equality of means, after controlling for the stratification variables used during randomization. The poverty 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 facilities index is the first component of a principal component analysis of indicator variables for: outer wall, staff room, playground, library, and kitchen. Standard errors are clustered at the school level for the test of equality. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.3 Empirical Specification

We estimate the effect of our interventions on students' test scores using the following OLS equation:

$$Z_{isd} = \delta_0 + \delta_1 Levels_s + \delta_2 P4Pctile_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_g + X_i\delta_3 + X_s\delta_4 + \varepsilon_{isd}, \quad (6)$$

where Z_{isdt} is the test score of student i in school s in district d at endline. *Levels* and *P4Pctile* are binary variables which capture the treatment assignment of each school. γ_g is a set of grade fixed effects, X_i is a series of student characteristics (age, gender and grade), X_s is a set of school characteristics including facilities, students per teacher, school committee characteristics, average teacher age, average teacher experience, average teacher qualifications, the fraction of female teachers, and the stratification controls such as previous intervention status.

We scale our test scores using an IRT model and then normalize them using the mean and variance of the control schools to facilitate a clear interpretation of our results. We include baseline test scores and district fixed effects in our specifications to increase precision.[27]

We examine the impacts of the incentives using both the non-incentivized and incentivized testing data. However, given the limited set of student characteristics in the incentivized test data, this analysis includes fewer student level controls. We use a similar specification to examine teachers' behavioral responses.

5 Results

In this section, we first explore how both incentive systems affected test scores (of both incentivized and non-incentivized subjects) in both incentivized and non-incentivized grades. We then explore whether the incentives increase observable teacher effort. We then turn to heterogeneity by baseline ability across students, and tie those results to the results in the theoretical framework. Finally, we explore some possible mechanisms to explain our results on test scores.

[27] We also balanced the timing of our survey activities, including the non-incentivized tests, across treatment arms. Hence, results are not driven by imbalanced survey timing.

5.1 Test Scores

First, we show the impact of the incentives on student learning in math, Kiswahili, and English using data from the non-incentivized test (Table 3, Panel A) as well as the incentivized test (Table 3, Panel B). In the first year, both incentive schemes resulted in small learning gains on the non-incentivized test. However, the treatment effects of the levels incentive scheme were consistently larger than those in the pay for percentile system (Panel A, columns 1, 2, and 3). In particular, the effect in Kiswahili was larger by $.078\sigma$ (p-value .084). In the second year of the program, the estimated treatment effects on the non-incentivized test are generally larger than the first-year estimates. Math test scores improved by $.068\sigma$ (p-value .089) in the levels system and $.072\sigma$ (p-value .05) in the pay for percentile system. Kiswahili test scores improved by $.11\sigma$ (p-value < 0.01) under the levels system, but only by $.058\sigma$ (p-value .1) under the pay for percentile system. Finally, English test scores improved by $.11\sigma$ (p-value .19) in the levels system, but improved by $.19\sigma$ (p-value .019) in the pay for percentile scheme. Although the results show that the estimated learning gains are generally larger under the levels system, formal hypothesis tests show that the differences are only significant for Kiswahili in year one (Panel A, Column 2).

Most of the existing literature on teacher incentives relies on data from incentivized tests that are used to determine teacher rewards (Muralidharan & Sundararaman, 2011; Fryer, 2013; Neal & Schanzenbach, 2010). Following this norm, we also present the treatment effects of our interventions using incentivized exams (Panel B). Generally, the estimated treatment effects are larger compared to those estimated using the non-incentivized test (Panel A). However, these differences are not statistically significant in most cases (Panel C). In addition, the differences between the estimated treatment effects across the two incentive designs tend to be larger in the incentivized test. For example, the effect of the levels scheme is larger for Kiswahili by $.11\sigma$ (p-value .026) in the first year and by $.093\sigma$ (p-value .045) in the second year.

The larger treatment effects found in the incentivized test are likely driven by test-taking effort, where teachers had incentives to motivate their students to take the tests seriously. The importance of student test-taking effort has been documented in other settings such as an evaluation of teacher and student incentives in Mexico City (Behrman, Parker, Todd, & Wolpin, 2015). As discussed in Section 4.2, administration of the incentivized test was tightly controlled by our implementation team. This mitigates any concerns about outright cheating. Assuming that all the differences between our incentivized and non-incentivized results are driven by test-taking effort, this suggests that student effort can increase test score results between 0.02σ and 0.2σ (see Panel C). This

is generally in line with the findings of Gneezy et al. (2017) and Levitt et al. (2016).

Given the reward structure, teachers in both treatment arms were motivated to ensure that their students took the incentivized test. In the second year of the study, teachers in the levels schools were able to increase student participation in the incentivized test by 5 percentage points. Their counterparts in pay for percentile schools increased participation by 3 percentage points (see Table A.4). Following Lee (2009), we compute bounds on the treatment effects by trimming the excess test takers from the left and right tails of the incentivized test distribution. Focusing on the year-two results for brevity, this bounding exercise suggests that the treatment effects for math range from -0.023 to 0.32 in the levels treatment and 0.014 to 0.17 in the pay for percentile treatment. The bounds for Kiswahili range from 0.027 to 0.35 in the levels and -0.0032 to 0.17 in the pay for percentile (see Table A.5).

As discussed previously, we had limited information to properly group grade 1 students in the pay for percentile arm. As this may limit the effectiveness of the pay for percentile scheme, we examine the effects of our interventions by focusing on grade 2 and 3 students, where we are able to appropriately group students by ability. Our results are generally robust to this sample restriction (see Table A.2 in the appendix). Our results are also robust to a composite measure of learning across all subjects (see Table **??** in the appendix). Hence, false discovery due to multiple hypothesis testing is not a substantial concern.

Table 3: Effect on test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Panel A: Non-incentivized								
Levels (α_1)	.039	.045	.012	.04	.068*	.11***	.11	.098**
	(.047)	(.048)	(.086)	(.047)	(.04)	(.039)	(.085)	(.038)
P4Pctile (α_2)	-.015	-.033	-.049	-.032	.072**	.058*	.19**	.086**
	(.04)	(.039)	(.076)	(.039)	(.037)	(.035)	(.081)	(.035)
N. of obs.	4,781	4,781	1,532	4,781	4,869	4,869	1,533	4,869
$\alpha_3 = \alpha_2 - \alpha_1$	-.053	-.078*	-.061	-.072*	.0047	-.056	.08	-.012
p-value ($H_0 : \alpha_3 = 0$)	.23	.084	.43	.092	.92	.17	.29	.77
Panel B: Incentivized								
Levels (β_1)	.11**	.13***	.18***	.19***	.14***	.18***	.28***	.25***
	(.047)	(.048)	(.067)	(.066)	(.045)	(.046)	(.069)	(.061)
P4Pctile (β_2)	.066*	.017	.16***	.081	.093**	.085*	.23***	.15**
	(.039)	(.043)	(.058)	(.057)	(.04)	(.045)	(.055)	(.058)
N. of obs.	48,077	48,077	14,664	48,077	59,680	59,680	15,458	59,680
$\beta_3 = \beta_2 - \beta_1$	-.047	-.11**	-.014	-.11*	-.044	-.093**	-.047	-.099
p-value ($H_0 : \beta_3 = 0$)	0.30	0.026	0.83	0.090	0.31	0.045	0.53	0.10
Panel C: Incentivized – Non-incentivized								
$\beta_1 - \alpha_1$	.065	.075	.14	.14	.063	.056	.15	.13
p-value($\beta_1 - \alpha_1 = 0$)	.13	.1	.12	.014	.12	.21	.14	.015
$\beta_2 - \alpha_2$	.075	.044	.2	.1	.019	.024	.04	.054
p-value($\beta_2 - \alpha_2 = 0$)	.083	.32	.017	.066	.64	.58	.66	.31
$\beta_3 - \alpha_3$	.01	-.031	.053	-.033	-.044	-.032	-.11	-.079
p-value($\beta_3 - \alpha_3 = 0$)	.81	.51	.54	.55	.28	.48	.28	.16

Results from estimating Equation 6 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. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.2 Spillovers to Other Grades and Subjects

As the teacher incentives only covered math, English, and Kiswahili in grades 1, 2, and 3, there are concerns that teachers and schools could focus on these grades and subjects to the detriment of other grades and subjects. For example, schools may shift resources such as textbook purchases from higher grades to grades 1, 2, and 3. In

addition, teachers may cut back on teaching non-incentivized subjects such as science. On the other hand, if our incentive programs improve literacy and numeracy skills, they may promote student learning in other subjects and these gains may persist over time. In order to asses possible spillovers, we examine learning outcomes in science for grades 1, 2, and 3. We also examine test scores in grade 4 to test for any negative spillovers in higher grades, as well as the persistence of any learning gains induced by the program (in the second year of the evaluation).

Overall, we do not see decreases in test scores of fourth graders, which suggests that schools were not disproportionately shifting resources away from higher grades (Table 4, Panel A). In the first year of the program, we find statistically significant increases in grade 4 test scores for math (0.13σ), and English (0.17σ), with a relatively small and statistically insignificant effect for Kiswahili (0.04σ). In contrast, we find relatively small effects and statistically insignificant effects across all three subjects in the pay for percentile arm. Further, the levels treatment effects were generally significantly larger than the pay for percentile effects. However, in the second year of the program we do not find any positive and significant treatment effects. Further, the estimated differences between our interventions are no longer statistically significant (Table 4, Panel A, Columns 3 and 4). Since we tested fourth-grade students and collected information on those students at baseline, it is possible that fourth-grade teachers assumed they would be included in the incentives. As a result of this belief, they may have exerted effort in the first year, but not in the second year once their non-eligibility had been confirmed. This type of spillover was also documented by Kremer, Miguel, and Thornton (2009), where student incentives for girls improved performance for non-eligible boys who believed they would also benefit from the program.

As third graders in the first year of our program transitioned to the fourth grade in the second year of the program, the fourth-grade results in the second year suggest that the learning gains from both incentive programs fade over time (Table 4, Panel A, Columns 3 and 4). Contrary to the concerns of teacher performance pay critics, the effects of both programs on science test scores are generally positive, suggesting that any estimated gains attributable to the incentives are not coming at the expense of learning in other subjects or domains (see Table 4, Panel B).

Table 4: Spillovers to other grades and subjects

Panel A: Grade 4

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Levels (α_1)	.13**	.044	.17**	.13**	.059	.042	.085	.07
	(.062)	(.051)	(.084)	(.062)	(.062)	(.065)	(.07)	(.061)
P4Pctile (α_2)	-.03	-.033	.031	-.014	-.0038	.025	.056	.029
	(.054)	(.054)	(.078)	(.057)	(.06)	(.061)	(.063)	(.057)
N. of obs.	1,513	1,513	1,513	1,513	1,482	1,482	1,482	1,482
$\alpha_3 = \alpha_2 - \alpha_1$	-.16**	-.077	-.14*	-.14**	-.063	-.017	-.029	-.041
p-value ($H_0 : \alpha_3 = 0$)	.011	.13	.079	.019	.27	.79	.67	.49

Panel B: Science (Grades 1-3)

	Year 1	Year 2
Levels (α_1)	.069	.083
	(.063)	(.06)
P4Pctile (α_2)	-.002	.078
	(.05)	(.057)
N. of obs.	4,781	4,869
$\alpha_3 = \alpha_2 - \alpha_1$	-.071	-.0055
p-value ($H_0 : \alpha_3 = 0$)	.26	.92

Results from estimating Equation 6 for grade 4 students (Panel B) and for grade 3 students in science (Panel A). 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). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.3 Teacher Effort

Since the treatments were designed to elicit teacher effort, in this section we examine teacher responsiveness to the incentives. We use teacher presence in school and in the classroom as broad measures of teacher effort. Teacher presence was measured by our survey team and was collected shortly after our team arrived at a school in the morning. Overall, we do not find any effect in this dimension of teacher effort across our treatments (see Table 5, Panel A). The effects are close to zero and precisely estimated. For example, we can rule out that teacher attendance decreased due to either treatment by more than 7 percentage points (from a base of 70%) at the 95% level. The difference in the treatment effect across treatments is also small and close to zero. We also examine student reports about teacher effort such as assigning homework and providing extra help. We do not find any effect of either treatment in providing extra help (and the dif-

ference across treatments is small). However, students report receiving more homework from teachers in the levels systems: An increase of 1.7 percentage points (p-value 0.24) relative to a base of 12% in the control group and an increase of 3.2 percentage points (p-value 0.065) relative to a base of 10.6% in the pay for percentile group.

Table 5: Treatment effects on teacher behavior

	(1)	(2)
Panel A: Spot checks		
	In school	In classroom
Levels (α_1)	-0.0065	0.015
	(0.038)	(0.039)
P4Pctile (α_2)	-0.0085	0.000017
	(0.032)	(0.036)
N. of obs.	360	360
Mean control	.7	.36
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.002	-.015
p-value ($H_0 : \alpha_3 = 0$)	.96	.69
Panel B: Student reports		
	Extra help	Homework
Levels (α_1)	0.0080	0.017
	(0.0097)	(0.015)
P4Pctile (α_2)	-0.0022	-0.014
	(0.0091)	(0.015)
N. of obs.	18,563	18,563
Mean control	.062	.12
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.01	-.032*
p-value ($H_0 : \alpha_3 = 0$)	.3	.065

Panel A presents teacher-level data on teacher absenteeism (Column 1), and time-on-task (Column 2). Panel B presents student-level data on teacher effort (as reported by students) on extra help (Column 1) and homework (Column 2). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In addition, we measure teacher effort with classroom observations. Since in-class observations are typically affected by Hawthorne effects, our survey teams collected data on teacher behavior by standing outside the classroom for several minutes, before teachers noticed they were being observed. Although these reports are not as detailed

as in-classroom observation protocols, they are arguably better able to capture broad measures of typical teacher behavior.[28] Our findings are shown in Table 6, and we focus on the estimated differences between the two incentive systems reported in the bottom row (α_3). We do not find any statistically significant differences in the likelihood that teachers were observed to be teaching, although the point estimates are larger for levels teachers (Column 1). Teachers in pay for percentile schools were 2.2 percentage points (almost 50 percent) less likely to be engaged in classroom management activities (such as taking attendance or disciplining students) compared to levels teachers (Column 2). Teachers in pay for percentile schools were also 7.7 percentage points (29 percent) more likely to be off-task or engaged in unrelated activities such as reading a newspaper or sending a text message (Column 3). Finally, we do not observe differences between the two incentives in distracted or off-task students, although the coefficient on pay for percentile schools shows a larger reduction in student distraction (Column 4).

Table 6: External classroom observation

	(1) Teaching	(2) Classroom management	(3) Teacher off-task	(4) Student off-task
Levels (α_1)	0.011	-0.0016	-0.011	-0.0068
	(0.043)	(0.010)	(0.042)	(0.018)
P4Pctile (α_2)	-0.048	-0.024**	0.066*	-0.023*
	(0.036)	(0.011)	(0.035)	(0.014)
N. of obs.	2,080	2,080	2,080	2,080
Control mean	.69	.041	.27	.048
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.059	-.022**	.077*	-.016
p-value ($H_0 : \alpha_3 = 0$)	.2	.037	.082	.28

The outcome variables in this table come from independent classroom observations performed by the research team for several minutes, before teachers noticed they were being observed. Teachers are classified doing one of three activities: Teaching (Column 1), managing the classroom (Column 2), and being off-task (Column 3). If students are distracted we classify the class as having students off-task (Column 4). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[28]Schools in Tanzania have open layouts where classrooms are built around an open space in the middle. This allows surveyors to simply stand in the open space and observe the class from a distance.

5.4 Heterogeneity by Student Characteristics

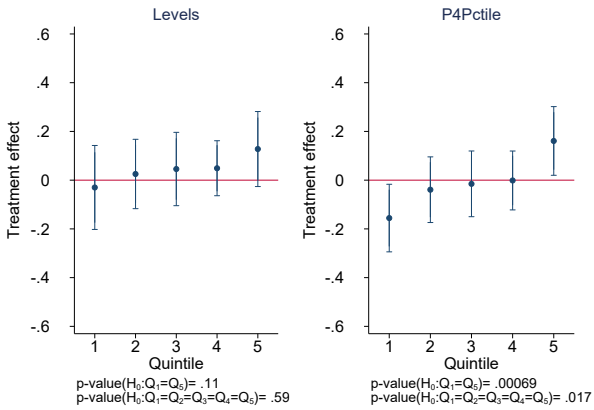
We explore the heterogeneity in treatment effects across the distribution of student baseline test scores in Figures 4 (math), 5 (Kiswahili), and 6 (English).[29] In addition to providing evidence on which students benefit more from the incentives, the analysis also sheds light on the functional form of the productivity of teacher effort (see Section 3). In particular, if the treatment effects in the pay for percentile system are equal in all student ability groups, then this would imply that the productivity of teacher effort does not vary by student ability, as shown in Figure 1. However, if better prepared students benefit more in the pay for percentile scheme, then this would suggest that the productivity of teacher effort is higher for better prepared students as shown in Figure 2.

In the first year of the program, both math and Kiswahili teachers in the pay for percentile system (labeled "P4Pctile") focused their attention on their best students, whereas teachers in the levels system (labeled "levels") focused equally on all students (Figures 4a and 5a). English teachers under both systems seemed to focus more on top students, although none of the individual quintile estimates are statistically significant (Figure 6a). In the second year of the program, we do not see such overt focus on top students in mathematics in either incentive system (Figure 4b). However, Kiswahili teachers under the levels system focused on all of their students in the second year, while teachers in the pay for percentile system focused on the very best students (Figure 5b). In contrast, English teachers in the levels scheme focused on the top students, while teachers in the pay for percentile seemed to focus more on the middle quintiles.

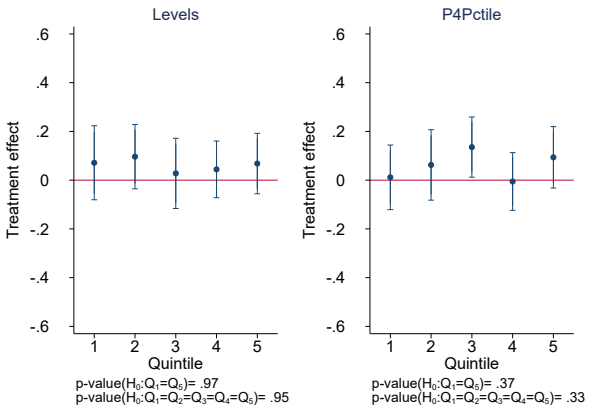
Overall, the pay for percentile results in math (Year 1) and Kiswahili (both years) suggest that the productivity of teacher effort is higher among better prepared students. Given the changes in the curriculum and the general difficulty of teaching the language in Tanzania, the results in English are less informative.

[29] We also explore heterogeneity by additional student characteristics such as gender, as well as school characteristics such as pupil teacher ratio, and find limited evidence of heterogeneity in those characteristics (see Tables A.6 and A.7 for details).

Figure 4: Math

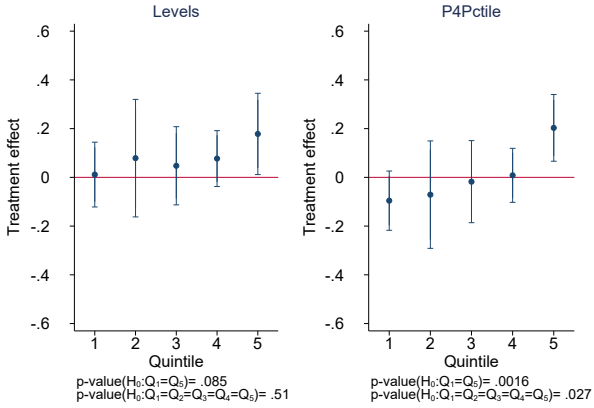


(a) Year 1

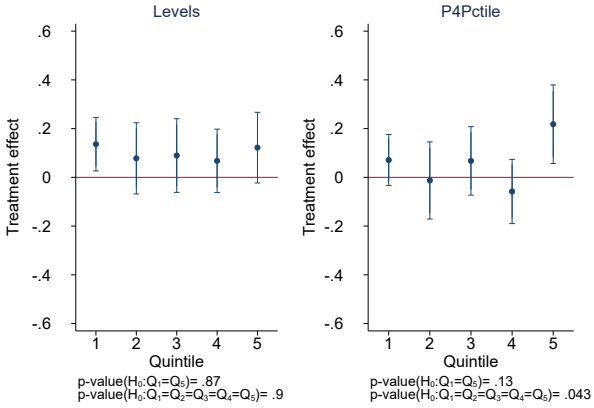


(b) Year 2

Figure 5: Kiswahili

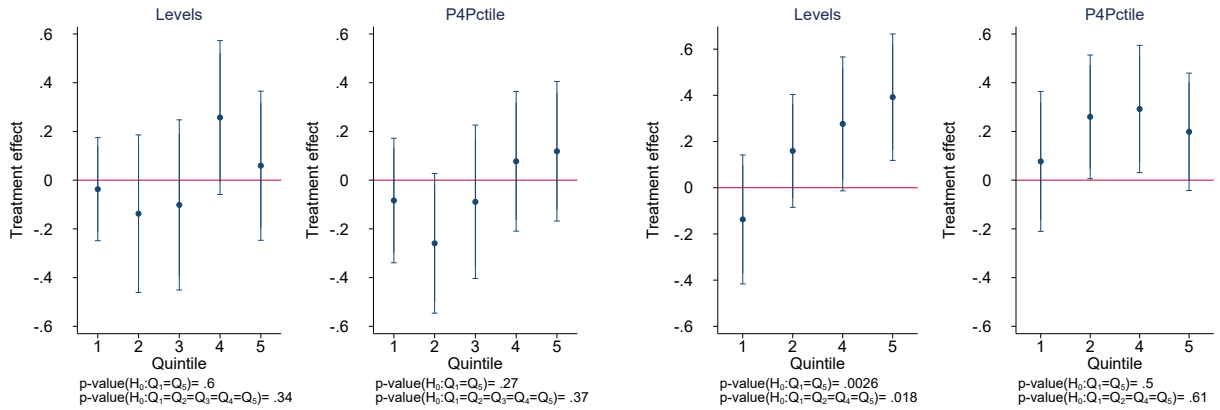


(a) Year 1



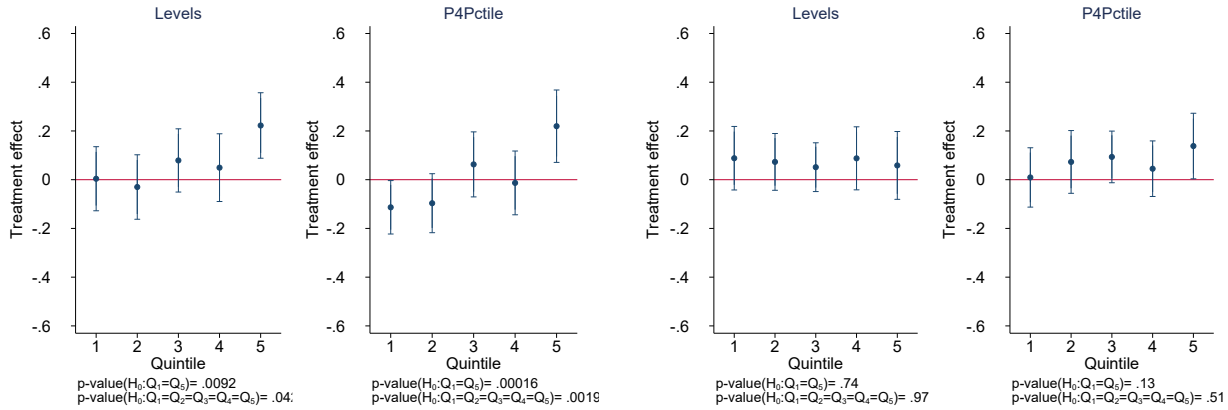
(b) Year 2

Figure 6: English



(a) Year 1 (b) Year 2

Figure 7: Composite (PCA)



(a) Year 1 (b) Year 2

5.5 Heterogeneity by Teacher Characteristics

Empirical evidence shows that women are averse to competition and exert relatively less effort than men in competitive situations such as rank-order tournaments (Niederle & Vesterlund, 2007, 2011). Although the competitive nature of the pay for percentile scheme seems more salient relative to the levels scheme, we do not find that women are less responsive to its competitive pressure (Table 7, Column 1). We also do not find any heterogeneous effects by teacher age, which proxies for experience.

Although studies such as Metzler and Woessmann (2012) and Bietenbeck, Piopiunik, and Wiederhold (2018) have shown that teacher content knowledge is predictive of stu-

dent learning outcomes, we do not find any significant heterogeneity in our treatment effects by teacher content knowledge (Column 3). More effective teachers, as measured by the head teacher's rating, are more responsive on average to the levels incentives compared to teachers in the pay for percentile system. These differences are significant for math (Panel A, Column 3) and English (Panel C, Column 3). Our findings could potentially reflect a greater discouragement effect among pay for percentile teachers relative to levels teachers.[30] We also examine heterogeneous effects by teacher beliefs about their individual efficacy in Column 5. Generally, teachers who believed they were more capable responded more to both incentives (Column 5, Panel A and Panel B). However, we find the reverse relationship in English (Column 5, Panel C). Overall, the patterns of heterogeneity by teachers' self-ratings are statistically indistinguishable across the two incentive designs.

[30]Previous studies on rank-order tournaments, such as Brown (2011) and Schotter and Weigelt (1992), have shown that heterogeneity in participant ability can negatively impact the efficacy of tournaments. For example, if a tournament features a number of strong players, then less-able players may (correctly) surmise that they have a limited chance of winning. As a result, such players may be discouraged from increasing their effort when faced with strong players. Similarly, strong players may reduce their effort when faced with less-able competitors (Brown, 2011). In the context of our study, this theory suggests that heterogeneity in ability among teachers may reduce the efficacy of the incentives among both the least and most able teachers in the pay for percentile scheme. In contrast, we would not expect to find a discouragement effect of heterogeneity in the levels incentives.

Table 7: Heterogeneity by teacher characteristics

	(1)	(2)	(3)	(4)	(5)
Panel A: Math					
	Male	Age	IRT	HT Rating	Self Rating
Levels*Covariate (α_2)	0.033	0.00080	0.016	0.073***	0.041
	(0.070)	(0.0016)	(0.037)	(0.021)	(0.035)
P4Pctile*Covariate (α_1)	-0.017	0.00056	-0.025	0.012	0.058*
	(0.060)	(0.0016)	(0.038)	(0.022)	(0.035)
N. of obs.	9,650	9,650	9,650	4,869	9,650
$\alpha_3 = \alpha_2 - \alpha_1$	-.05	-.00024	-.041	-.062**	.017
p-value ($H_0 : \alpha_3 = 0$)	.49	.88	.2	.012	.61
Panel B: Kiswahili					
	Male	Age	IRT	HT Rating	Self Rating
Levels*Covariate (α_2)	-0.081	-0.0000038	0.0022	0.069**	0.085**
	(0.069)	(0.0011)	(0.034)	(0.031)	(0.034)
P4Pctile*Covariate (α_1)	0.013	0.000058	0.0053	0.051	0.076**
	(0.067)	(0.0011)	(0.030)	(0.034)	(0.032)
N. of obs.	9,650	9,650	9,650	4,869	9,650
$\alpha_3 = \alpha_2 - \alpha_1$	.094	.000062	.0031	-.019	-.0092
p-value ($H_0 : \alpha_3 = 0$)	.19	.95	.93	.56	.8
Panel C: English					
	Male	Age	IRT	HT Rating	Self Rating
Levels*Covariate (α_2)	0.082	0.0039	0.071	0	-0.091
	(0.12)	(0.0024)	(0.098)	(.)	(0.081)
P4Pctile*Covariate (α_1)	-0.011	0.0013	-0.068	0	-0.15**
	(0.12)	(0.0024)	(0.088)	(.)	(0.076)
N. of obs.	6,314	6,314	6,314	0	6,314
$\alpha_3 = \alpha_2 - \alpha_1$	-.093	-.0025	-.14	0	-.064
p-value ($H_0 : \alpha_3 = 0$)	.46	.29	.14	.	.34

The outcome variables are student test scores. The data include both follow-ups. Each column shows the heterogeneous treatment effect by different teacher characteristics: sex (Column 1), age (Column 2), content knowledge scaled by an IRT model (Column 3), head teacher rating (Column 4) — only requested for math and Kiswahili teachers at the end of the second year — and self rating (Column 5). We use three different measures of teacher ability to explore the heterogeneity in treatment effects. Teachers were tested on all three subjects and we created an index of content knowledge using an IRT model. Head teachers were asked to rate teacher performance in seven dimensions, including the ability to ensure that students learn, and classroom management skills. To create the self-perception metric, we create an index based on teacher responses to the following five statements: "I am capable of motivating students who show low interest in school", "I am capable of implementing alternative strategies in my classroom", "I am capable of getting students to believe they can do well in school", "I am capable of assisting families in helping their children do well in school", and "I am capable of providing an alternative explanation or example when students are confused". Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.6 Mechanisms

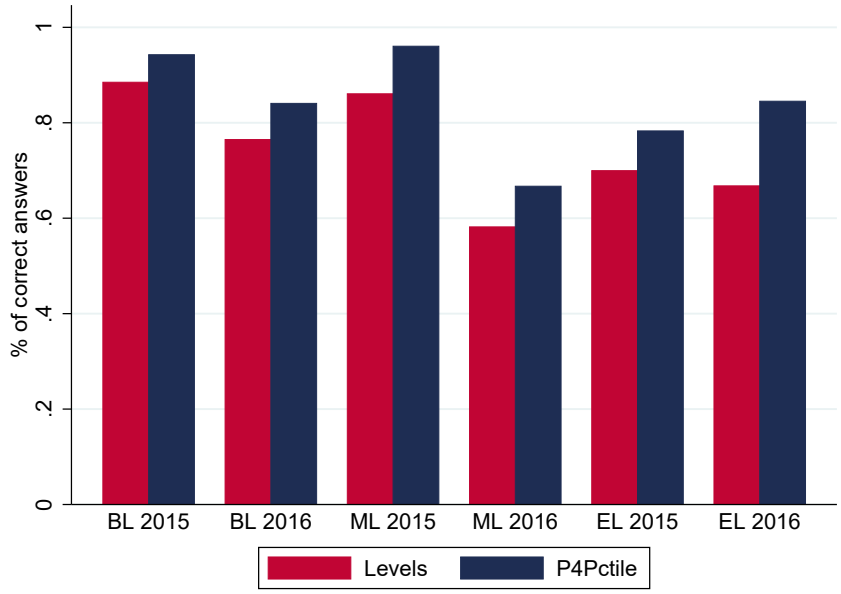
We examine potential mechanisms that could drive differences in behavior and outcomes between the two types of incentives, with a particular focus on differences in the incentive structures of the two systems. For instance, the levels system is easier to understand and could provide clear learning targets for classrooms, compared to the pay for percentile system. This difference in clarity could also affect teachers' expectations about their potential rewards from the incentive programs, which would ultimately affect the level of effort exerted. We also examine the potential for our interventions to differentially encourage sabotage or promote divisions among teachers (see the appendix).

Teachers understand both programs. Complex teacher incentive programs may be less effective if teachers cannot understand the program details and therefore do not optimally allocate their effort (Goodman & Turner, 2013; Loyalka et al., in press). These concerns are potentially more important in contexts with weak state capacity, which may be less able to effectively disseminate the details of a complex incentive program to teachers. Because the pay for percentile system is more complex, our results may reflect differences in teacher understanding of the incentive systems. To ameliorate these concerns, we developed culturally appropriate materials, including examples, analogies, and illustrations, which we used to communicate the details of the incentive program to teachers.[31] We also sent teams to visit schools multiple times to reinforce teachers' familiarity with the main features of the program. During our visits, we tested teachers to ensure they understood the details of the incentive program they were assigned to. We then conducted a review session to discuss the answers to the test questions to further ensure that teachers understood the design details. The results of the teacher comprehension tests are shown in Figure 8. Since we asked different questions during each survey round (baseline, midline and endline), we cannot compare the trends in understanding over time. However, despite the lack of temporal comparability, teacher comprehension was generally high and roughly equal across both types of incentive programs. [32]

[31] We worked closely with Twaweza's communications unit to develop our dissemination strategy and communications. The communications unit is experienced and highly specialized in developing materials to inform and educate the general public in Tanzania.

[32] Although teacher understanding was relatively high, we also test for heterogeneity in treatment effects by teachers' understanding (at endline). We do not find any significant relationship between teacher understanding and student test scores. The results are shown in Table A.8 in the appendix.

Figure 8: Do Teachers Understand the Interventions?



Teachers expect higher earning in the levels system. Even though we equalized the budgets across treatments, it is still possible that teachers' beliefs about their potential earnings could differ across the two incentive systems. In the pay for percentile system, the fact that the final bonus payment depends on the relative performance of other teachers is more salient. Hence, teachers may be less confident about their ability to receive larger payouts compared to their peers in the levels treatment, where payouts are determined students' proficiency levels. Prior to payout of the bonuses, we collected data on teachers' earnings expectations from the incentives, as well as their beliefs about their performance relative to other teachers in the district. As these questions were only applicable to teachers in the incentive programs, we compare teachers in the pay for percentile arm to the levels program, which serves as the omitted category in Table 8.

Teachers in pay for percentile schools had lower bonus earnings expectations compared to their peers in the levels system. They expected almost 95,000 TZS (US$ 42) less in bonus payments than teachers in the levels system. This represents an 18 percent difference in bonus expectations relative to the mean expectations of teachers in the levels system (Column 1) and 36 percent of the realized mean bonus payment in 2016. The lower expectations among pay for percentile teachers could be driven by the greater uncertainty of earnings in rank-order tournaments such as pay for percentile systems. While the competitive pressure can be motivating, it can also be demotivating if an individual teacher has low subjective beliefs of winning relative to other competitors.

We also examine differences in teachers' beliefs about their relative ranking within their district based on their (expected) bonus winnings in columns 2 to 4. Overall, we do not find any differences across the treatments in teachers' beliefs about their rankings. The results suggest that teachers were quite optimistic about their projected earnings: Only 9 percent of teachers expected to be among the bottom earners (Column 2) and 7 percent were worried about earning a low bonus (Column 5). On the other hand, 80 percent expected to be among the top earners in the district (Column 4).

Table 8: Teachers' earning expectations

	Bonus (TZS)	Bottom of the district	Middle of the district	Top of the district	Worried low bonus
	(1)	(2)	(3)	(4)	(5)
P4Pctile (α_2)	-94,330**	-.029	-.0092	.035	-.02
	(37,169)	(.03)	(.059)	(.045)	(.026)
N. of obs.	653	676	676	676	676
Mean Levels	525,641	.086	.48	.8	.074

This table show the effect of treatment on teacher self-reported expectations: The expected payoff (Column 1), the expected relative ranking in the district (Columns 2-4), and whether the teacher is worried about receiving a low bonus payment (Column 5). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Goal setting is easier for teachers in the levels system. In addition to being relatively easier to understand, the levels system provides teachers with a set of clear learning targets and goals for their students. This can help guide their instructional strategies and areas of focus in the classroom, and perhaps even support individualized coaching.[33] We ask teachers about whether they have professional goals for the academic year, and what this goals are. We do not find any differences in the likelihood of setting up goals for the general school exams (Table 9, Column 1). However, teachers in the levels system were almost 8 percentage points more likely (or twice as likely) to have set clear goals for the incentivized Twaweza test (Column 2). In contrast, teachers in pay for percentile schools were 2.5 percentage points (p-value, 0.35) more likely to have set clear goals (Column 2). Although we cannot reject the equality of the two estimates, the results

[33]Recent papers in the (behavioral) economics literature provide evidence on general productivity effects of setting goals, for example Koch and Nafziger (2011); Gómez-Minambres (2012) and Dalton, Gonzalez, and Noussair (2015).

provide some suggestive evidence that the levels systems facilitated more goal-setting and targeting on the incentivized (Twaweza) test. In regard to the incentivized Twaweza test, which was administered in all schools, teachers in both types of incentives schools were approximately 7 percentage points (roughly 8 percent) more likely to set a general goal for the test that teachers in control schools (Column 3). Additionally, teachers in levels schools were almost 10 percentage points (50 percent) more likely to set a specific numerical target for the Twaweza incentivized test, compared to just under 4 percent of teachers in pay for percentile schools (Column 4). Although these differences are not statistically distinguishable, the point estimates suggest greater incidences of goal-setting among teachers in the levels design.

Table 9: Goal-setting

	Goals		Twaweza test goals	
	School exam	Twaweza exam	General	Specific (number)
	(1)	(2)	(3)	(4)
Levels (α_1)	-.02	.076**	.067**	.095*
	(.053)	(.029)	(.031)	(.052)
P4Pctile (α_2)	-.047	.025	.076***	.036
	(.048)	(.027)	(.022)	(.042)
N. of obs.	1,016	1,016	1,016	1,016
Mean control	.46	.078	.89	.19
$\alpha_3 = \alpha_2 - \alpha_1$	-.027	-.05	.0094	-.059
p-value($\alpha_3 = 0$)	.58	.14	.7	.27

This table shows the effect of treatment on whether teachers set professional goals (columns 1-2) and specific goals for the Twaweza exam (columns 3-4); specifically, whether they set goals for the school exams (Column 1) and the Twaweza exams (Column 2). In addition, it indicates whether they have general goals for student performance on the Twaweza exam (Column 3) or specific (numeric) goals (Column 4). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6 Cost-effectiveness

We use accounting records to examine the cost-effectiveness of our interventions, following the framework outlined in Dhaliwal, Duflo, Glennerster, and Tulloch (2013). The

total annual cost of the teacher incentive programs was US$ 7.23 per student. This cost estimate includes both the direct costs (value of incentive payments) as well as the implementation costs (test design and implementation, communications, audit, transfer costs, etc.) of the program. [34]

For each intervention, we use the highest and lowest estimated treatment effects on the incentivized test in the second year to compute the cost-effectiveness range. We focus on the incentivized test to facilitate comparability with other teacher incentive studies. As the pay for percentile treatment effects range from 0.085σ to 0.23σ in the second year, the cost-effectiveness of the intervention ranges from 1.18σ to 3.18σ per US$ 100 spent per child. The levels treatment effects range from 0.14σ to 0.28σ. Thus, the cost-effectiveness of the levels treatment ranges from 1.94σ to 3.87σ per US$ 100 spent per child. These estimates suggest that both programs are cost-effective compared to several other interventions in developing countries analyzed in the overview by (Kremer, Brannen, & Glennerster, 2013). For instance, using the lower-bound estimates for the levels treatments, our intervention is more cost-effective than a computer-assisted learning program evaluated in India (1.54σ per US$ 100), but less effective than the incentive program on attendance in India (2.28σ per US$ 100).

7 Conclusion

We use a randomized controlled trial to compare the effectiveness of two different teacher incentive programs aimed at improving early-grade learning in Tanzanian public schools. Specifically, we compare the effectiveness of a simple, multiple-threshold proficiency incentive design relative to a more complex, rank-order, tournament-style pay for percentile system in terms of their impact on independently measured test scores of students in grades 1-3.

Our main findings are summarized as follows: first, we find that both programs lead to increases in test scores in the focal grades, compared to students in the control group. Second, we do not find any negative effects on test scores for non-incentivized subjects or grades. Third, despite the theoretical advantage of the pay for percentile system, we find that the simpler proficiency system was just as effective, and sometimes more effective,

[34] The costs of pre-treatment testing required in pay for percentile are not included in the cost figure, since this cost would only be incurred once (ability groups could be based on endline data after the first year of implementation). Our calculations also assume similar data management costs for both programs, even though in reality the pay for percentile data costs were higher due to tasks such as preparing the ability groups and programming the payment calculations. However, these are largely fixed costs and relatively small relative to the variable costs, especially at scale.

than the pay for percentile system. In particular, we find that point estimates of impact on incentivized test scores are higher under the proficiency system for all incentivized subjects, and significantly higher for Kiswahili. Finally, our findings suggest that the pay for percentile system may encourage teachers to focus more on better-prepared students, especially in Kiswahili. Our findings are consistent with a model where the added uncertainty and complexity of competing with other teachers reduces teachers' perceptions of the expected reward from the pay for percentile system. Further, the results suggest that better prepared students benefit more from teacher effort in the education production function.

Our results allow us to explore the trade-offs faced by education authorities who have to consider the effectiveness, feasibility of implementation, and equity of different incentive designs with limited information about the education production function. This demonstrates the challenge of designing effective teacher incentive schemes in settings with limited administrative capacity. While proficiency systems may be more effective than complex systems in the short run, the reverse may be true in the long run since pay for percentile schemes are harder to game over time. Thus, additional experiments conducted over several years are needed to experimentally examine whether pay for percentile systems would dominate simpler systems over a longer-run horizon.

References

Alger, V. E. (2014). *Teacher incentive pay that works: A global survey of programs that improve student achievement.*

Bandiera, O., Barankay, I., & Rasul, I. (2007). Incentives for managers and inequality among workers: Evidence from a firm-level experiment. *The Quarterly Journal of Economics*, *122*(2), 729–773.

Banerjee, A., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. PublicAffairs.

Barlevy, G., & Neal, D. (2012). Pay for percentile. *American Economic Review*, *102*(5), 1805-31. Retrieved from http://www.aeaweb.org/articles.php?doi=10.1257/aer.102.5.1805 doi: 10.1257/aer.102.5.1805

Barrera-Osorio, F., & Raju, D. (in press). Teacher performance pay: Experimental evidence from pakistan. *Journal of Public Economics*.

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in mexican

high schools. *Journal of Political Economy*, *123*(2), 325-364. Retrieved from https://doi.org/10.1086/675910 doi: 10.1086/675910

Bettinger, E. P., & Long, B. T. (2010, August). Does cheaper mean better? the impact of using adjunct instructors on student outcomes. *The Review of Economics and Statistics*, *92*(3), 598-613. Retrieved from http://ideas.repec.org/a/tpr/restat/v92y2010i3p598-613.html

Bietenbeck, J., Piopiunik, M., & Wiederhold, S. (2018). Africa's skill tragedy does teachers' lack of knowledge lead to low student performance? *Journal of Human Resources*, *53*(3), 553–578.

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. *Journal of Political Economy*, *119*(5), 982-1013.

Bruns, B., Filmer, D., & Patrinos, H. A. (2011). *Making schools work: New evidence on accountability reforms*. World Bank Publications.

Bruns, B., & Luque, J. (2015). *Great teachers: How to raise student learning in latin america and the caribbean*. World Bank Publications.

Carroll, G. (2015). Robustness and linear contracts. *American Economic Review*, *105*(2), 536–63.

Carroll, G., & Meng, D. (2016). Locally robust contracts for moral hazard. *Journal of Mathematical Economics*, *62*, 36–51.

Charness, G., & Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? In *Handbook of labor economics* (Vol. 4, pp. 229–330). Elsevier.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–79.

Contreras, D., & Rau, T. (2012). Tournament incentives for teachers: evidence from a scaled-up intervention in chile. *Economic development and cultural change*, *61*(1), 219–246.

Dalton, P. S., Gonzalez, V., & Noussair, C. N. (2015). *Paying with self-chosen goals: Incentives and gender differences* (Tech. Rep.).

Dhaliwal, I., Duflo, E., Glennerster, R., & Tulloch, C. (2013). Comparative cost-effectiveness analysis to inform policy in developing countries: a general framework with applications for education. *Education Policy in Developing Countries*, 285–338.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, *102*(4), 1241–78.

Ferraz, C., & Bruns, B. (2012). Paying teachers to perform: The impact of bonus pay in pernambuco, brazil. *Society for Research on Educational Effectiveness*.

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from new york city public schools. *Journal of Labor Economics*, *31*(2), 373–407.

Gilligan, D., Karachiwalla, N., Kasirye, I., Lucas, A., & Neal, D. (2018). *Educator incentives and educational triage in rural primary schools.* (mimeo)

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2*(3), 205-27. Retrieved from http://www.aeaweb.org/articles.php?doi=10.1257/app.2.3.205 doi: 10.1257/app.2.3.205

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

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

Gómez-Minambres, J. (2012). Motivation through goal setting. *Journal of Economic Psychology*, *33*(6), 1223 - 1239. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167487012000967 doi: https://doi.org/10.1016/j.joep.2012.08.010

Goodman, S. F., & Turner, L. J. (2013). The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. *Journal of Labor Economics*, *31*(2), 409 - 420. Retrieved from http://ideas.repec.org/a/ucp/jlabec/doi10.1086-668676.html

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annu. Rev. Econ.*, *4*(1), 131–157.

Imberman, S. A. (2015). How effective are financial incentives for teachers? *IZA World of Labor*.

Imberman, S. A., & Lovenheim, M. F. (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics*, *97*(2), 364–386.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008, December). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education Review*, *27*(6), 615-631. Retrieved from http://ideas.repec.org/a/eee/ecoedu/v27y2008i6p615-631.html

Koch, A., & Nafziger, J. (2011). Self-regulation through goal setting*. *The Scandinavian*

Journal of Economics, *113*(1), 212-227. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9442.2010.01641.x doi: 10.1111/j.1467-9442.2010.01641.x

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

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

Ladd, H. F. (1999). The dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review*, *18*(1), 1–16.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy*, *110*(6), pp. 1286-1317. Retrieved from http://www.jstor.org/stable/10.1086/342810

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

Leigh, A. (2012). The economics and politics of teacher merit pay. *CESifo Economic Studies*, *59*(1), 1–33.

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

Loyalka, P. K., Sylvia, S., Liu, C., Chu, J., & Shi, Y. (in press). Pay by design: Teacher performance pay design and the distribution of student achievement. *Journal of Labor Economics*.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teacher's understanding offundamental mathematics in china and the united states*. Lawrence Erlbaum Associates, Incorporated.

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

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2017). *Inputs, incentives, and complementarities in primary education: Experimental evidence from tanzania.* (mimeo)

Metzler, J., & Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics*, *99*(2), 486–496.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. *Journal of Political Economy*, *119*(1), pp. 39-77. Retrieved from

http://www.jstor.org/stable/10.1086/659655

Neal, D. (2011). Chapter 6 - the design of performance pay in education. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 4, p. 495 - 550). Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780444534446000067 doi: https://doi.org/10.1016/B978-0-444-53444-6.00006-7

Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. *The Journal of Economic Education*, *44*(4), 339–352.

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

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101.

Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, *3*(1), 601–630.

OECD. (2017). *Teachers' salaries (indicator).* (data retrieved from https://data.oecd.org/eduresource/teachers-salaries.htm) doi: 10.1787/f689fb91-en

PRI. (2013). *Tanzanian teachers learning education doesn't pay.* Retrieved 13/09/2017, from https://www.pri.org/stories/2013-12-20/tanzanian-teachers-learning-education-doesnt-pay

Reuters. (2012). *Tanzanian teachers in strike over pay.* Retrieved 13/09/2017, from http://www.reuters.com/article/ozatp-tanzania-strike-20120730-idAFJOE86T05320120730

Schotter, A., & Weigelt, K. (1992). Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *The Quarterly Journal of Economics*, *107*(2), 511–539.

Uwezo. (2012). *Are our children learning? annual learning assessment report 2011* (Tech. Rep.). Author. Retrieved from http://www.twaweza.org/uploads/files/UwezoTZ2013forlaunch.pdf (Accessed on 05-12-2014)

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

Vigdor, J. (2008). *Teacher salary bonuses in north carolina.*

Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review*, *30*(3), 404 - 418. Retrieved from http://www.sciencedirect.com/science/article/pii/S0272775710001731 doi: http://dx.doi.org/10.1016/j.econedurev.2010.12.008

World Bank. (2011). *Service delivery indicators: Tanzania* (Tech. Rep.). The World Bank, Washington D.C.

World Bank. (2014). *Service delivery indicators: Tanzania* (Tech. Rep.). The World Bank, Washington D.C.

World Bank. (2017). *World development indicators.* (data retrieved from, https://data.worldbank.org/data-catalog/world-development-indicators)

World Bank. (2018a). *Systems approach for better education results (saber).* (data retrieved from, http://saber.worldbank.org/index.cfm?indx=8&pd=1&sub=1)

World Bank. (2018b). *World development report 2018: Learning to realize education's promise.* The World Bank. Retrieved from https://elibrary.worldbank.org/doi/abs/10.1596/978-1-4648-1096-1 doi: 10.1596/978-1-4648-1096-1

A Appendix

A.1 Randomization Details

From a previous RCT (KiuFunza I), we have the baseline data necessary to implement the pay for percentile incentive scheme (to split students into groups, and properly seed each contest) for 180 schools. There were two treatments and a control group in this experiment. The treatment was stratified by district (and we continue this practice in this experiment). In each district, there are seven schools in each of the previous treatments (seven schools in C1 and seven in C2) and four in the control group (C3).

We randomly assign schools from the previous treatment groups into the new treatments groups. 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[35]. 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, P4pctile, 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.

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

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

A.2 Additional Tables

A.2.1 Properly seeded contests

Table A.2: Effect on test scores (without grade 1)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Panel A: Non-incentivized								
Levels (α_1)	.052	.031	.012	.035	.11**	.13***	.11	.12***
	(.047)	(.053)	(.086)	(.049)	(.05)	(.046)	(.085)	(.046)
P4Pctile (α_2)	-.008	-.047	-.049	-.04	.11**	.071*	.19**	.12***
	(.044)	(.049)	(.076)	(.046)	(.045)	(.041)	(.081)	(.041)
N. of obs.	3,120	3,120	1,532	3,120	3,163	3,163	1,533	3,163
$\alpha_3 = \alpha_2 - \alpha_1$	-.06	-.078	-.061	-.076*	-.0026	-.055	.08	-.0066
p-value ($H_0 : \alpha_3 = 0$)	.18	.12	.43	.08	.96	.25	.29	.89
Panel B: Incentivized								
Levels (β_1)	.13***	.12**	.18***	.22***	.17***	.14**	.28***	.26***
	(.05)	(.054)	(.067)	(.072)	(.051)	(.055)	(.069)	(.072)
P4Pctile (β_2)	.079*	.034	.16***	.12*	.09**	.063	.23***	.15**
	(.045)	(.048)	(.058)	(.065)	(.045)	(.045)	(.055)	(.061)
N. of obs.	30,206	30,206	14,664	30,206	32,956	32,956	15,458	32,956
Gains-Levels (β_3) $= \beta_2 - \beta_1$	-.054	-.09	-.014	-.1	-.083*	-.073	-.047	-.11
p-value ($H_0 : \beta_3 = 0$)	0.26	0.10	0.83	0.16	0.097	0.19	0.53	0.12
Panel C: Incentivized – Non-incentivized								
$\beta_1 - \alpha_1$	.07	.08	.14	.16	.053	.0053	.15	.12
p-value($\beta_1 - \alpha_1 = 0$)	.16	.12	.12	.015	.27	.92	.14	.065
$\beta_2 - \alpha_2$	.081	.073	.2	.14	-.02	-.0089	.04	.023
p-value($\beta_2 - \alpha_2 = 0$)	.11	.14	.017	.026	.66	.85	.66	.7
$\beta_3 - \alpha_3$	.012	-.0066	.053	-.015	-.074	-.014	-.11	-.097
p-value($\beta_3 - \alpha_3 = 0$)	.82	.9	.54	.82	.13	.8	.28	.16

Results from estimating Equation 6 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. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.2 Balance in Teacher Turnover

Table A.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
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.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). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.3 Effects on Test Takers and Lee Bounds on the Incentivized Test

Table A.4: Number of test takers, incentivized test

	(1)	(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. Clustered standard errors, by school, in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: 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). Clustered standard errors, by school, in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.4 Additional Heterogeneity in Treatment Effects

Table A.6: Heterogeneity by student characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Math			Kiswahili			English		
	Male	Age	Test(Yr0)	Male	Age	Test(Yr0)	Male	Age	Test(Yr0)
Levels*Covariate (α_2)	-0.025	0.011	0.026	0.011	-0.024	0.023	-0.059	-0.0021	0.091
	(0.039)	(0.015)	(0.033)	(0.039)	(0.016)	(0.027)	(0.081)	(0.041)	(0.055)
P4Pctile*Covariate (α_1)	0.0095	0.0089	0.063**	0.0023	-0.0051	0.040	-0.048	0.032	0.066
	(0.042)	(0.016)	(0.027)	(0.039)	(0.016)	(0.026)	(0.082)	(0.042)	(0.057)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650	3,065	3,065	3,065
$\alpha_3 = \alpha_2 - \alpha_1$	.035	-.0024	.037	-.009	.019	.017	.011	.034	-.024
p-value ($H_0 : \alpha_3 = 0$)	.4	.88	.23	.82	.22	.52	.89	.35	.59

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

Table A.7: Heterogeneity by school characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Math			Kiswahili			English		
	Facilities	PTR	Fraction Weak	Facilities	PTR	Fraction Weak	Facilities	PTR	Fraction Weak
Levels*Covariate (α_2)	0.031	-0.00015	-0.16	0.033	-0.0019	-0.23	0.063	-0.0040*	-0.42
	(0.023)	(0.0015)	(0.18)	(0.031)	(0.0013)	(0.17)	(0.043)	(0.0022)	(0.26)
P4Pctile*Covariate (α_1)	-0.027	-0.0025**	-0.24	0.0024	-0.0021	-0.32**	0.072	-0.0026	-0.34
	(0.026)	(0.0012)	(0.15)	(0.032)	(0.0013)	(0.16)	(0.044)	(0.0024)	(0.30)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650	3,065	3,065	3,065
$\alpha_3 = \alpha_2 - \alpha_1$	-.057**	-.0024	-.079	-.031	-.00025	-.088	.0093	.0014	.075
p-value ($H_0 : \alpha_3 = 0$)	.023	.18	.62	.28	.88	.55	.82	.64	.78

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

A.2.5 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 A.8: 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. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.6 Cooperation

Table A.9: Teacher behavioral responses: cooperation

	Help from other teachers (Instances last month) (1)	Help from other teachers (last month>0) (2)	Help/advice from other teachers (very good/good) (3)	Help/advice from head teacher (very good/good) (4)
Levels (α_1)	-.32**	-.047	-.058*	-.025
	(.15)	(.036)	(.031)	(.031)
P4Pctile (α_2)	-.42**	-.046	-.0015	.026
	(.18)	(.034)	(.026)	(.026)
N. of obs.	1,991	1,991	1,998	1,940
Mean control	1.3	.4	.75	.78
$\alpha_3 = \alpha_2 - \alpha_1$	-.094	.0012	.057*	.05
p-value($\alpha_3 = 0$)	.5	.97	.081	.14

This table shows the effect of treatment on teacher reports of help and cooperation from other teachers: the number of times the teacher received help from other teachers (Column 1), whether the teacher received any help from other teachers (Column 2), whether the teacher rated the advice or help from the other teacher as very good or good (Column 3) and whether the teacher rated the advice or help from the head teacher as very good or good (Column 4). Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.7 Pass Rates

Table A.10: Pass rates using levels thresholds in Kiswahili

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

Table A.11: Pass rates using levels thresholds in math

	Counting (1)	Numbers (2)	Inequalities (3)	Addition (4)	Subtraction (5)	Multiplication (6)	Division (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. Clustered standard errors, by school, in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.12: Pass rates using levels thresholds in English

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

A.2.8 National Assessments

We test the effect of both interventions using data from national standardized tests. In particular, we look at the average performance of students enrolled in our schools on the Standard Four National Examination (SFNA) taken by students in grade 4 and 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.

While we have data for both the SFNA and the PSLE, we focus on the latter for several reasons. First, the PSLE is a high-stakes test for students: their progression to secondary school is related to the results of this test. Schools are ranked based on the results of these tests. On the other hand, the SFNA is low-stakes with no consequences for either school administrators or students. Second, we collected higher quality data for grade 4 students, since the data are at the student level, we developed the test and thus are able

to verify that it is not susceptible to bottom- or top-coding, nor is it prone to student selection issues (unlike the PSLE and the SFNA). Finally, the SNFA data are known for being low-quality () (ISAAC SOME CITATION HERE?).

Overall, we do not find any impact of our treatment on PSLE test scores, pass rates, or the number of test takers (see Table A.13).[36]

[36]We do find that test scores decrease on the SNFA examination in 2015. However, this is not consistent with our higher-quality data on grade 4 students (see Table 4). 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.

Table A.13: 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

Clustered standard errors, by school, in parentheses.

A.2.9 "Wasted" Money

To estimate how much money was "wasted" in the levels scheme for student's passing certain thresholds regardless of teachers' effort, we do the following.

- First, we estimate the following model: $Y_i = \beta X_i + \varepsilon_i$, where Y_i is whether a student passed (acquired a skill) or not, X_i is a set of student controls (including region, grade, and school characteristics). The sample is restricted to the control group.
- We then estimate the probability of passing (acquiring each skill) in the treatment group using this model. This assumes that in the absence of the treatment, the treatment group would behave like the control group.
- We then estimate the average *estimated* pass rate and compare it to the *actual* pass rate. Specifically, we estimate $\max 0, \frac{\overline{\hat{Y}_i}}{\overline{Y_i}}$. This is the proportion of the money paid for results that were not related to additional effort exerted by the teacher. The results are shown below.

Table A.14: "Wasted" money

	Kiswahili (1)	Math (2)	English (3)
Year 1	7%	4%	0%
Year 2	5%	10%	0%

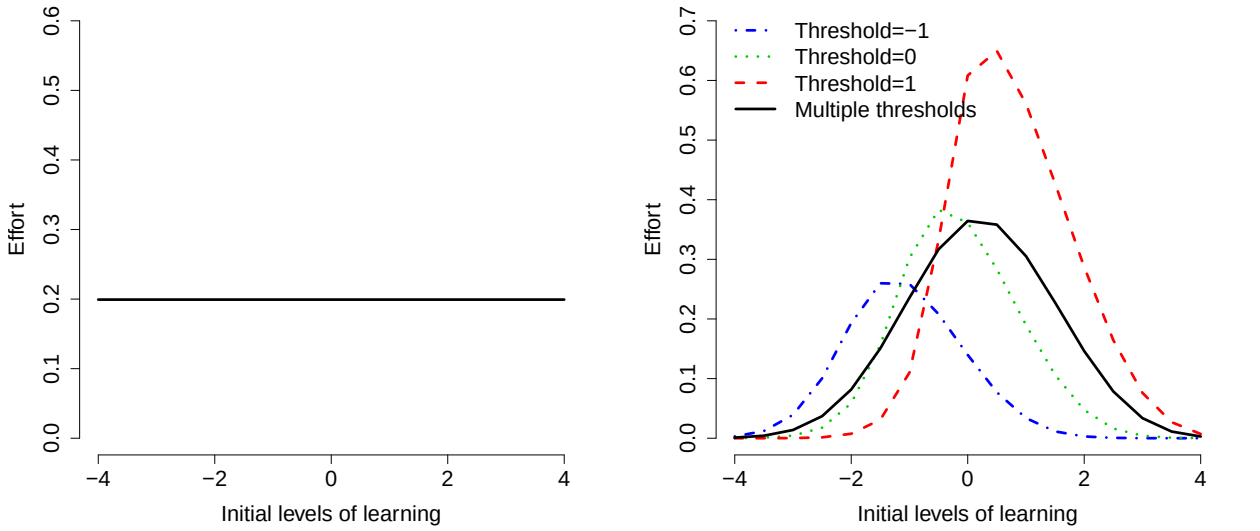
A.2.10 Robustness of Simulation Results

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 3.4. 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.[37] Figures A.1 and A.2 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 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 A.1: Incentive design and optimal effort with constant productivity of teacher effort

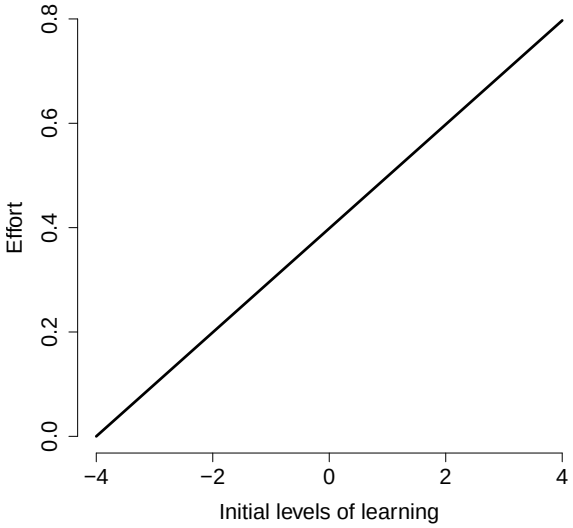


(a) P4Pctile - γ constant across initial levels of learning

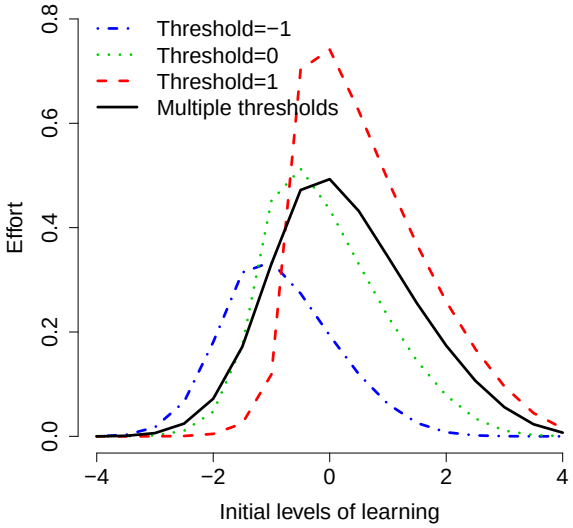
(b) Levels - γ constant across initial levels of learning

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

Figure A.2: 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