Left Behind by Optimal Design: The Challenge of Designing Effective Teacher Performance Pay Programs[*]

Isaac Mbiti[†] Mauricio Romero[‡] Youdi Schipper[§]

June 19, 2018

Abstract

A growing body of evidence suggests that 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 randomly assigned teacher performance pay systems: a pay for percentile system, which is more complex, but can (under certain conditions) induce optimal effort among teachers and a simple system that rewards teachers based on student proficiency levels. We find that both systems improve student test scores. However, despite the theoretical advantages of the pay for percentile system, the proficiency system is at least as effective in raising student learning. Moreover, we find suggestive evidence that the pay for percentile system favors students from the top of the 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 especially thank Karthik Muralidharan for his involvement as a collaborator in the early stages of this project and for subsequent discussions. We are grateful to Bobby Pakzad-Hurson, John Friedman, Joseph Mmbando, Molly Lipscomb, Omer Ozak, Jay Shimshack and seminar participants at UC San Diego and Universidad del Rosario for comments. Erin Litzow and Jessica Mahoney provided excellent research assistance through Innovations for Poverty Action. We are also grateful to EDI Tanzania for their excellent 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 started to shift from increasing primary school enrollment to promoting policies that improve learning. This shift has been driven in part by the growing body of 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; Mbiti, 2016). By strengthening the links between teacher remuneration and performance indicators, such as objectively measured student learning outcomes, 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. For instance, the share of US school districts featuring teacher performance pay programs has increased by more than 40% from 2004 to 2012 (Imberman, 2015). Moreover, there is a global trend towards increased adoption of such programs 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 depend on key elements of the incentive program design (Neal & Schanzenbach, 2010; Neal, 2011; Mbiti et al., 2017; Bruns & Luque, 2015; Loyalka, Sylvia, Liu, Chu, & Shi, in press).[2] Incentive schemes based on proficiency thresholds are commonly used (for example by 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 are theoretically less effective than more complex alternatives and can encourage teachers to focus on students who are close to the proficiency threshold (Neal & Schanzenbach, 2010; Neal, 2011). In contrast, more

[1] 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] These elements include: whose performance is incentivized (individual or collective); what is the incentive metric (test scores, or input measures like attendance or preparation); how large is the expected incentive payment; and the mapping between the performance metric and the teacher incentive payment. This last design element is the focus of our study.

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, 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 (Goodman & Turner, 2013; Fryer, 2013; Neal, 2011). Furthermore, under certain conditions (or specifications of the education production function), these complex systems can potentially increase inequality in learning by encouraging teachers to focus on students with higher baseline test-scores. Given the challenge of designing effective incentive systems, 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.

In this paper, we compare the effectiveness of a more complex teacher incentive scheme to a simpler system using a randomized experiment in a set of 180 Tanzanian public primary schools. In particular, we compare the effectiveness of the pay for percentile scheme proposed by Barlevy and Neal (2012) to a simpler design with multiple proficiency thresholds ('levels'). We compare the student learning outcomes in both treatment groups to our control group. Both types of incentive programs rewarded teachers in math, Kiswahili, and English in first, second, and third grade. In addition, the per-student bonus budget was equalized 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 percent of the annual net salary. Following Neal (2013), we evaluate the incentive programs using data from both the 'high-stakes' test that is administered to all students in order to determine teacher bonuses, and the 'low-stakes' test that is administered to a sample of students for research purposes. Both types of tests are collected in control schools, although the results of the 'high-stakes' test do not trigger any payments in these schools.

In the 60 schools assigned to the pay for percentile arm, students are 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 are re-tested and ranked within their assigned group. Teachers are then rewarded in proportion to their students' ranks within each group. By effectively 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 the pay for percentile can induce socially optimal levels of effort among teachers. In addition, they also show that the system can lead to equitable learning gains under certain specifications of the education production function. However, the system can be difficult for teachers to comprehend, and challenging to implement, as school systems would need to create and manage databases that track student learning over time.

In the 60 schools assigned to receive incentives based on proficiency targets, teachers earn bonuses based on their students' mastery of several grade-specific skills that are outlined in the national curriculum. As teacher incentive programs that use single proficiency thresholds typically encourage teachers to focus on students close to the passing threshold, we included several passing thresholds to encourage teachers to broaden their focus. In our design, the skill thresholds range from very basic skills to more complex skills, which allow teachers to earn rewards from a wide range of students. As reward payments for each skill are 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, this system is easier to understand and simpler to implement, as it only requires that school systems administer one test at the end of the year. 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.

Despite the theoretical advantages of the pay for percentile system, 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 low-stakes test, we find that 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 $.056\sigma$ (p-value .11) 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 .02) in pay for percentile schools, although these estimates are not statistically distinguishable. The results using the high-stakes 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 high-stakes exam increased by $.14\sigma$ (p-value < 0.01) under the proficiency levels system compared to a $.093\sigma$ (p-value .02) increase 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 high-stakes data only).

In addition, Loyalka et al. (in press) find that pay for percentile teacher incentives led to equitable learning gains across the distribution of students.[3] In contrast, we find suggestive evidence that better prepared students benefited more under the pay for percentile treatment. However, we find that the learning gains under the levels systems 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.[4] 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.[5] 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), our results suggest that teacher performance pay programs need to be carefully designed to account for equity concerns.[6]

We use our comprehensive set of survey data collected from school administrators, teachers, and students to shed light on theoretically relevant mechanisms. Consistent

[3] Muralidharan and Sundararaman (2011) found equitable learning gains from a value-added teacher incentive design.

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

[5] 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.

[6] 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 workers because managers focused their efforts on high ability workers.

with our results on test scores, we find that teacher effort was generally higher under the levels systems compared to the pay for percentile system. For instance, teachers were more likely to be off task in pay for percentile schools relative to levels schools. Given the well-documented concerns about teacher misunderstandings 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 driving factor. Although previous studies have shown that women exert less effort in tournaments (Niederle & Vesterlund, 2007, 2011), we do not find any heterogeneity by teacher gender. We also explore the heterogeneity in treatment effects by teacher ability and find that more able teachers were more responsive to the levels design, while there was no differential pattern by teacher ability in the pay for percentile system. We also find that teachers in the pay for percentile design expected to earn lower bonuses, perhaps due to the increased uncertainty and complexity of the pay for percentile system relative to the levels system. In addition, teachers in the levels system were better able to articulate clear and specific targets for their students on the high stakes exams, perhaps due to the clearer reward structure.

Our results highlight the challenges of designing teacher incentives that are both effective and equitable. Program design features such as the mapping of test-scores to rewards, the size of the bonus, the potential for free-riding, and the measure of student learning used to reward teachers play important roles in determining the effectiveness of the program (Neal, 2011; Loyalka et al., in press; Ganimian & Murnane, 2016; Bruns & Luque, 2015). These differences are the primary drivers of the heterogeneity in the effectiveness of teacher performance pay programs (Imberman & Lovenheim, 2015; Neal, 2011; Ganimian & Murnane, 2016). For instance, Glewwe, Ilias, and Kremer (2010); Lavy (2002, 2009); Muralidharan and Sundararaman (2011); Balch and Springer (2015); Loyalka et al. (in press) find that teacher performance pay increases student test scores. However, another set of studies find that these programs have limited effects on student learning (Fryer, 2013; Barrera-Osorio & Raju, in press; Goldhaber & Walch, 2012; Goodman & Turner, 2013; Springer et al., 2011).

Our study contributes to a growing literature on the potential of teacher incentives to improve learning outcomes in developing countries. By comparing the pay for percentile system with a simpler budget equivalent proficiency system, our study provides some of the first empirical evidence of the effectiveness of pay for percentile in sub-Saharan Africa, benchmarked against a simpler system.[7] This comparison allows us to

[7] 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

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.

As our results show, under certain specifications of the education production function, 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 function specifications. This is consistent with a large body of literature from contract theory, such as Carroll (2015) and Carroll and Meng (2016), who show that simpler incentive schemes are more robust mechanisms to resolve principal agent problems when there is uncertainty about the specification of the production function.

2 Experimental Design

2.1 Context

Tanzania allocates about a fifth of overall government spending (roughly 3.5 percent of GDP) on education (World Bank, 2017). Much of this spending has been devoted to promoting educational access. As a consequence, net enrollment rates in primary school have 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, only 14 percent of schools had access to electricty, 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. Moreover, approximately 5 pupils shared a single mathematics textbook, while 2.5 pupils shared a reading textbook in 2013 (World Bank, 2017). Consequently, student learning levels are quite low. In 2012, data from nationwide assessments show 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 the 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 can promote learning (Mbiti, 2016). The lack of accountability is further reflected in

student learning, except in schools with textbooks. They do not compare their pay for percentile system to other incentive designs teacher absence 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 was 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 (roughly US$300), or roughly 4.5 times GDP per capita (World Bank, 2017).[8] 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).[9]

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. The intervention was part of a series of projects launched under a broader program umbrella named KiuFunza ('Thirst for learning' in Kiswahili).[10] 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 prize 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. To ensure school support for the incentive scheme, we also offered a performance bonus to Head Teachers. This bonus equals 20 percent of the combined bonus of all incentivized teachers in his or her school. The head teacher bonus was communicated transparently but did not affect the bonus calculations of the teachers, since the teacher bonus budgets were communicated excluding the head teacher budget.

[8] 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).

[9] In recent years, other teacher strikes have occurred in South Africa, Kenya, Guinea, Malawi, Swaziland, Uganda, Benin and Ghana.

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

Within each intervention arm, Twaweza distributed information describing the program to schools and the communities via public meetings in early 2015 and 2016. The implementation teams also conducted additional 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, are tested in Kiswahili, English and, math. As this test was used to determine teacher incentive payments, it is 'high-stakes' (from the teacher's perspective). Our 'low stakes' 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.[11]

2.2.1 Proficiency thresholds (levels) design

Proficiency based systems are easier for teachers to understand and provide more actionable targets. Consequently, such systems are likely to increase motivation among teachers and head teachers, but 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. In addition, 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, rather than a single threshold, ranging from very basic skills to more advanced skills on the curriculum. This design allows teachers to earn bonuses for helping a broader set of students, including students with lower baseline test scores and those with 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 between skills, while the per pass bonus paid ex-post equals the skill budget divided by the number of students passing the skill. Consequently, harder 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 skills.

Table 1 shows the skills to be tested in each grade-subject combination. The total budget is split across grades in proportion to the number of students enrolled in each

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

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 of 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, 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.

For each skill, teachers earn more money as more students in their class score higher than the passing threshold. The payment for the skill is higher if fewer students are above the threshold. In other words, the reward is higher for teachers if students master a 'difficult' skill, which is defined by the overall passing rate of each skill.

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
		Multiply
	Grade 3	
		Addition
Story Comprehension	Story Comprehension	Subtraction
		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 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 a 'unknown' ability group. 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 with a similar baseline level of knowledge.

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 (into 100 ranks) within each group according to their endline test score a_i^t, and within each group we gave teachers points proportional to the rank of their students. A teacher would receive 99 points for a student in the top 1% of group k, and no 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 is proportional to the number of students in each grade and is 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, the total prize money was $70,820 and total enrollment was 22,296 in pay of percentile schools. Therefore, the payment per 'rank' was $0.0178. For a student in the top 1%, teachers earned $1.77 and for a student in the top 50% teachers earned $0.89.

Although this design can deliver socially optimal levels of effort, it can be challenging to implement at scale, particularly in settings with weak administrative capacity, such as Tanzania. For instance, maintaining the child-level databases of learning required to calculate teacher value-added, and ensuring the integrity of the testing system are non-trivial administrative challenges. 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 pay-off 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

As Kiswahili is the official language of instruction in Tanzania, English is taught as a second language in primary schools. As English is rarely spoken outside of the classroom, 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 Grades 1 and 2 in 2015 to allow teachers to focus on numeracy and literacy in Kiswahili in those grades. English would still be taught in Grade 3, under a revised curriculum. However, there was little guidance from the Ministry of Education on how to transition to the new curriculum. Hence, there was substantial variation in how the curriculum changes were actually implemented by schools. Some schools stopped teaching English in 2015, while others stopped in 2016. In addition, there was no official guidance on whether to use Grade 1 English materials in Grade 3, as there were no new books issued to reflect the curriculum changes. As a result, we 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 is unclear how to interpret the results for English in Grade 1 and 2 in 2015, and for Grade 3 in both years.

3 Theoretical Framework

We present a set of simple models to clarify the potential behavioral responses by teachers and schools in our interventions. We first characterize equilibrium effort levels of teachers in both incentive systems, and then impose some additional structure 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, students vary in their initial level of learning and are indexed by l. Further, each classroom of students is taught by a single teacher, indexed by j. We assume student learning (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 with learning level l taught by teacher j, and $a_{j(t-1)}^l$ is the student's baseline level of learning. 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.[12] 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.

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:

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

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

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

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

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 N is large and teachers ignore their own effect on the overall pass rates, the first order conditions for the teacher's maximization problem becomes:

$$\sum_t \gamma^l h^l(T_t - a_{j(t-1)}^l - \gamma^l e_j^l) \frac{\Pi_t}{\sum_l \sum_k C_n^l P\left(v_k^l > T_t - a_{k(t-1)}^l - \gamma^l e_k^l\right)} = c_l'(e_j^l) \tag{4}$$

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

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

for all l.

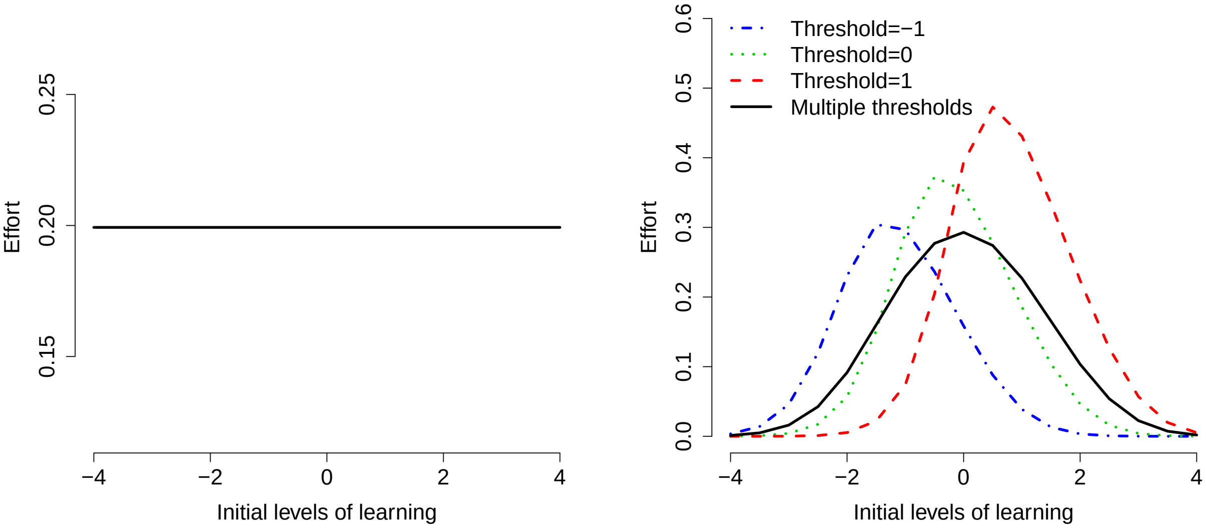
3.4 A Comparison of Optimal Teacher Effort

We impose additional structure on our conceptual framework and numerical methods to generate qualitative predictions about the level of teacher effort across the baseline test-score distribution of students. 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 importance 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, and each teacher has 17 students in his or her class. Within each class, we let student baseline learning levels to be uniformly distributed from -4 to 4, in 0.5 intervals (i.e., there is one student with each baseline learning level in each class). We set the prize per student in both schemes at \$1. Therefore, in the pay for percentile scheme the prize per contest won is $\$\frac{2}{99}$ (see Section 3.2) and in the levels the total prize is such that there is \$1 per student. In the multiple threshold scenario the prize is held constant and split evenly across all three thresholds. For simplicity, we assume that there are three proficiency thresholds. We first compute the optimal teacher response assuming a single proficiency threshold and then vary the threshold value from -1 to 1. We then compute the multiple threshold case by considering all three cases simultaneously.

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 to all 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 would not exert much effort to students in the tails of the initial learning distribution (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



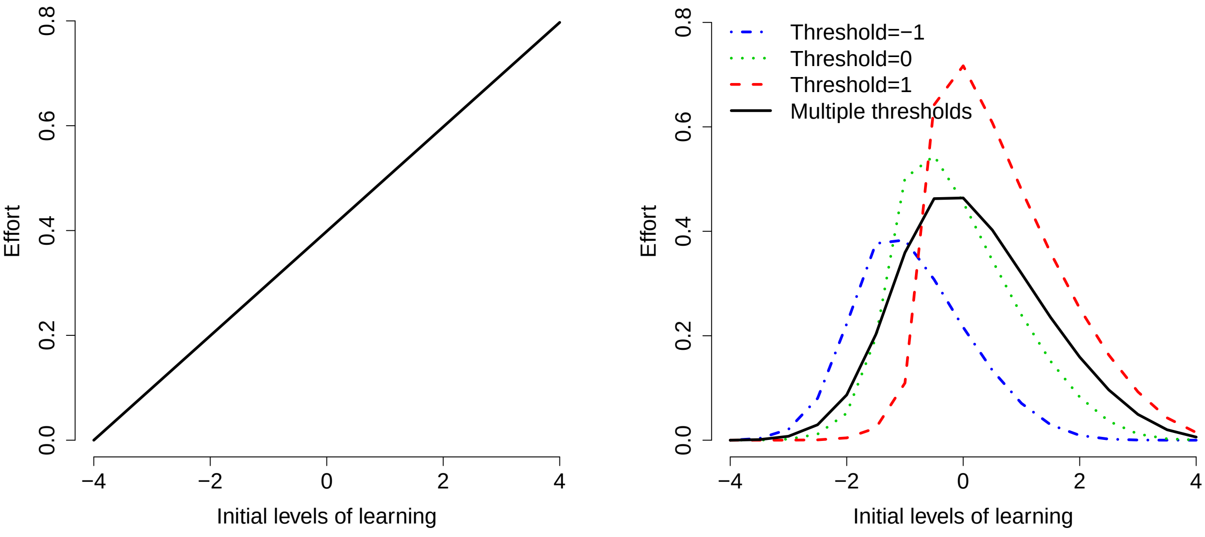
(a) Gains - γ constant across initial levels of learning

(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 the 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) Gains - γ increases with initial levels of learning

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

4 Data and Empirical Specification

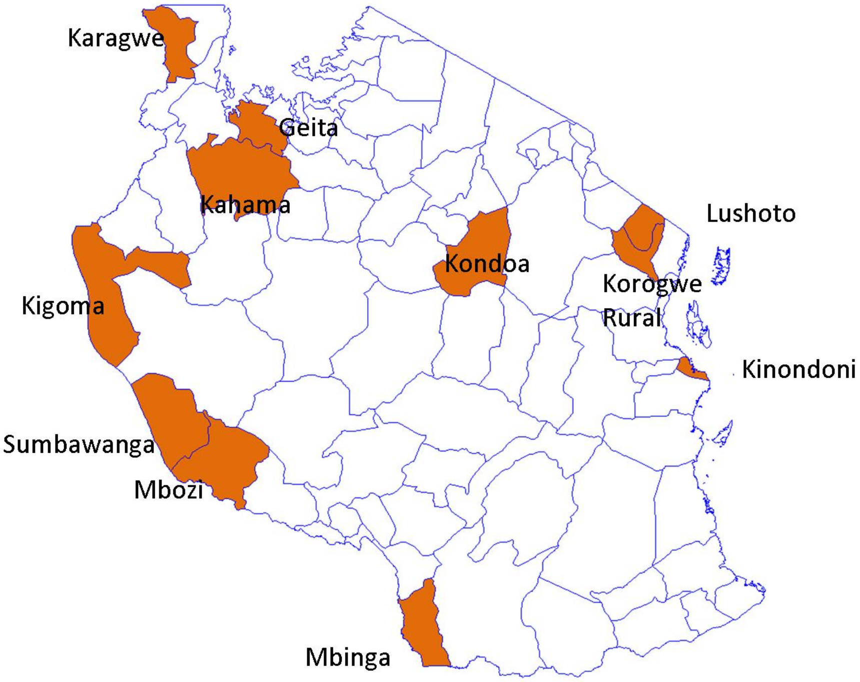
4.1 Sample Selection

The teacher incentive programs were evaluated using a randomized design. First, ten districts were randomly selected (see Figure 3).[13] 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. Within each district, we randomly allocated schools to one of our

[13] The program was administered in 11 districts, as one district was included non-randomly by Twaweza for piloting and training. We do not survey schools in this district.

three experimental groups. Thus, in each district, 6 schools were assigned to the 'levels' treatment, 6 schools to the pay for percentile treatment, and 6 schools served as controls. In total, there are 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.

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.

4.2 Data and Balance

Over the two-year evaluation our survey teams visited each school at the beginning and end of the 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 each teacher in our evaluation, including personal characteristics such as 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 'low-stakes' student tests with the results from the 'high-stakes' tests which are used to determine teacher bonus payments, and are conducted in all schools including control schools.

Although the content (subject order, question type, phrasing, difficulty level) is consistent across low- and high-stakes tests, there are a number of important differences in the test administration. The low-stakes test took longer (40 minutes) than the high-stakes test (15 minutes). The low-stakes test had more questions in each section (Kiswahili, English and math) to avoid bottom- and top-coding, and also included an 'other subject' module at the end to test spillover effects. The testing environment was different. Low-stakes tests were administered by an enumerator, who took children out of the classroom and tested them one by one during a regular school day. In the high-stakes test, all students in Grades 1-3 were tested on an agreed test day. As most schools used the high-stakes test as the end of year test, students in higher grades were often given a day off, while Twaweza test teams administered the one-on-one tests in designated classrooms. 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. Each student listed at baseline received an individual pre-printed test form. One test out of ten test sets was randomly assigned to each student for additional test security. Tests were handled, administered, and electronically scored by Twaweza teams without any interference from teachers.

Most students, school, teachers, 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 , 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.

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				
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)	
Poverty index (PCA)	0.01	-0.08	0.01	0.42
	(1.99)	(1.94)	(1.98)	
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)	
Experience (Yrs)	13.97	12.91	13.54	0.11
	(11.93)	(11.47)	(11.14)	
Private school experience	0.03	0.01	0.03	0.05*
	(0.17)	(0.11)	(0.17)	
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 parenthesis) 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). Randomization was stratified by district, previous treatment arm, and quality strata. The quality strata variable for schools was created using principal component analysis on students' test scores. Schools were categorized into one of two strata depending on whether they were above or below the median for the first principal component. 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 from 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 from a Principal Component Analysis of indicator variables for: Outer wall, staff room, playground, library, and kitchen. Standard errors are clustered

4.3 Empirical Specification

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

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

where Z_{isdt} is the test score of student i in school s in district d at time t. $Levels$ 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 teachers age, average teacher experience, average teacher qualifications, and the fraction of female teachers.

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. We also balanced the timing of our survey activities, including the low-stakes tests, on a weekly basis across treatment arms. Hence, our results are not driven by imbalanced survey timing.

We examine the impacts of the incentives using both the low stakes (survey) and high stakes (intervention) testing data. However, given the limited set of student characteristics in the high stakes data, this analysis includes fewer student level controls.

We also use a similar specification to examine teacher behavioral responses. We further explore heterogeneity in treatment effects by interacting our treatment indicators with a variety of school, teacher, and student characteristics. These include student gender and age, teacher content knowledge, and school facilities and pupil-teacher ratios. We also explore the heterogeneity in treatment effects by baseline student ability to examine if teachers focus their efforts on a particular type of student. Specifically, we assign students into quintiles based on their baseline test scores, and then estimate the difference between the treatment and the control group within each quintile.

5 Results

5.1 Test Scores

Table 3 shows the impact of the incentives on student learning in math and Kiswahili using data from the low-stakes test (Panel A), as well as the high-stakes test (Panel B).

In the first year, both incentive schemes resulted in small learning gains on the low-stakes test (see Panel A in Table 3). However, the treatment effects of the levels incentive scheme were consistently larger than those in the pay for percentile system (Panel A, Columns 1 and 2). In particular, the effect in Kiswahili was larger by .08σ (p-value .077). In the second year of the program, we find that the estimated treatment effects on the low-stakes test are generally larger than the first year estimates. Math test scores improved by .067σ (p-value .09) in the levels system, and .07σ (p-value .056) in the pay for percentile system. Kiswahili test scores improved by .11σ (p-value < 0.01) under the levels system, but only by .056σ (p-value .11) under the pay for percentile system. Finally, English test scores improved by .11σ (p-value .19)) in the levels system, but improved by .19σ (p-value .02) 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 the high-stakes 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 the high-stakes exams (Panel B). Generally, the estimated treatment effects are larger compared to those estimated using the low-stakes data (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 high-stakes data. For example, the effect of the levels scheme is larger for Kiswahili by .11σ (p-value .026) in the first year, and by .093σ (p-value .045) in the second year.

The larger treatment effects found in the high-stakes data are likely driven by test-taking effort, where teachers have 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, the administration of the high stakes test was tightly controlled, and conducted by our implementation team. This mitigates any concerns about outright cheating.[14] Assuming that all the differences between our high stakes and low stakes 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).

Given the reward structure, teachers in both treatment arms would be motivated to

[14] Details of the Twaweza test data integrity process are available on request.

ensure that their students took the high-stakes exam. In the second year of the study, teachers in the levels schools were able to increase student participation in the high-stakes exam by 5 percentage points. Their counterparts in pay for percentile schools increased participation by 3 percentage points (see Table A.3). Following Lee (2009), we compute bounds on the treatment effects by trimming the excess test takers from the left and right tails of the high stakes test distribution respectively. 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.4).

Table 3: Effect on test scores

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	English	Math	Kiswahili	English
Panel A: Low-stakes						
Levels (α_1)	.038	.044	.014	.067*	.11***	.11
	(.047)	(.047)	(.086)	(.039)	(.039)	(.085)
P4Pctile (α_2)	-.017	-.035	-.049	.07*	.056	.19**
	(.04)	(.039)	(.076)	(.037)	(.035)	(.081)
N. of obs.	4,781	4,781	1,532	4,869	4,869	1,533
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.055	-.08*	-.063	.003	-.057	.079
p-value ($H_0 : \alpha_3 = 0$)	.21	.077	.41	.95	.16	.3
Panel B: High-stakes						
Levels (β_1)	.11**	.13***	.18***	.14***	.18***	.28***
	(.047)	(.048)	(.067)	(.045)	(.046)	(.069)
P4Pctile (β_2)	.066*	.017	.16***	.093**	.085*	.23***
	(.039)	(.043)	(.058)	(.04)	(.045)	(.055)
N. of obs.	48,077	48,077	14,664	59,680	59,680	15,458
Gains-Levels (β_3) $= \beta_2 - \beta_1$	-.047	-.11**	-.014	-.044	-.093**	-.047
p-value ($H_0 : \beta_3 = 0$)	0.30	0.026	0.83	0.31	0.045	0.53
Panel C: High-stakes – Low-stakes						
$\beta_1 - \alpha_1$	.065	.075	.14	.063	.056	.15
p-value($\beta_1 - \alpha_1 = 0$)	.13	.097	.12	.11	.2	.14
$\beta_2 - \alpha_2$	.078	.046	.2	.021	.025	.041
p-value($\beta_2 - \alpha_2 = 0$)	.072	.29	.017	.6	.55	.64
$\beta_3 - \alpha_3$	.012	-.029	.056	-.042	-.031	-.11
p-value($\beta_3 - \alpha_3 = 0$)	.78	.53	.52	.3	.51	.28

Results from estimating Equation 6 for different subjects at both follow-ups. Panel A uses data from the low-stakes exam 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 is the school to different facilities, and an indicator for whether the school is single shift or not). Panel B uses data from the high-stakes exam 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 is the school 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 parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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. On the one hand, schools may shift resources such as textbook purchases from higher grades to Grades 1, 2, and 3. Additionally, 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). As third graders in the first year of our program transitioned to the fourth grade in the second year of the program, we can use the fourth grade results in the second year (Table 4, Panel A, Columns 3 and 4) to explore the persistence of any learning gains produced by the incentive programs. Although the point estimates are mostly positive, they are not statistically significant.[15] This is consistent with learning gains from both incentive programs fading out over time. 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).

[15] The standard errors are larger than those in Panel A of Table 3 due to smaller sample sizes.

Table 4: Spillovers to other grades and subjects

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Grade 4						
	Year 1			Year 2		
	Math	Kiswahili	English	Math	Kiswahili	English
Levels (α_1)	.13**	.044	.17**	.061	.041	.082
	(.062)	(.05)	(.084)	(.063)	(.065)	(.069)
P4Pctile (α_2)	-.03	-.032	.032	-.0054	.026	.058
	(.054)	(.054)	(.077)	(.06)	(.061)	(.063)
N. of obs.	1,513	1,513	1,513	1,482	1,482	1,482
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.16**	-.077	-.14*	-.067	-.014	-.025
p-value ($H_0 : \alpha_3 = 0$)	.011	.13	.077	.25	.82	.72
Panel B: Science (Grades 1-3)						
	Year 1	Year 2				
Levels (α_1)	.069	.083				
	(.063)	(.06)				
P4Pctile (α_2)	-.005	.079				
	(.05)	(.057)				
N. of obs.	4,781	4,869				
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.074	-.0044				
p-value ($H_0 : \alpha_3 = 0$)	.24	.94				

Results from estimating Equation 6 for grade 4 students (Panel B) and for grade 3 students in science (Panel B). 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 is the school to different facilities, and an indicator for whether the school is single shift or not). Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.3 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 is measured by our survey team and is collected shortly after our team arrives at a school in the morning. Overall, we do not find any differences in this dimension of teacher effort across our treatments (see Table 5, Panel A). We also examine student reports about teacher effort such as assigning homework and providing extra help. In the first year of the program, students report receiving more help from teachers in the levels systems relative to the pay for percentile system. This difference is statistically significant at the 7 percent level (see Table 5, Panel B, Column 1). Students also report receiving more homework from teachers in the levels systems relative to the pay for percentile system, although the difference is not statistically significant (Panel B, Column 2). However, in the second year, students report receiving similar levels of extra help and homework assignments from teachers in both incentive systems; the point estimate on receiving extra help from pay for percentile teachers is significant (Panel B, Column 3).

Table 5: Teacher Presence and Teaching Strategies

	(1)	(2)	(3)	(4)
Panel A: Spot-checks				
	Year 1		Year 2	
	In school	In classroom	In school	In classroom
Levels (α_1)	0.012	0.0061	-0.025	0.025
	(0.053)	(0.057)	(0.050)	(0.053)
P4Pctile (α_2)	-0.012	-0.023	-0.0050	0.023
	(0.044)	(0.050)	(0.044)	(0.044)
N. of obs.	180	180	180	180
Mean control	.71	.32	.67	.37
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.024	-.029	.02	-.0021
p-value ($H_0 : \alpha_3 = 0$)	.65	.6	.71	.97
Panel B: Student reports				
	Year 1		Year 2	
	Extra help	Homework	Extra help	Homework
Levels (α_1)	0.011	0.033	0.0052	0.0029
	(0.018)	(0.024)	(0.0096)	(0.018)
P4Pctile (α_2)	-0.022	-0.0055	0.016*	-0.023
	(0.017)	(0.024)	(0.0097)	(0.019)
N. of obs.	9,006	9,006	9,557	9,557
Mean control	.12	.1	.018	.093
Gains-Levels $\alpha_3 = \alpha_2 - \alpha_1$	-.034*	-.038	.011	-.026
p-value ($H_0 : \alpha_3 = 0$)	.073	.16	.29	.24

Panel A presents teacher-level data on teacher absenteeism (Columns 1 and 3), and time-on-task (Columns 2 and 4). Panel B presents student-level data on teacher effort (as reported by students) on extra help (Columns 1 and 3) and homework (Columns 2 and 4). Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As in class observations are typically affected by Hawthorne effects, our survey teams collected data on teacher behavior standing outside the classroom for a few minutes, before teachers noticed they were being observed. Although these reports are not as detailed as within-classroom observation protocols, they are arguably better able to capture broad measures of typical teacher behavior.[16] 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 the likelihood that we observed teachers 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 student) 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 irrelevant 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 more precise 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 a few 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 parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[16] Schools in Tanzania have open layouts where classrooms are built in blocks, with 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).[17] 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. 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.

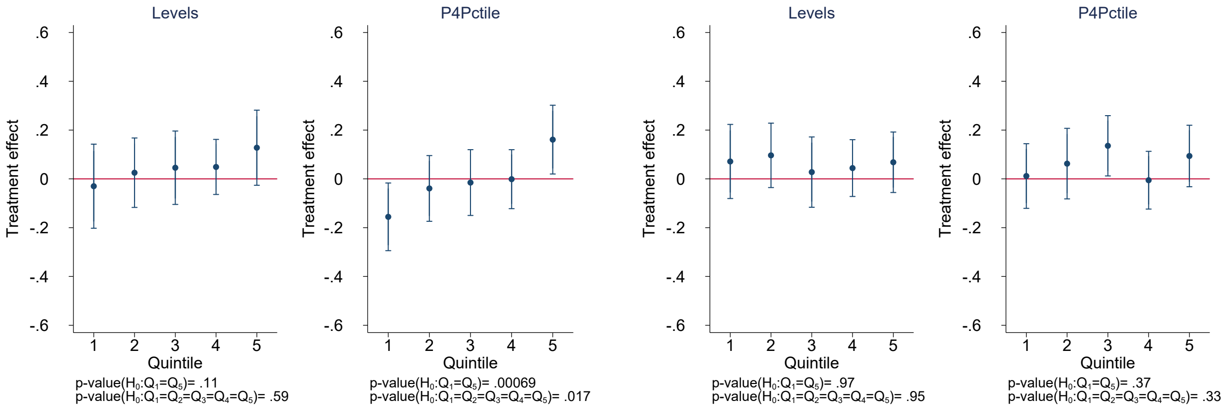
In the first year of the program, both math and Kiswahili teachers in the pay for percentile system (labeled "P4Pctile") focused their attention on their best students, whereas teachers in the levels system (labeled "levels") focused on the top half of their class (Figures 4a and 5a). English teachers under both systems seem 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 seem 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. The results in English are less informative, given the changes in the curriculum, and the general difficulty of teaching the language in Tanzania.[18]

[17] 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 those characteristics (see Tables A.5 and A.6 for details).

[18] Previous research that examined a simple single proficiency incentive scheme for teachers found that teachers in Kiswahili and Math focused on students in the middle of the distribution, while teachers in English focused on the top students (Mbiti et al., 2017).

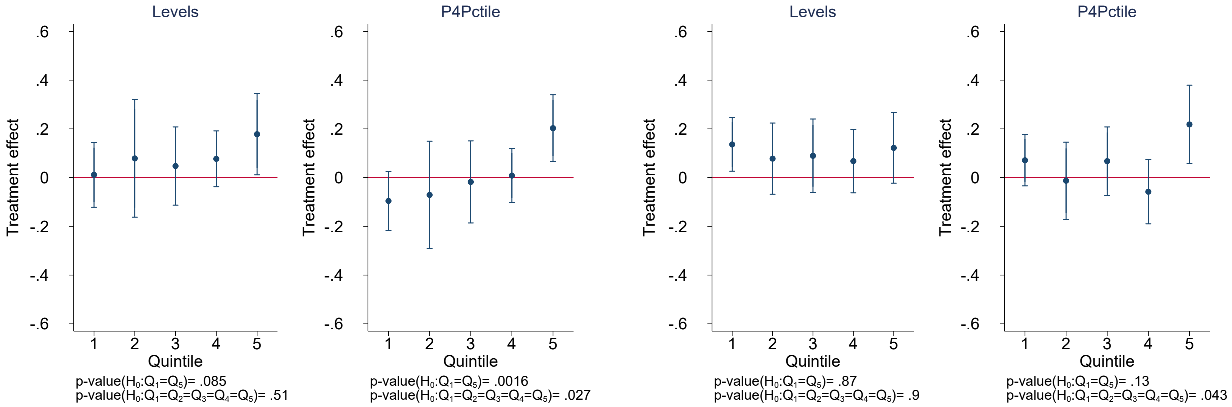
Figure 4: Math



(a) Year 1

(b) Year 2

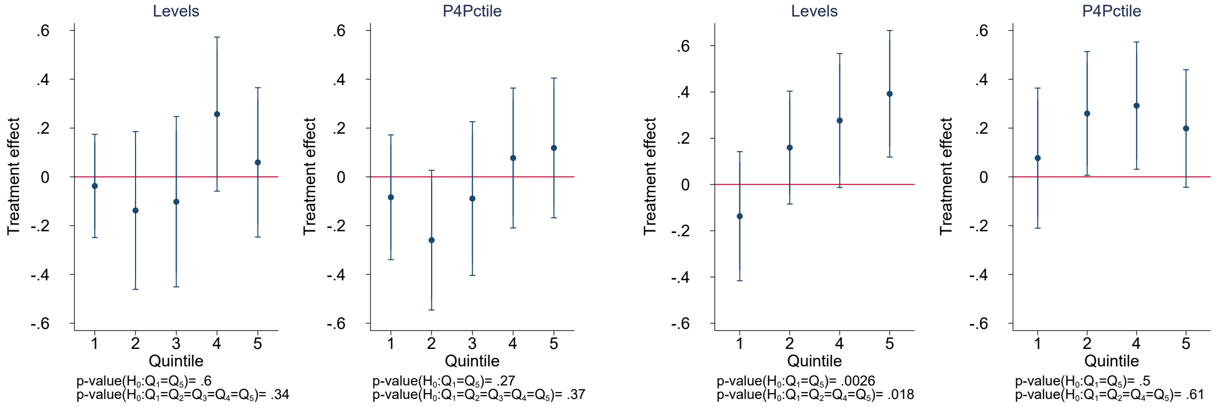
Figure 5: Kiswahili



(a) Year 1

(b) Year 2

Figure 6: English



(a) Year 1

(b) Year 2

5.5 Mechanisms

Given the limited empirical evidence on pay for percentile schemes, we rely on the large body of theoretical and empirical evidence on rank-order tournaments to guide our analysis on potential mechanisms that drive the differences in behavior and outcomes between the two types of incentives. We focus particularly on differences in the incentive structure 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. In addition, we explore the importance of a number of theoretically relevant teacher characteristics, such as teacher ability, as well as differences in expectations, cooperation, and goal-setting.

5.5.1 Heterogeneity by Teacher Characteristics

Table 7 explores heterogeneity by teacher characteristics including gender, age, content knowledge, and measures of teacher effectiveness across all three subjects. There is a growing body of evidence that 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 pay for percentile scheme is more competitive than the levels scheme, we do not find that women are less responsive to its competitive pressure (Column 1). We also do not find any heterogeneous effects by teacher age, which proxies for experience.

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 also 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. We use three different measures of teacher ability to explore the heterogeneity in treatment effects. Our measures include an index of teacher content knowledge, which is scaled by an IRT model, an index of head teacher evaluations of individual teacher effectiveness, and an index of teachers' perceptions of their own self-efficacy.[19]

Although studies such as Metzler and Woessmann (2012) have shown that teacher content knowledge is predictive of student learning outcomes, we do not find any significant heterogeneity in our treatment effects by teacher content knowledge (Column 3). We further find that 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. We also examine heterogeneous effects by teacher beliefs about their individual efficacy in Column 5. We generally find that teachers who believed they were more capable responded more to both incentives (Column 5, Panel A and Panel B). Although 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.

[19] 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 on seven dimension including the ability to ensure that students learn, and classroom management skills. 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'.

Table 7: Heterogeneity by teacher characteristics

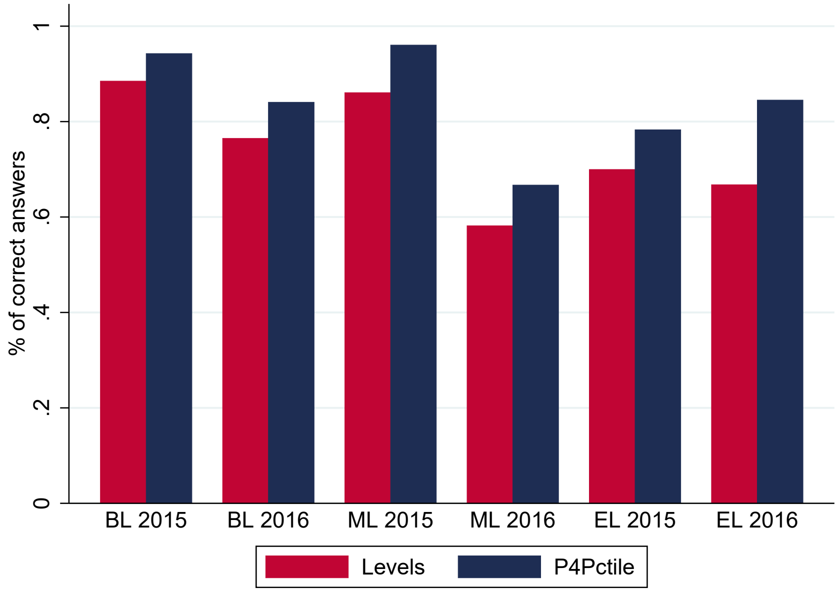
	(1)	(2)	(3)	(4)	(5)
Panel A: Math					
	Male	Age	IRT	HT Rating	Self Rating
Levels*Covariate (α_2)	0.033	0.00080	0.016	0.073***	0.041
	(0.070)	(0.0016)	(0.037)	(0.021)	(0.035)
P4Pctile*Covariate (α_1)	-0.017	0.00056	-0.025	0.012	0.058*
	(0.060)	(0.0016)	(0.038)	(0.022)	(0.035)
N. of obs.	9,650	9,650	9,650	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 variable are student test scores. The data includes 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 asked for math and Kiswahili teachers at the end of the second year — and self-rating (Column 5). . Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.5.2 Teacher Understanding

Complex teacher incentive programs may be less effective if teachers are unable to understand the details of the design, and thus cannot 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. As 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.[20] 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 quiz questions to further ensure that teachers understood the design details. The results of the teacher comprehension tests are shown in Figure 7. As we asked different questions during each survey round (Baseline, Midline and Endline), we cannot compare the trends in understanding over time. Despite the lack of temporal comparability, teacher comprehension was generally high and roughly equal across both types of incentive programs. This provides some assurance that our results are not driven by differences in program comprehension.

Figure 7: Do Teachers Understand the Interventions?



Although teacher understanding was relatively high, we also test for heterogeneity in

[20] 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.

treatment effects by teachers' understanding (at endline). Since there is no comparable test for control group teachers, we cannot interact the treatment variable with teacher understanding. Rather, 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. 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.[21] The results are shown in Table 8 for math and Kiswahili. Focusing on the differences between understanding within each incentive group, we cannot reject the equality of the coefficients, which suggests that better program understanding is not associated with higher treatment effects. This is likely due to the extensive communication efforts that repeatedly reinforced the main details of the program design to teachers.

[21] 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 8: Heterogeneity by teacher's understanding

	(1) Math	(2) Swahili
Levels (high-understanding)	0.031	0.075*
	(0.044)	(0.042)
Levels (low-understanding)	0.073*	0.082**
	(0.041)	(0.037)
P4Pctile (high-understanding)	0.0066	0.027
	(0.035)	(0.036)
P4Pctile (low-understanding)	0.049	-0.0080
	(0.043)	(0.041)
N. of obs.	9,650	9,650
Levels:High-Low	-.042	-.0073
p-value (Levels:High-Low=0)	.27	.84
P4Pctile:High-Low	-.043	.035
p-value (P4Pctile:High-Low=0)	.31	.41
P4Pctile:High-Levels:High	-.025	-.048
p-value (P4Pctile:High-Levels:High=0)	.59	.26
P4Pctile:Low-Levels:Low	-.024	-.09
p-value (P4Pctile:Low-Levels:Low=0)	.64	.052

The outcome variable are student test scores in math (Column 1) and Kiswahili (Column 2). Each regression pools the data for both follow-ups. Teacher's are classified as above or below the median in each follow-up. Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.5.3 Cooperation

Critics of teacher performance pay systems often argue that individual teacher incentives can negatively impact cooperative behavior among teachers. These concerns may be especially relevant for teachers in the pay for percentile scheme, or more generally in rank-order tournaments schemes, as teachers could potentially sabotage other teachers to improve their relative ranking. As our incentive program was only implemented in Grades 1, 2, and 3, the exclusion of other teachers may reduce cooperation, and increase jealousy within treatment schools. In addition, tournament style incentive programs, such as pay for percentile schemes, can encourage sabotage in extreme cases.[22]

We examine the extent to which teacher incentives reduce cooperative behavior between teachers in the treatment schools relative to the control schools in Table 9. Coop-

[22] We examined the potential for sabotage. However, we found very low rates (5 percent or less) of reported sabotage attempts by other teachers, with no differences across treatments

eration, measured by the levels of assistance provided by other teachers, was generally lower for both types of incentives. Program teachers reported receiving between 0.32 and 0.42 fewer instances of assistance from their fellow teachers (see Table 9, Column 1). Relative to the control group mean, this translates to a 23 percent to 30 percent reduction in the levels of assistance. The differences between the levels treatment and the pay for percentile treatment are not statistically significant. We do not find any statistically significant treatment effects in the extensive margin of receiving help from other teachers (Column 2). However, when we examine the quality of assistance, we find that teachers in the levels treatment report they are almost six percentage points (or 8 percent) less likely to receive quality assistance from their peers, while teachers in the pay for percentile treatment report a negligible reduction (Column 3). The reductions in the quality of assistance is significantly lower in levels schools, which could reflect more jealousy among non-participating teachers. As the levels systems is easier to understand, non-participating teachers are perhaps better able to judge the potential pay-off of the incentive system relative to teachers in pay for percentile systems. As head teachers were included in both incentives designs, they may be more motivated to assist program-eligible teachers in their schools. However, our results in Column 4 show that there was no significant change in head teacher assistance to program teachers in either treatment relative to control schools.

Table 9: Teacher behavioral responses: cooperation

	Help from other teachers (# last month) (1)	Help from other teachers (last month>1) (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 show the effect of treatment on teacher reports of help and cooperation from other teachers: The number of times the teacher receives help from other teachers (Column 1), whether the teacher received any help from other teachers or not (Column 2), a high rating of the advice or help received from other teachers (Column 3), and a high rating of the advice or help received from the head teacher (Column 4). Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.5.4 Teacher Expectations

Even though we equalized the budgets across our treatments, 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 the realization of the bonus payments, we collected teachers' expected earnings 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 simply compare teachers in the pay for percentile arm to the levels program, which serves as the omitted category in Table 10. 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 (42 USD) less in bonus payments than teachers in the levels system. This represents an 18 percent reduction 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.[23]

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 suggests that teachers were quite optimistic about their projected earnings; 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 of the district (Column 4).

Table 10: 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 payments (Column 5). Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.5.5 Goal Setting

In addition to being 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.[24] Recall that all teachers in our study, including control teachers, are provided with student baseline reports, which detail initial levels of student performance. Although teachers in both treatment groups were equally likely to use these reports, there is suggestive evidence that teachers in the levels system were slightly more

[23] As mentioned above, since the budget per student was equal across both systems, the average payoff is the same across both systems.

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

likely to use these baseline reports to set goals (results not shown). We explore differences in goal setting behavior between teacher in our incentive systems in Table 11. We do not find any differences in goals set for general school level exams (Column 1). However, we find that teachers in the levels system were almost 8 percentage points more likely (or twice as likely) to have set clear goals for the high stakes Twaweza exam (Column 2). In contrast, teachers in pay for percentile schools were 2.5 percentage points more likely to have set clear goals, although this effect is not statistically significant (Column 2). Although we cannot reject the equality of the two estimates, the results provide some suggestive evidence that the levels systems facilitated more goal setting and targeting on the high-stakes (Twaweza) exam. Teachers in both systems are less likely to set goals for general student learning (Column 3), as well as their own knowledge (Column 4). In terms of the high-stakes Twaweza test, which was administered to all schools, teachers in both incentive schools were approximately 7 percentage points (roughly 8 percent) more likely to set a general goal for the test (Column 5). Additionally, teachers in levels schools were almost 10 percentage points (50 percent) more likely to set a specific numerical target for the Twaweza high stakes test, compared to just under 4 percent of teachers in pay for percentile schools (Column 6). Although these differences are not statistically distinguishable, the point estimates suggest greater incidences of goal setting among teachers in the levels design.

Table 11: Goal Setting

	Goals				Twaweza test goals	
	School exam	Twaweza exam	Student learning	Own knowledge	General	Specific (number)
	(1)	(2)	(3)	(4)	(5)	(6)
Levels (α_1)	-.02	.076**	-.088**	-.097**	.067**	.095*
	(.053)	(.029)	(.04)	(.037)	(.031)	(.052)
P4Pctile (α_2)	-.047	.025	-.077*	-.066*	.076***	.036
	(.048)	(.027)	(.042)	(.037)	(.022)	(.042)
N. of obs.	1,016	1,016	1,016	1,016	1,016	1,016
Mean control	.46	.078	.34	.25	.89	.19
$\alpha_3 = \alpha_2 - \alpha_1$	-.027	-.05	.011	.031	.0094	-.059
p-value($\alpha_3 = 0$)	.58	.14	.78	.42	.7	.27

This table show the effect of treatment on whether teacher set professional goals (Column 1-4) and specific goals for the twaweza exam (Column 5-6). Specifically, whether they set goals for the school exams (Column 1), the twaweza exams (Column 2), student learning (Column 3), and improving content knowledge (Column 4). In addition, whether have general goals for student performance on the twaweza exam (Column 5) or specific (numeric) goals (Column 6). Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As the high-stakes Twaweza test covers the major skills that are rewarded in the levels treatment, we can compute skill-specific pass rates for each student in each treatment. For example, we can compute the fraction of students who pass addition on the second grade math test. The ability to create skill-specific pass rates for all students enables us to use the high-stakes Twaweza testing data to explore differential patterns in student pass rates by treatment in Appendix Tables A.7, A.8, and A.9. Generally, the results show that teachers in levels schools are better able to improve specific skills on the high-stakes test. This is consistent with the idea that the levels treatment enables teachers to better focus instruction using the explicit incentive structure as a target. In addition, because the levels treatment is clear about which skills will be tested, and how skill proficiency is rewarded, it allows teachers to better teach to the test. However, as the treatment effects for teachers in the levels design are also higher, these patterns could be entirely driven by those gains. Thus, these results are merely suggestive of increased targeting or teaching to the test in levels schools.

5.6 Cost effectiveness

The total annual cost of the teacher incentive programs was 7.23 USD 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. We do not have accounting data at the treatment arm level and do not attempt to separate the costs by program. XX TO DISCUSS XX [25].

While cost-effectiveness comparisons require several assumptions (e.g., external validity) (Dhaliwal, Duflo, Glennerster, & Tulloch, 2013), we estimate the increase in test scores per dollar spent, as this is the relevant policy question. The parameters for this calculation are presented in XX.

Using the treatment effects from the high-stakes exam, the pay for percentile program increases test scores by 1.17 SD per 100 USD spent and the levels program increases them by 2.47 SD per 100 USD spent. These assumptions make the cost-effectiveness comparable to most programs since they are evaluated using high-stakes exams and cost-effectiveness is calculated using the long-term cost of the program.

Education administrators face pressure from teachers to increase salaries but are increasingly confronted with questions about learning progress and performance accountability. In this context it is not unrealistic to assume that a Government contemplating the introduction of a teacher performance pay scheme would finance it out of budget that otherwise would have paid for unconditional pay increase. In such a scenario, the principal cost of the incentive program is the administrative cost of implementing the program (costs of communicating the bonus offer, independent measurement and recording of student learning, organising the payments) and not the cost of the bonus itself. If we exclude the costs of the bonus, our estimates of cost-effectiveness are XX for the pay for percentile program and YY for the levels program.

These estimates suggest that both programs are cost-effective compared to several other interventions in developing countries in the overview by (Kremer, Brannen, & Glennerster, 2013). However, teacher incentives in Western Kenya have been shown to increase test scores by over 6 SD per 100 USD spent.

6 Conclusion

We present estimates of the impact on early grade learning of two teacher incentive programs implemented as a randomised trial in a nationally representative sample of

[25] Costs of the pre-treatment testing required in pay for percentile is 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). Pay for percentile also took a little longer to explain but reports from field team leaders indicate that the difference with levels was small. The main cost difference is in data management: preparing the ability groups, programming the payment calculations. However, these are largely fixed costs and so unit costs would be relatively small in steady state. Overall the pay for percentile is the more expensive program to implement

180 public primary schools in Tanzania. Specifically, we compare a multiple thresholds proficiency incentive design with a pay for percentile system in terms of impact on independently measured test scores of students in Grades 1-3, relative to a control group.

We report X four X main findings. First, we find that both programs led to increases in test scores in the focal grades. Second, we do not find any negative effects on test scores for non-incentivised subjects or grades. Third, despite the theoretical advantage of the pay for percentile system, overall this design is not more effective than the simple proficiency system in improving mean student test-scores. In particular, we find that point estimates of impact on high stakes test scores are higher under the proficiency system for all incentivised subjects, and significantly higher for Kiswahili. Fourth, and contrary to expectations, the pay for percentile system leads to learning improvements primarily among the best students, while the levels system benefitted student from a wider range of initial abilities.

To help interpret our results we report on potential mechanisms. We do not find that teacher understanding of the programs, as measured by scores on quizzes about program rules, is associated with impact. However, we do find that teachers in the percentile pay program had significantly lower *expectations of bonus earnings compared to teachers in the levels system (equal to 36 percent of realised mean bonus in 2016). We find some evidence of reduced cooperation between teachers as a result of the incentives....*

Goal setting mechanism.

Overall, a simpler system with multiple thresholds can actually outperform a more complex incentive system, especially in countries with low administrative capacity such as Tanzania.

References

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

Balch, R., & Springer, M. G. (2015). Performance pay, test scores, and student learning objectives. Economics of Education Review, 44*(0), 114 - 125. Retrieved from* http://www.sciencedirect.com/science/article/pii/S0272775714001034 *doi: http://dx.doi.org/10.1016/j.econedurev.2014.11.002*

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. Retrieved from* http://books.google.com/books?id=Tj0TF0IHIyAC

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

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.*

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.*

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.*

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.*

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

Gilligan, D., 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*

Goldhaber, D., & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's procomp teacher pay initiative. Economics of Education Review*, 31(6), 1067 - 1083. Retrieved from* http://www.sciencedirect.com/science/article/pii/S0272775712000751 *doi: http://dx.doi.org/10.1016/j.econedurev.2012.06.007*

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

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

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

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. American Economic Review*, 99(5), 1979-2011. Retrieved from* http://www.aeaweb.org/articles.php?doi=10.1257/aer.99.5.1979 *doi: 10.1257/aer.99.5.1979*

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

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.

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. Annu. Rev. Econ., *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.*

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J., McCaffrey, D. F., . . . Stecher, B. M. (2011). Teacher pay for performance: Experimental evidence from the project on incentives in teaching (point). Society for Research on Educational Effectiveness.

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)*

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. (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 are two treatments and a control group in this experiment, and 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 assign a higher proportion of schools in treatments C1 (which involved threshold teacher incentives) to both "levels". Similarly, we assign a higher proportion of schools in the control group of 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[26]*. Table A.1 summarizes the number of schools randomly allocated to each treatment arm based on their assignment in the previous experiment. In short, in each district, we have 18 schools. In each district,there are six schools in each of the new treatment groups (levels, gains, and control). In each one of the new treatments, there are 60 schools. 30 of these schools are above the median in baseline learning and 30 are 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	Gains	Control	Total
KiuFunza I	C1	40	20	10	70
	C2	10	30	30	70
	C3	10	10	20	40
	Total	60	60	60	180

[26] We create an overall measure of student learning, and split schools as above or below the median

A.2 Additional Tables

A.2.1 Balance in Teacher Turnover

Table A.2: 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 teaching math, English or Kiswahili in grades 1, 2, and 3 at the beginning of 2015 that are still teaching them (in the same school) at the end of 2015 (Column 1) and 2016 (Column 2). Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.2 Effects on Test-Takers and Lee Bounds on High-Stakes Test

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

	(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) in the high-stakes exam. The unit of observation is at the school-grade level. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

	(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 Gains schools so that the proportion of test takes is the same as the number in control schools). Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.3 Additional Heterogeneity in Treatment Effects

Table A.5: Heterogeneity by student characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Math			Swahili			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 (Column 3, 6, and 9). Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: Heterogeneity by school characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Math			Swahili			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 (Column 3, 6, and 9). Clustered standard errors, by school, in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.4 Pass Rates

Table A.7: 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 passed a given skill in the high-stakes exam. Clustered standard errors, by school, in parenthesis.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: Pass rates using ‘levels’ thresholds in math

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

The independent variable is whether a student passed a given skill in the high-stakes exam. Clustered standard errors, by school, in parenthesis.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.9: 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 passed a given skill in the high-stakes exam. Clustered standard errors, by school, in parenthesis.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.5 National Assessments

Table A.10: National assessments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Grade 4 - SFNA									
	Grade 4 SFNA 2015			Grade 4 SFNA 2016			Grade 4 SFNA 2017		
	Pass	Score	Test takers	Pass	Score	Test takers	Pass	Score	Test takers
Levels (α_1)	-0.06*	-0.20**	3.04	-0.05**	-0.21**	15.77	0.02	-0.06	25.61*
	(0.04)	(0.08)	(8.79)	(0.03)	(0.08)	(9.84)	(0.03)	(0.08)	(13.75)
P4Pctile (α_2)	-0.05	-0.18**	-6.20	0.01	-0.01	0.93	0.01	-0.11	4.83
	(0.03)	(0.08)	(8.24)	(0.03)	(0.08)	(8.03)	(0.03)	(0.07)	(10.98)
N. of obs.	13,853	13,853	166	12,487	12,487	153	14,575	14,575	148
N. of schools	168	168	166	157	157	153	153	153	148
Mean control group	0.67	2.87	63.6	0.81	3.24	65.3	0.79	3.30	75.8
$\alpha_3 = \alpha_2 - \alpha_1$	0.018	0.019	-9.23	0.059**	0.19**	-14.8	-0.012	-0.049	-20.8
p-value ($H_0 : \alpha_3 = 0$)	0.61	0.82	0.33	0.032	0.018	0.13	0.67	0.48	0.15
Panel B: Grade 7 - PSLE									
	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 parenthesis.

A.2.6 Wasted money

To estimate how much money was 'wasted' in the level's 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 an ability or not,* X_i *is a set of student controls (including region, grade, and school characteristics), and we restrict the sample to the control group.*

- *We then estimate the probability of passing each ability 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 given for results that are not related to additional effort excreted by the teacher. The results are below.*

Table A.11: Wasted money

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