NBER WORKING PAPER SERIES

DESIGNING EFFECTIVE TEACHER PERFORMANCE PAY PROGRAMS: EXPERIMENTAL EVIDENCE FROM TANZANIA

Isaac Mbiti
Mauricio Romero
Youdi Schipper

Working Paper 25903
http://www.nber.org/papers/w25902

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019

We are especially grateful to Karthik Muralidharan for his collaboration in the early stages of this project and subsequent discussions. We thank the leadership and staff at Twaweza for their collaboration and support on this project. We would also like to thank Austin Dempewolff, Mitch Downey, David Evans, John Friedman, Delia Furtado, Guthrie Gray-Lobe, Ronak Jain, Joseph Mmbando, Molly Lipscomb Johnson, Terence Johnson, Michael Kremer, Derek Neal, Omer Ozak, Bobby Pakzad-Hurson, Wayne Aaron Sandholtz, Enrique Seira, Daniela Scur, Jay Shimshack, Bryce Millet Steinberg, Tavneet Suri, and seminar/conference participants at UC San Diego, Universidad del Rosario, NEUDC, PacDev, RISE, SOLE, and SREE for their comments. Erin Litzow and Jessica Mahoney provided excellent research assistance through Innovations for Poverty Action. We are also grateful to EDI Tanzania for their thorough data collection and implementation efforts. The EDI team included Respichius Mitti, Phil Itanisia, Timo Kyessey, Julius Josephat, Nate Sivewright, and Celine Guimas. This evaluation was funded by Twaweza with supplemental funding 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. Financial support from the Asociacion Mexicana de Cultura, A.C. is gratefully acknowledged by Romero. This paper was partly written while Mbiti was visiting the Brown University Economics Department and the Population Studies and Training Center (PSTC). We received IRB approval from Innovations for Poverty Action, UC San Diego, and University of Virginia. The protocol was also reviewed and approved by the Tanzania Commission for Science and Technology (COSTECH). A randomized controlled trials registry entry and the pre-analysis plan are available at: https://www.socialscienceregistry.org/trials/1009 The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w25903.ack

Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania
Isaac Mbiti, Mauricio Romero, and Youdi Schipper
NBER Working Paper No. 25903
May 2019
JEL No. C93,H52,I21,M52,O15

ABSTRACT

We use a field experiment in Tanzania to compare the effectiveness on learning of two teacher performance pay systems. The first is a Pay for Percentile system (a rank-order tournament). The second rewards teachers based on multiple proficiency thresholds. Pay for Percentile can (under certain conditions) induce optimal effort among teachers, but our threshold system is easier to implement and provides teachers with clearer goals and targets. Both systems improved student test scores. However, the multiple-thresholds system was more effective in boosting student learning and is less costly.

Isaac Mbiti
Batten School of Leadership and Public Policy
University of Virginia
P. O. Box 400893
Charlottesville, VA 22904
and NBER
imbiti@virginia.edu

Mauricio Romero
Centro de Investigacion Economica
ITAM
Mexico
mtromero@itam.mx

Youdi Schipper
Twaweza
127 Mafinga Road
off Kinondoni Road
P.O. Box 38342
Dar es Salaam
Tanzania
yschipper@twaweza.org

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/1009

1 Introduction

Over the past two decades, global education priorities have shifted from increasing primary school enrollment to promoting policies that improve learning. This shift has been driven in part by evidence revealing poor and stagnant levels of learning among students in developing countries, despite significant investments in education (World Bank, 2018). Given the central role of teachers in the education production function (Hanushek & Rivkin, 2012; Chetty, Friedman, & Rockoff, 2014b, 2014a), as well as the large share of the education budget devoted to their compensation, policy makers and researchers are increasingly interested in interventions that increase the effectiveness of teachers. Teacher performance pay programs are seen as a potential policy response to address these concerns because they strengthen the links between teacher renumeration and student learning outcomes (World Bank, 2018; Bruns, Filmer, & Patrinos, 2011).[1] However, there is limited evidence on how to best structure teacher incentives.

Insights from economic theory suggest that sophisticated teacher incentive designs, such as those based on rank-order tournaments, are more effective and may induce greater — and potentially socially optimal — levels of effort among teachers as opposed to those based on proficiency thresholds (Neal & Schanzenbach, 2010; Neal, 2011; Barlevy & Neal, 2012; Loyalka, Sylvia, Liu, Chu, & Shi, in press). However, the theoretical advantages of rank-order tournament incentive schemes may not materialize in practice because participants need to "think strategically about their co-workers' efforts to find a Nash equilibrium" (Charness & Kuhn, 2011). That is, it may be more difficult for teachers to determine how to react optimally to such schemes. In contrast, incentive schemes based on proficiency thresholds are perceived as more transparent and easier to implement compared to tournaments. Such incentive designs are commonly used in education systems, despite their well-known shortcomings — including encouraging teachers to focus on marginal students (Neal & Schanzenbach, 2010). Since each incentive system has its own set of practical and theoretical advantages and disadvantages, there is a need for empirical studies that compare the effectiveness of different designs.

We conduct a randomized experiment that examines the effectiveness of two different incentive schemes in a nationally representative set of 180 Tanzanian public schools. The first is a "Pay for Percentile" (a rank-order tournament) scheme based on research

[1] Teacher performance pay programs have been implemented in both developed and developing contexts. For instance, the share of US school districts with teacher performance pay programs has increased by more than 40% from 2004 to 2012 (Imberman, 2015). Less developed countries such as Brazil, Chile, and Pakistan have also implemented performance pay programs, often as large pilot programs (Alger, 2014; Ferraz & Bruns, 2012; Barrera-Osorio & Raju, in press; Contreras & Rau, 2012).

by Barlevy and Neal (2012). The second is a "Levels" scheme that features multiple proficiency thresholds. To the best of our knowledge, this is the first documented implementation of proficiency-based teacher incentives with multiple (curricular-based) thresholds. Both types of incentive programs rewarded teachers for the performance of their students on externally administered tests in math, Kiswahili, and English in first, second, and third grade.[2] To facilitate comparisons, the per-student bonus budget was equalized (ex-ante) across grades, subjects, and treatment arms. The average teacher bonus was equal to approximately 3.5% of the annual net salary (roughly half a month's pay).[3] Further, all teachers in our study were provided with baseline student reports so they were aware of each students' initial skill (or proficiency) level. Following Neal (2013) and Mbiti et al. (in press), we evaluate the incentive programs using data from both the incentivized (or "high-stakes") test that was administered to all students to determine teacher bonuses and a non-incentivized (or "low-stakes") test that was administered to a sample of students for research purposes.[4]

In the 60 schools assigned to the Pay for Percentile arm, students were first tested and then assigned to one of several "baseline ability groups" based on their learning level (across all schools). At the end of the school year, students were re-tested and ranked within their assigned group based on their test scores. Teachers were rewarded in proportion to their students' rankings within each group. By handicapping the differences in initial student performance across teachers, the system does not penalize teachers who serve disadvantaged students. To ensure that teachers understood the incentive scheme, we developed information packets that used culturally appropriate scripts and examples, and also budgeted extra time to explain the design details in this treatment.

In the 60 schools assigned to receive incentives based on proficiency targets (the "Levels" arm), teachers earned bonuses based on their students' mastery of several grade-specific skills. As incentive programs using single-proficiency thresholds encourage teachers to focus on students close to the passing threshold, we included several thresholds to mitigate this concern. The skill thresholds were salient milestones based on the national curriculum and ranged from very basic (e.g., number recognition) to more complex skills (e.g., multiplication) in order to allow teachers to earn rewards from a wide

[2] English was dropped from the national curriculum in first and second grades in the middle of the experiment. We therefore focus on our analysis on math and Kiswahili test scores and present the analysis of English scores in the Appendix.

[3] Similar incentive sizes were used in Fryer (2013); Glewwe, Ilias, and Kremer (2010); Mbiti et al. (in press); Muralidharan and Sundararaman (2011); Lavy (2002); Ladd (1999); Vigdor (2008). See Leigh (2012) for additional details.

[4] Both types of tests were conducted in control schools. However, the results of the "incentivized" test did not trigger any payments in these schools.

range of students. This system retains the clarity of more basic single-proficiency threshold systems. Since teachers in developing contexts are generally unfamiliar with incentive schemes, such clarity can be an important factor in encouraging the widespread acceptance of the system. Further, Miller and Babiarz (2013) argue that threshold (or "bright-line") designs are well-suited for situations where the thresholds correspond to important goals or objectives. As reward payments for each skill were inversely proportional to the number of students that attained the skill, harder-to-obtain skills were rewarded more.[5] Since the Levels system only requires students to be tested at the end of the year, it is simpler to implement than the Pay for Percentile system which requires linked beginning and end of year student test scores. However, as rewards are based on absolute learning levels, the system may disadvantage teachers who serve students from poorer backgrounds.

We report two main findings. First, both types of teacher incentives are effective at improving learning outcomes compared to the control group, especially when we examine the results from the incentivized tests. Focusing on the results at the end of the second year of the program, we find modest test score increases for students in Pay for Percentile schools relative to students in control schools. Using a composite measure of test-scores across subjects (math and Kiswahili), students in this treatment arm scored $.13\sigma$ higher (p-value .027) compared to the control group. The Levels treatment was also effective at increasing student learning. At the end of the second year, the composite test scores of students in the Levels treatment were $.22\sigma$ higher (p-value < 0.01) compared to the control group.

Test score gains in both treatments were lower on the non-incentivized tests. Composite test scores increased by $.044\sigma$ (p-value .31) and $.096\sigma$ (p-value .037) in Pay for Percentile schools and Levels schools. As the test content was similar, the differences in treatment effects are likely due to differences in student effort on test taking. The incentivized tests involved the whole classroom and were typically used as official end of school year tests. This could induce teachers to encourage their students to exert (relatively) more effort on these tests. In contrast, the non-incentivized tests involved a smaller number of students and were conducted in a more inconspicuous manner.[6] The effects of student test-taking effort on test scores has been documented in previous

[5] As bonus payments were determined ex-post based on pass rates, teachers face some uncertainty about the exact bonus sizes. We abstract from this by assuming an individual teacher's effort would have a negligible effect on the aggregate pass rate and that teachers have sufficient ex-ante information (e.g., through experience) to have reasonable predictions about the pass rates. Appendix B provides more details.

[6] Appendix C provides more details about the testing procedures.

studies such as Levitt, List, Neckermann, and Sadoff (2016), Gneezy et al. (2017), and Mbiti et al. (in press).

Our second main finding suggests that the details of teacher incentive design matter. Despite the theoretical advantages of the Pay for Percentile system, in this setting the Levels incentive system was more effective at improving learning than the Pay for Percentile system. At the end of the second year, the estimated treatment effect on the incentivized composite test score in Levels schools was .096σ higher (p-value .097) relative to the estimates for Pay for Percentile schools. The treatment effect on the non-incentivized composite test score shows a similar pattern, although the difference is smaller (.052σ) and statistically insignificant (p-value .29). The greater learning gains in Levels schools are also reflected in lower grade repetition rates in those schools relative to Pay for Percentile schools. At the end of the second year, students in Levels schools were 3.3 percentage points less likely (p-value .048) to repeat a grade compared the students in the control group. As 14 percent of students in control schools repeated a grade, this represents a 25 percent reduction. This reduction in repetition among Levels schools was significantly different (p-value .041) from the limited (and insignificant) effect in the Pay for Percentile treatment.

We also use a comprehensive set of survey data collected from school administrators, teachers, and students to shed light on theoretically relevant mechanisms. We do not find any evidence of negative treatment effects on non-incentivized subjects such as science, suggesting that learning gains in incentivized subjects were not at the expense of learning in other subjects. Despite the concerns that the Levels system may induce teachers to focus on marginal students, we find similar learning gains across all five quintiles of the student baseline test score distribution in the second year (using composite test scores) in both treatment arms.

Given the well-documented concerns about teachers misunderstanding incentive designs (Goodman & Turner, 2013; Fryer, 2013), teacher comprehension was high under both systems, allowing us to rule out a differential lack of understanding as a major driving factor. However, even if teachers understood how payments were made, those in the Pay for Percentile schools may have had a relatively harder time determining how to best react to the incentive. For instance, if teachers had limited information (or ambiguity) about the efforts and capabilities of other teachers in the tournament, they may have reduced their own effort in the Pay for Percentile scheme if they believed they were not competitive. Consistent with this notion, teachers in the Pay for Percentile schools reported they expected to receive 18 percent lower bonus payments, on average, compared to their Levels counterparts. These lower expectations could have dampened teachers' responsiveness to the incentives and reduced effort. In addition, teachers in the Levels system were better able to articulate clear and specific targets for their students on the incentivized tests, perhaps due to the clearer reward structure and salient thresholds.

Our study contributes to debate on the optimal design of teacher incentives. There are only a limited set of adequately powered experimental studies that compare different teacher incentive designs (e.g., Muralidharan and Sundararaman (2011); Loyalka et al. (in press); Fryer Jr, Levitt, List, and Sadoff (2012)).[7] Our study provides the first evidence on the effectiveness of a novel multiple threshold incentive system tied to specific curriculum objectives. In addition, it shows that such system can elicit greater effort from teachers and deliver greater learning outcomes in early grades when compared to a more sophisticated cost-equivalent rank-order tournament scheme. This highlights the importance of the practical limitations of tournaments outlined in Charness and Kuhn (2011). In addition, salient learning targets, such as those used in our Levels design, can promote teacher effectiveness, especially in settings where teacher capacity is relatively limited.

Overall, our results are consistent with the existing evidence suggesting that teacher incentives tend to be more effective in contexts where there is low accountability in the education system (Imberman, 2015; Ganimian & Murnane, 2016; Glewwe & Muralidharan, 2016). However, only a limited number of studies have examined tournament style teacher incentives, and we are only aware of two studies in developing country contexts that specifically evaluate Pay for Percentile schemes (see Loyalka et al. (in press) and Gilligan, Karachiwalla, Kasirye, Lucas, and Neal (2018)).[8]

Since education systems in developing countries face numerous challenges including accountability constraints and lower teacher capacity relative to global scales, evidence that can provide policy makers with insights on the relative cost-effectiveness of programs that can improve learning outcomes is especially important (World Bank, 2014, 2018). Our results highlight the trade-offs faced by education authorities who have to consider the effectiveness and feasibility of implementation of different teacher incentive designs, often with limited information about the education production function.

[7] There is a small but growing set of studies that compare the effectiveness of different types of provider incentives in the healthcare sector in developing countries. Examples include Singh and Masters (2018) and Mohanan, Donato, Miller, Truskinovsky, and Vera-Hernández (2019).

[8] Loyalka et al. (in press) find that Pay for Percentile incentives increased test scores among math teachers in Chinese schools. Gilligan et al. (2018) find that Pay for Percentile have no impact on student learning in Ugandan schools, except for top students in schools with textbooks. In the US context, Fryer Jr et al. (2012) find that Pay for Percentile schemes are most effective when the rewards are framed as losses, where teachers are first given a lump sum payment, and then required to return part of the payment if their students do not meet the required targets.

2 Experimental Design

2.1 Context

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

The poor quality of education is driven in part by limited accountability in the education system. Quality assurance systems (e.g., school inspectors) typically focus on superficial issues such as the state of the school yard, rather than on issues that may affect learning (Mbiti, 2016). The lack of accountability is further reflected in teacher absence rates. Data from unannounced spot checks shows that almost a quarter of teachers were absent from school, and only half of the teachers who were at school were in the classroom (World Bank, 2011). As a result, almost 60 percent of planned instructional time is lost (World Bank, 2011).

Tanzanian teachers' unions have been actively lobbying for better pay as a way to address quality concerns in the education system. Yet, the correlation between teacher compensation and student learning is extremely low (Kane, Rockoff, & Staiger, 2008; Bettinger & Long, 2010; Woessmann, 2011; de Ree, Muralidharan, Pradhan, & Rogers, 2018). Moreover, teachers salaries are currently relatively high — approximately 500,000 TZS per month ($\sim$ US$300) or roughly 4.5 times GDP per capita (World Bank, 2017) — and approximately 60 percent of the education budget is devoted to teacher compensation.[9] Despite the relatively attractive wages of Tanzanian teachers, the teachers' union called a strike in 2012 to demand a 100 percent increase in pay (Reuters, 2012; PRI,

[9] 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).

2013).[10]

2.2 Interventions and Implementation

The interventions in this study were developed in close collaboration with Twaweza, an East African civil society organization that focuses on citizen agency and public service delivery. The interventions were part of a series of projects launched under a broader program umbrella known as KiuFunza ('Thirst for learning' in Kiswahili).[11]

The KiuFunza program targets teachers in focal grades 1, 2 and 3 who are responsible for teaching the focal subjects Kiswahili, English and math (arithmetic). A budget of US$150,000 per year for teacher and head teacher incentives was split between the two treatment arms in proportion to the number of students enrolled. As a result, the prize money in each treatment arm was approximately US$3 per student. All interventions were implemented by Twaweza in partnership with EDI (a Tanzanian research firm) and a set of local district partners. Head teachers were offered a bonus of 20 percent of the combined bonus of all incentivized teachers in his or her school.[12]

Within each intervention arm, Twaweza distributed information describing the program in early 2015 and 2016: first to focal grade and subject teachers and head teachers, then to their respective communities via public meetings. From the program's onset Twaweza informed teachers the program would last two years. The implementation teams also conducted mid-year school visits to re-familiarize teachers with the program, gauge teacher understanding of the bonus payment mechanisms, and answer any remaining questions.

At the end of the school year, all students in grades 1, 2, and 3 in every school, including control schools, were tested in Kiswahili, English, and math. Because this test was used to determine teacher incentive payments, it was considered "high-stakes" (from the teachers' perspective). Our non-incentivized research test was conducted on a different day, but within a few weeks from the incentivized test. Both sets of tests were based on the Tanzanian curriculum and were developed by Tanzanian education professionals using the Uwezo learning assessment framework.[13] We provide additional

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

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

[12] Twaweza included head teachers in the incentive design to ensure that they would be stakeholders in improving learning outcomes. Likewise, any scaled-up teacher incentive program would feature bonuses for head teachers.

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

details about the design and implementation of both types of tests in Appendix C.

2.2.1 Pay for Percentile design

The Pay for Percentile design used in our intervention is based on research by Barlevy and Neal (2012). They show that this incentive structure can, under certain conditions, induce teachers to exert socially optimal levels of effort. One important necessary condition for Pay for Percentile to induce optimal effort is that teachers believe they are competing in properly seeded (or fair) contests. To achieve this, the Pay for Percentile uses a modified rank-order tournament structure that accounts for the heterogeneity in students baseline learning levels across classrooms (and teachers). Specifically, the system divides students into groups based on their academic achievement (or "ability"), and a separate rank-order tournament is conducted for each group. Teacher's are then rewarded on the basis of their students' rank-order within each ability group. Without this adjustment, teachers in schools that served students from affluent backgrounds would be advantaged, and those serving less-affluent students may be discouraged from exerting effort.

In order to implement this system in practice, we created student groups with similar initial learning levels based on test score data from the previous school year for each subject-grade combination. Students without test scores in second and third grade were grouped together in an "unknown" ability group.[14] Since none of the first grade students had incoming test scores, we created broad country-level ability groups and assigned all first grade students within a school to the same group based on the historical average test scores for the school. Thus, all first-grade students within a school were assigned to the same group.[15]

To compute the payment structure, we divide the total prize money in this treatment arm equally across grades and subjects. We then apportion the subject-grade budget to each ability-group in proportion to the total number of students in the grade who are in each ability-group. At the end of the year, we ranked students within each ability-group according to their endline test score. Within each ability-group we assigned teachers points proportional to the rank of their students. For a given ability-group, a teacher would receive 99 points for a student in the top 1% of the group and zero points for a student in the bottom 1% of the group. In other words, the rewards increase linearly in rank. The total amount of money paid per point is the same across all groups, in all

[14] Roughly 20% of students are grouped into the "unknown" ability group. This includes newly enrolled students, as well as students who were enrolled but for some reason were not tested at baseline.

[15] Our results are robust to excluding grade 1 students from the sample. See Table D.1 in Appendix D.

subjects, and in all grades.

For example, suppose there is a total of US$1,000 for teacher incentives and that there are two ability groups with 40 and 60 students. Accordingly, the total budget for teacher bonuses in each ability group would be US$400 and US$600. In each ability-group, the total bonus would be equal to the sum of all teacher rewards or

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

where X is the total budget for teacher bonuses in each ability group, N is the number of students in each ability group, i indexes a student's percentile rank on the endline test, and b is the teacher reward per point. Since $\sum_{i=1}^{100}(i-1) = 4,950$, the reward per point ($b$) is roughly $\sim$US$0.20 for both groups. Thus, in this example if a student was in the top 1% of the either ability-group, their teacher would earn $99 * 0.2$ or US$20. Conversely, a median student would earn their teacher $50 * 0.2$ or US$10. In the first year of our study, the total bonus available to teachers in Pay for Percentile schools was US$70,820 and total enrollment was 22,296. For each grade and subject, teachers earned US$1.77 for each student in the top 1% and US$0.89 for each student in the 50th percentile.

Although this design can deliver socially optimal levels of effort, it may be challenging to implement at scale, particularly in settings with weak administrative capacity such as Tanzania. For instance, maintaining child-level panel databases is a non-trivial administrative challenge. Moreover, the Pay for Percentile system may prove difficult to grasp for teachers. It presents each teacher with a series of tournaments (for each ability group in each subject that they teach) and therefore the bonus payoff is relatively hard to predict, even if the design guarantees a fair system. Furthermore, the uncertainty introduced by competing against teachers from schools across the whole country may dilute the incentive.

2.2.2 Proficiency thresholds (Levels) design

Proficiency based systems are easier for teachers to understand and provide more actionable targets than rank-order or value-added tournaments. Consequently, such systems are likely to increase motivation among teachers and head teachers; however, they have well-known limitations. For example, they are unable to adequately account for differences in the initial distribution of student preparation across schools and classrooms. Moreover, this type of system can encourage teachers to focus on students close to the proficiency threshold, at the expense of students who are well above or below the threshold (Neal & Schanzenbach, 2010). To mitigate this concern, our Levels design features multiple thresholds ranging from very basic skills to more advanced skills in the curriculum. This design allows teachers to earn bonuses for helping a broader set of students, including students with lower and higher baseline test scores.[16] Miller and Babiarz (2013) argue that incentive designs based on "bright-line" performance thresholds (and goals) can be effective in helping service providers to focus on achieving these goals. They also argue that bright-line designs are well suited to helping providers focus on achieving important outcomes.[17]

In Levels schools, teachers are paid in proportion to the number of skills students in grades 1-3 are able to master in mathematics, Kiswahili, and English. The total budget is split across grades in proportion to the number of students enrolled in each grade. The budget is then divided equally among subjects and skills within each subject. For example, suppose the budget allocated to one grade for demonstrating proficiency in addition (a math skill) is US$1,000. If there are 500 students in the grade, and 250 pass the addition portion of the math test, then a teacher would receive US$4 for every student in her class that was proficient in addition.

Table 1 shows the skills (i.e., the thresholds) tested in each grade-subject combination and the corresponding (ex-post) payment per student that each teacher would receive. Since the per pass bonus paid ex-post is equal to the skill budget divided by the number of students passing the skill, the budget for easier-to-obtain skills is spread across more students — resulting in a lower per-pass bonus. Conversely, harder-to-obtain skills have a higher per pass bonus. Thus, teachers have the potential to earn larger bonuses if their students are proficient in a larger number of skills, especially harder-to-obtain skills.[18]

[Table 1 about here.]

[16] As discussed in Appendix B, a key practical challenge is ensuring that the thresholds are sufficiently close together to prevent teachers from ignoring students who fall between two thresholds.

[17] In the health sector, Miller and Babiarz (2013) argue bright-lines may be especially appropriate when thresholds have clinical significance (e.g., vaccination rates). In our early grade education setting, the fundamental nature of the numeracy and literacy thresholds in our design corresponds with this criteria.

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

2.3 A Note on English Language Teaching

As Kiswahili is the official language of instruction in primary schools in Tanzania, English is taught as a second language. However, English is rarely spoken outside of the classroom, so English language skills are quite low in Tanzania. For instance, only 12 percent of grade 3 students were proficient at the grade 2 level in English (Uwezo, 2012). Given the challenges of teaching English in Tanzania, the subject was removed from the national curriculum in grade 1 and 2 in 2015 to allow teachers to focus on numeracy and literacy in Kiswahili in those grades. English was still taught in grade 3, under a revised curriculum. However, the Education Ministry provided little guidance on how to transition to the new curriculum and as a result, there was substantial variation in its implementation. Some schools stopped teaching English in 2015, while others continued until 2016. In addition, there was no official guidance on whether to use grade 1 English materials in grade 3, as no new books were issued that reflected the curriculum changes. To maintain consistency between the curriculum and KiuFunza incentives, Twaweza dropped English from the incentives in grade 1 and 2 in 2016, but included grade 3 English teachers. To avoid confusion, we also communicated that our end-of-year English test in 2016 would still use the pre-reform grade 3 curriculum. Given these issues in the implementation of the curriculum reform, it is unclear how to interpret the results for English. In addition, these estimates are less policy relevant after the reform. Therefore, in order to facilitate a clearer analysis, we only present results for mathematics and Kiswahili in the main text. Table D.2 in Appendix D presents the effects of our treatments on English test scores in grade 3.

3 Data and Empirical Specification

3.1 Sample Selection

The teacher incentive programs were evaluated using a randomized design. First, 10 districts were randomly selected (see Figure 1).[19] The study sample of 180 schools was taken from a previous field experiment — studied by Mbiti et al. (in press) — 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. As mentioned above, the Pay for Percentile design will deliver optimal levels

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

of effort only if teachers believe they are competing in fair contests. Thus, having reliable information about student initial learning levels was key.[20]

Within each district, we randomly allocated schools to one of our three experimental groups. Thus, in each district six schools were assigned to the Levels treatment, six schools to the Pay for Percentile treatment, and six schools served as controls. In total, there were 60 schools in each group. The treatment assignment 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 provided in Appendix A.

[Figure 1 about here.]

3.2 Data and Balance

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

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 supplemented the results from this set of non-incentivized student tests with the results from the incentivized tests that were used to determine teacher bonus payments and were conducted in all schools, including control schools. Most articles studying teacher performance pay use incentivized tests to measure the overall treatment effects. However, it is unclear whether incentivized or

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

non-incentivized tests are better for measuring treatment effects. We therefore present results from both tests for completeness.[21]

Although the content (subject order, question type, phrasing, difficulty level) is consistent across the incentivized and non-incentivized tests, there are a number of important differences in the test administration. The non-incentivized test took longer (40 minutes) than the incentivized test (15 minutes). The non-incentivized test had more questions in each subject to avoid bottom- and top-coding, and also included an "other subject" module at the end to test spillover effects. Further, even though both tests were administered individually to students, the testing environment was different. Non-incentivized tests were administered during a regular school day by survey enumerators. In contrast, the incentivized test was more "official" as all students in grades 1-3 were tested on a specified day. On the test day, a Twaweza test team would administer the tests in dedicated classrooms, with head teachers and teachers managing the flow of students. In addition, most schools used the incentivized test as the official end-of-year test. A number of measures were introduced to enhance test security. First, to prevent test-taking by non-target grade candidates, students could only be tested if their name had been listed and their photo taken at baseline. Second, each student was assigned one test randomly selected out of ten test versions to prevent copying during the test and to reduce the benefits of leakage. Finally, tests were handled, administered, and electronically scored by Twaweza teams without any teacher involvement.

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

[Table 2 about here.]

[21] As argued by Mbiti et al. (in press): "The confirmation that test-taking effort is a salient component of measured test scores by Levitt et al. (2016) and Gneezy et al. (2017) presents a conundrum for education researchers as to what the appropriate measure of human capital should be for assessing the impact of education interventions. On one hand, low-stakes tests may provide a better estimate of a true measure of human capital that does not depend on external stimuli for performance. On the other hand, test-taking effort is costly, and students may not demonstrate their true potential under low-stakes testing, in which case, an 'incentivized' testing procedure may be a better measure of true human capital."

3.3 Empirical Specification

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

$$Z_{isdt} = \delta_0 + \delta_1 Levels_s + \delta_2 P4Pctile_s + \delta_3 Z_{isd,t=0} + X_i\delta_4 + X_s\delta_5 + \gamma_d + \gamma_g + \varepsilon_{isdt}, \qquad (1)$$

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

We scale our test scores using an Item Response Theory (IRT) model and then normalize them using the mean and standard deviation 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.[22]

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

4 Results

In this section, we first explore how both incentive systems affected student test scores and grade repetition. We then explore whether the incentives increase observable teacher effort. We then turn to heterogeneity by students and teacher characteristics. Finally, we explore some possible mechanisms that could explain our results on test scores.

4.1 Test Scores

We present the estimated treatment effects of the incentive programs on student learning using data from both the non-incentivized test (Table 3, Panel A) and the incentivized test (Table 3, Panel B). As discussed earlier, we focus our main analysis on math and

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

Kiswahili due to the curriculum changes that occurred. We provide estimates of the intervention on English test scores in Table D.2 in Appendix D. To ameliorate concerns due to multiple testing, we present a composite index of learning computed using an Item Response Theory model.

In the first year, both incentive schemes resulted in small but imprecisely estimated changes in test-scores on the non-incentivized test. Focusing on the composite learning index (Panel A, Column 3), test-scores increased by about .057σ (p-value .24) in Levels schools relative to the control group. Test scores were -.029σ smaller (p-value .46) in Pay for Percentile schools relative to control schools. In the second year of the program, the estimated treatment effects on the non-incentivized test are generally larger than the first-year estimates (Panel A, Columns 4-6). Test scores on the composite index increased by .096σ (p-value .037) in Levels schools and .044σ (p-value .31) in Pay for Percentile schools.

Most of the existing literature on teacher incentives relies on data from incentivized tests that are used to determine teacher rewards (Muralidharan & Sundararaman, 2011; Fryer, 2013; Neal & Schanzenbach, 2010). Following this norm, we also present the treatment effects of our interventions using incentivized exams (Panel B). Generally, the estimated treatment effects are larger compared to those estimated using the non-incentivized test (Panel A). In the first year of the program our composite measure of learning was .17σ higher (p-value < 0.01) in Levels schools relative to the control group, and .059σ higher in Pay for Percentile schools but this was not statistically significant (p-value .28, see Column 3). In the second year, learning was .22σ higher (p-value < 0.01) in Levels schools and .13σ higher (p-value .027) in Pay for Percentile schools.[23]

The estimated treatment effects (on the incentivized test) for Levels schools are comparable with those found in previous RCTs in India and Mexico (Muralidharan & Sundararaman, 2011; Behrman, Parker, Todd, & Wolpin, 2015). The estimated effects for the Pay for Percentile design are lower than those found in Loyalka et al. (in press), but larger than those in Gilligan et al. (2018).[24] In addition, the results suggest that the Levels design outperforms the Pay for Percentile design. Focusing on the composite test scores, the estimated differences between the incentive designs (α_3 and β_3 in Columns 3 and 6) are always negative (i.e., Levels outperforms Pay for Percentile), and statistically significant in three out of four cases.

[23] The treatment effects on threshold specific pass rates are shown in Tables D.4- D.7 in Appendix D.

[24] For the full sample Gilligan et al. (2018) find that Pay for Percentile incentives have a small (0.01σ) and statistically insignificant effect on learning. However, there is important heterogeneity in treatment effects. In schools with books, Pay for Percentile incentives improve learning outcomes by 0.11σ on the grade-relevant sub-test. In schools without books, there is no significant treatment effect on learning.

The larger treatment effects found in the incentivized test are likely driven by test-taking effort, where teachers had incentives to motivate their students to take the tests seriously. The importance of student test-taking effort has been documented in other settings such as an evaluation of teacher and student incentives in Mexico City (Behrman et al., 2015). As discussed in Section 3.2, administration of the incentivized test was tightly controlled by our implementation team. This mitigates any concerns about outright cheating. Assuming that all the differences between our incentivized and non-incentivized results are driven by test-taking effort, student effort can increase test score results between 0.016σ and 0.11σ (see Panel C). This is generally in line with the findings of Gneezy et al. (2017) and Levitt et al. (2016).

Given the reward structure, teachers in both treatment arms were motivated to ensure that their students took the incentivized test. There were no incentives to exclude academically-weaker students because learning gains from all students would be rewarded. In the second year of the study, teachers in the Levels schools were able to increase student participation in the incentivized test by 5 percentage points. Their counterparts in Pay for Percentile schools increased participation by 3 percentage points (see Table D.8 in Appendix D). Following Lee (2009), we compute bounds on the treatment effects by trimming the excess test takers from the left and right tails of the incentivized test distribution. Focusing on the year-two results for brevity, the 95% confidence interval for the treatment effects from this bounding exercise for math is 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 D.9 in Appendix D).

As discussed previously, we had limited information to properly group grade 1 students in Pay for Percentile schools. As this may limit the effectiveness of the Pay for Percentile scheme, we examine the effects of our interventions by focusing on grade 2 and 3 students, where we are able to appropriately group most students by ability. Our results are generally robust to this sample restriction (see Table D.1 in Appendix D.).

[Table 3 about here.]

4.2 Grade repetition

Cross-country comparisons reveal there is a negative correlation between income per capita and the grade repetition rate in primary school (Manacorda, 2012). Grade repetition is commonplace in developing countries, and is thought to impose significant individual and social costs, such as an increase in the probability that a student drops out of school (Manacorda, 2012).

In Tanzania, the introduction of the 3R (Reading, wRiting and aRithmetic) curriculum in 2015 was accompanied by a change in grade repetition policy in grades 1, 2 and 3. Promotion is no longer automatic and pupils who do not master basic skills can be forced to repeat, based on a decision by the school committee (automatic promotion remains in place after grade 3). Thus, we can use grade repetition as an additional outcome measure of learning in early grades.

We examine the impact of both treatments on grade repetition in Table 4. In 2015, the first year of both the incentive program and the new retention policy, we do not find any statistically significant changes in repetition rates in Levels or Pay for Percentile schools (Column 1). At the end of the second year, repetition rates in Levels schools were 3.3 percentage points lower than the control group (p-value .048), a 24 percent reduction. Among students in Pay for Percentile schools there was a small positive and statistically insignificant effect on grade repetition. Formal hypothesis tests show that the estimated reduction in repetition in Levels schools was significantly lower (p-value .041) compared to the estimated change in Pay for Percentile schools. Given that repetition rates reflect academic performance, this provides additional evidence that the Levels system leads to greater learning improvements than the Pay for Percentile system.

[Table 4 about here.]

4.3 Spillovers to Other Grades and Subjects

As the teacher incentives only covered numeracy and literacy in grades 1, 2, and 3, a potential concern is that teachers and schools focus on these grades and subjects to the detriment of other grades and subjects. For example, schools may shift resources such as textbook purchases from higher grades to grades 1, 2, and 3. In addition, teachers may cut back on teaching non-incentivized subjects such as science. On the other hand, if our incentive programs improve literacy and numeracy skills, they may promote student learning in other subjects and these gains may persist over time. In order to assess possible spillovers, we examine test scores 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 5, Panel A). In the first year of the program, composite test scores for grade 4 students in Levels schools increased by .099σ (p-value .052) (Column 3). In Pay for Percentile schools, we find relatively small (-.027σ) and statistically insignificant (p-value .59) effects on composite test scores. Since we tested fourth-grade students and collected information on those students at baseline, we conjecture that fourth-grade teachers assumed they would be included in the incentives. As a result of this belief, they may have exerted effort in the first year, but not in the second year once their non-eligibility had been confirmed. This type of spillover was also documented by Kremer, Miguel, and Thornton (2009), where a student incentive program for girls improved the performance of non-eligible boys who believed they would also benefit from the program.[25]

As third graders in the first year of our program transitioned to the fourth grade in the second year of the program, the fourth-grade results in the second year suggest that the learning gains from both incentive programs fade over time (Table 5, Panel A, Columns 4 to 6).

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 that are not directly incentivized (see Table 5, Panel B).

[Table 5 about here.]

4.4 Teacher Effort

Since the treatments were designed to elicit teacher effort, in this section we examine teacher responsiveness to the incentives. We use teacher presence in school and in the classroom as broad measures of teacher effort. Teacher presence was measured by our survey team and was collected shortly after our team arrived at a school in the morning. Overall, we do not find any effect in this dimension of teacher effort across our treatments (see Table 6, Panel A). We use data from student reports to examine additional dimensions of teacher effort in Panel B. According to students, teachers do not

[25] We also test for any spillover effects on the seventh grade primary school national exit exam (PSLE). We do not find any evidence that our incentives affected students performance on those tests. See Table D.10 in Appendix D.

provide any extra support in response to either treatment (Panel B, Column 1). Relative to the control group, teachers do not assign more (or less) homework to students in either treatment. However, there are diverging patterns in the estimated coefficients. Students in the Levels treatment report a small (but statistically insignificant) increase in homework, while students in Pay for Percentile schools report a small (but statistically insignificant) reduction in homework. Formal hypothesis tests show that the difference in these estimates is statistically significant (p-value .065), suggesting that teachers assigned relatively more homework in Levels schools (Panel B, Column 2). There is also evidence that the treatments altered the interactions between students and teachers. Students in Levels schools were 8 percentage points (p-value .033) more likely to report that the teacher called them by name in class. Students in Pay for Percentile schools reported a 4.7 percentage point (p-value .14) increase. Corporal punishment is very common in Tanzania. Almost 40 percent of students in the control group report had experienced some form of corporal punishment during the school year. Our results suggest that both incentives reduced incidences of corporal punishment. Students in Levels schools reported a 3 percentage point reduction (p-value .39) in corporal punishment. The reported reduction in Pay for Percentile schools was 6.1 percentage points (p-value .056). Overall, the results suggest that teacher incentives can foster a more positive learning environment in the classroom.

[Table 6 about here.]

In addition, we measure teacher effort using two sets of classroom observations. First, we conducted "external classroom observations," where our survey teams observed teacher behavior by standing outside the classroom for several minutes to prevent disruptions.[26] We also conducted within-classroom observations following the World Bank Service Delivery Indicator protocols. However, in-class observations are often affected by Hawthorne effects, which can reduce the usefulness of these protocols (Muralidharan & Sundararaman, 2010). Even though the external observations are less detailed, they are arguably better able to capture broad measures of teacher behavior because they are not affected by Hawthorne effects. We therefore focus on measures from the external observations and present the results from the in-class observations in Table D.11 in Appendix D.

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

Our findings using the external observations are shown in Table 7. For brevity, we focus on the estimated differences between the two incentive systems reported in the bottom row (α_3). We do not find any statistically significant differences in the likelihood that teachers were observed to be actively teaching, although the point estimates are larger for Levels teachers (Column 1). Teachers in Pay for Percentile schools were 2.2 percentage points (almost 50 percent) less likely to be engaged in classroom management activities (such as taking attendance or disciplining students) compared to Levels teachers (Column 2). Teachers in Pay for Percentile schools were also 7.7 percentage points (29 percent) more likely to be off-task or engaged in unrelated activities such as reading a newspaper or sending a text message (Column 3). Finally, we do not observe differences between the two incentives in distracted or off-task students, although the coefficient on Pay for Percentile schools shows a larger reduction in student distraction (Column 4).

[Table 7 about here.]

4.5 Heterogeneity by Student Characteristics

In this section, we explore the heterogeneity in treatment effects across the distribution of student baseline composite test scores in Figure 2 (for the non-incentivized tests) and Figure 3 (for the incentivized tests).[27]

In the first year of the program, we find suggestive evidence that teachers in both systems focused their attention on the best students. This pattern is more pronounced in Pay for Percentile schools where we can reject that the estimated learning gains are the same for all quintiles (p-value 0.016). In the second year of the program, the treatment effects were more balanced across the distribution of students and we fail to reject the hypothesis that the treatment effects of each quintile are equal.[28] Our first year results for the Pay for Percentile treatment are in line with Gilligan et al. (2018) who find that learning gains were greater for above median students, especially in schools with books. Our second year results for the Pay for Percentile treatment are in line with Loyalka et al. (in press) who find learning gains across the entire distribution of students.

[Figure 2 about here.]

[27] We also explore heterogeneity by additional student characteristics such as gender, as well as school characteristics such as pupil teacher ratio, and find limited evidence of heterogeneity in those characteristics (see Tables D.12 and D.13 in Appendix D for details).

[28] Subject specific results are available in Figures D.1 - D.4 in Appendix D.

[Figure 3 about here.]

4.6 Heterogeneity by Teacher Characteristics

Empirical evidence shows that women are more averse to competition and exert relatively less effort than men in competitive situations such as rank-order tournaments (Niederle & Vesterlund, 2007, 2011). However, we do not find any significant heterogeneous treatment effects by gender (Table 8, Column 1). We also do not find any heterogeneous effects by teacher age, which proxies for experience.

Although previous studies (e.g., Metzler and Woessmann (2012) and Bietenbeck, Piopiunik, and Wiederhold (2018)) 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, measured by our survey team through math and word association tests (Column 3). More effective teachers, as measured by the head teacher's performance rating, were more responsive on average to the Levels incentives compared to teachers in Pay for Percentile schools (Column 4). Teacher's who were more confident in their teaching abilities responded more to both incentives (Column 5).

[Table 8 about here.]

4.7 Why are treatment effects different across both programs?

We examine potential mechanisms that could drive differences in behavior and outcomes between the two types of incentives, with a particular focus on differences in the incentive structures of the two systems. For instance, the Levels system is easier to understand and could provide clear learning targets for classrooms, relative to the Pay for Percentile system. This difference in clarity could also affect teachers' expectations about their potential rewards from the incentive programs, which would ultimately affect the level of effort exerted.

Teachers understand both programs. Complex teacher incentive programs may be less effective if teachers cannot understand the program details and therefore do not optimally allocate their effort (Goodman & Turner, 2013; Loyalka et al., in press). These concerns are potentially more important in contexts with weak management capacity, which may be less able to effectively disseminate the details of a complex incentive program to teachers. Because the Pay for Percentile system is more complex, our results may reflect differences in teacher understanding of the incentive systems. To reduce these concerns, we developed culturally appropriate materials, including Q&A formats, examples, and illustrations, which we used to communicate the details of the incentive program to teachers. For example, in Pay for Percentile schools we explained that students would be grouped into separate contests based on their initial abilities, ensuring that each contest would be fair. To make our explanation clear, we used an analogy of a footrace. We explained that a race featuring one fast runner competing against slower opponents would not be fair. A fairer system would group runners into separate races based on their speed.[29]

During baseline and midline school visits, teams reinforced teachers' familiarity with the main features of the program. During our visits, we tested teachers to ensure they understood the details of the incentive program they were assigned to. We then conducted a review session to discuss the answers to the test questions to further ensure that teachers understood the design details. Because we asked different questions during each survey round (baseline, midline and endline), we cannot compare the trends in understanding over time. However, despite the lack of temporal comparability, teacher comprehension was generally high and roughly equal across both types of incentive programs. For example, at the end of the second year 70% of teachers in Level schools knew that the amount of money paid per skill obtained by their students depended on the total number of students that pass across Tanzania. Over 90% of teachers in Pay for Percentile schools were aware that a student from a low ability group ranked at the top of his group at the end of the year would give him or her a larger bonus than a student in the highest ability group ranked low among their peers.[30]

Teachers expect higher earnings in the Levels system. Even though we equalized the budgets across treatments, it is still possible that teachers' beliefs about their potential earnings could differ across the two incentive systems. In the Pay for Percentile system, the fact that the final bonus payment depends on the relative performance of other teachers in schools across the country is more salient. Hence, teachers may be less

[29] 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.

[30] Although teacher understanding was relatively high, we also test for heterogeneity in treatment effects by teachers' understanding (at endline). We do not find any significant relationship between teacher understanding and student test scores. The results are shown in Table D.14 in Appendix D.

confident about their ability to receive large payouts compared to their peers in the Levels treatment, where payouts are determined by students' proficiency levels. Prior to payout of the bonuses, we collected data on teachers' earnings expectations from the incentives, as well as their beliefs about their performance relative to other teachers in the district. As these questions were only applicable to teachers in the incentive programs, we compare teachers in the Pay for Percentile arm to the Levels scheme, which serves as the omitted category in Table 9.

Teachers in Pay for Percentile schools had lower bonus earnings expectations compared to their peers in the Levels system. They expected almost 95,000 TZS (US$ 42) less in bonus payments than teachers in the Levels system. This represents an 18% reduction in bonus expectations relative to the mean expectations of teachers in the Levels system (Column 1) and 36% 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 about their probability to win relative to the probability of competitors winning.

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. Teachers were optimistic about their projected earnings: Only 9 percent of teachers expected to be among the bottom earners (Column 2) and 7 percent were worried about earning a low bonus (Column 5). On the other hand, 80 percent expected to be among the top earners in the district (Column 4).

[Table 9 about here.]

Goal setting is easier for teachers in the Levels system. In addition to being relatively easier to understand, the Levels system provides teachers with a clear set of 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.[31] Our surveys collected information on the professional goals that teachers had set for the academic year. In Table 10 we test whether the treatments affected teachers' goal

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

setting behavior. We do not find any differences in the likelihood of setting goals for the general school exams between teachers in the treatment schools and their counterparts in the control group (Column 1). However, teachers in the Levels system were almost 8 percentage points more likely to have set goals for the incentivized Twaweza test than control group teachers (Column 2). In contrast, teachers in Pay for Percentile schools were 2.5 percentage points (p-value .34) more likely to have set goals for the Twaweza test (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 on the incentivized (Twaweza) test. Our surveys also collected information about specific teacher goals on the Twaweza test. Because Twaweza tests were administered in all schools, we were able to collect this information from teachers in treatment and control schools. Teachers in both types of incentives schools were approximately 7 percentage points more likely to set a general goal (e.g., "I want my students to pass") for the test than teachers in control schools (Column 3). Additionally, teachers in Levels schools were almost 10 percentage points more likely to set a specific numerical target (e.g., "I want 50 percent of my students to pass") for the Twaweza incentivized test, compared to just under 4 percent of teachers in Pay for Percentile schools (Column 4). Although these differences are not statistically significant, the point estimates in Column 2 and 4 suggest greater incidences of goal-setting among teachers in the Levels design.[32]

[Table 10 about here.]

5 Cost-effectiveness

We use accounting records to examine the cost-effectiveness of our interventions, following the framework outlined in Dhaliwal, Duflo, Glennerster, and Tulloch (2013). The total annual cost of the teacher incentive programs was US$ 7.23 per student. This cost estimate includes both the direct costs (value of incentive payments) as well as the implementation costs (test design and implementation, communications, audit, transfer costs, etc.) of the program. However, the cost in the long run of the Pay for Percentile scheme

[32] This finding, in combination with the overall incentive effects, provides a link between our paper and a fast growing empirical literature that finds positive associations between quality of management practices (including provision of clear and well-defined targets, performance measurement and feedback, and setting performance related rewards and sanctions) and organizational goal achievement (Bloom, Lemos, Sadun, Scur, & Van Reenen, 2014).

is US$ 1.50 higher (US$ 8.73 total) due to pre-testing costs to determine ability groups.[33]

For each intervention, we use the treatment effect on the composite index in the incentivized test in the second year to compute cost-effectiveness. We focus on the incentivized test to facilitate comparability with other teacher incentive studies. Since the Pay for Percentile treatment effects is 0.13σ in the second year, the cost-effectiveness of the intervention is 1.48σ per US$ 100 spent per child. The Levels treatment effects is 0.22σ, implying a cost-effectiveness of 3.04σ per US$ 100 spent per child. These estimates suggest that both programs are cost-effective compared to several other interventions in developing countries analyzed in the overview by (Kremer, Brannen, & Glennerster, 2013). For instance, the Levels treatments, our intervention is more cost-effective than a computer-assisted learning program evaluated in India (1.54σ per US$ 100), but less effective than the incentive program on attendance in India (2.28σ per US$ 100).

6 Conclusion

We use a randomized controlled trial to compare the effectiveness of two different teacher incentive programs aimed at improving early-grade learning in Tanzanian public schools. Specifically, we compare the effectiveness of an innovative multiple-threshold proficiency incentive design relative to an ostensibly more sophisticated, rank-order tournament-style Pay for Percentile system in terms of their impact on student test scores.

We report two main findings. First, both programs lead to increases in test scores, compared to students in the control group. Second, despite the theoretical advantage of the Pay for Percentile system, our multiple threshold proficiency system was more effective at increasing test scores and reducing grade repetition than the Pay for Percentile system.

Our results demonstrate some of the theoretical and practical considerations facing education authorities interested in adopting teacher incentive programs. Although rank-order tournament schemes can provide powerful incentives to increase effort, such systems can be more opaque, making it harder for teachers to determine how to best exert effort. By contrast, the multiple threshold proficiency system used in this study com-

[33] The costs of pre-treatment testing required in Pay for Percentile for Grades 2 and 3 are not included in the cost figure, since this cost would only be incurred once (ability groups could be based on endline data after the first year of implementation). Our calculations also assume similar data management costs for both programs, even though in reality the Pay for Percentile data costs were higher due to tasks such as preparing the ability groups and programming the payment calculations. However, these are largely fixed costs and relatively small relative to the variable costs, especially at scale.

municates clear student-level targets. These salient targets provide teachers with clear signals about how to allocate their effort in the class. Since developing countries are often faced with implementation capacity constraints, the multiple threshold system may be particularly well suited for these contexts given its relative administrative simplicity. Further, such a system is arguably better suited for early grades, where the curriculum is focused on a narrower set of key learning milestones such as number recognition and subtraction. Consequently, this incentive system can serve as an important complement to "teaching at the right level" programs, and education reforms that scale back overly ambitious curricula in early grades (Cunningham, 2018).

An important caveat is that our results focus on short run outcomes. In the long run, concerns about gaming the system (e.g., teaching to the test) will be greater. Since rank-order tournaments (such as Pay for Percentile) allow education systems to use different tests and test-formats, they can minimize these concerns if administrators have the willingness and capacity to implement such testing changes. Longer run studies conducted at scale will be needed to better understand the long run advantages and disadvantages of different teacher incentive systems.

References

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

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

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

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools. *Journal of Political Economy*, *123*(2), 325-364. Retrieved from https://doi.org/10.1086/675910 doi: 10.1086/675910

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

Bietenbeck, J., Piopiunik, M., & Wiederhold, S. (2018). Africa's skill tragedy does teachers' lack of knowledge lead to low student performance? *Journal of Human Re-*

sources, *53*(3), 553–578.

Bloom, N., Lemos, R., Sadun, R., Scur, D., & Van Reenen, J. (2014). The new empirical economics of management. *Journal of the European Economic Association*, *12*(4), 835–876.

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

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

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

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

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

Cunningham, R. (2018). *Unicef think piece series: Curriculum reform.* Retrieved from https://www.unicef.org/esaro/EducationThinkPieces_5_CurriculumReform.pdf (UNICEF Eastern and Southern Africa Regional Office, Nairobi)

Dalton, P. S., Gonzalez, V., & Noussair, C. N. (2015). *Paying with self-chosen goals: Incentives and gender differences.* (CentER Discussion Paper Series No. 2016-036)

de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. (2018). Double for nothing? experimental evidence on an unconditional teacher salary increase in indonesia*. *The Quarterly Journal of Economics*, *133*(2), 993-1039.

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.

Fryer Jr, R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. *NBER Working Paper*(18237).

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., & Muralidharan, K. (2016). Chapter 10 - improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In S. M. Eric A. Hanushek & L. Woessmann (Eds.), (Vol. 5, p. 653 - 743). Elsevier.

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

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

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

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

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

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

Koch, A., & Nafziger, J. (2011). Self-regulation through goal setting. *The Scandinavian Journal of Economics*, *113*(1), 212-227. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9442.2010.01641.x doi: 10.1111/j.1467-9442.2010.01641.x

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

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

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

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

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

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

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

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

Manacorda, M. (2012). The cost of grade retention. *Review of Economics and Statistics*, *94*(2), 596–606.

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. (in press). Inputs, incentives, and complementarities in education: Experimental evidence from tanzania. *The Quarterly Journal of Economics*.

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.

Miller, G., & Babiarz, K. S. (2013). Pay-for-performance incentives in low-and middle-income country health programs. *NBER Working Paper*(No. 18932).

Mohanan, M., Donato, K., Miller, G., Truskinovsky, Y., & Vera-Hernández, M. (2019). Different strokes for different folks: Experimental evidence on the effectiveness of input and output incentive contracts for health care providers with varying skills. *NBER Working Paper*(25499).

Muralidharan, K., & Sundararaman, V. (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India. *Economic Journal*, *120*, F187–F203.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*(1), pp. 39-77. Retrieved from http://www.jstor.org/stable/10.1086/659655

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

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

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

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

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

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

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

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

Singh, P., & Masters, W. A. (2018). Performance bonuses in the public sector: Winner-take-all prizes versus proportional payments to reduce child malnutrition in india. *Journal of Development Economics*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0304387818300610 doi: https://doi.org/10.1016/j.jdeveco.2018.10.003

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

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

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

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

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

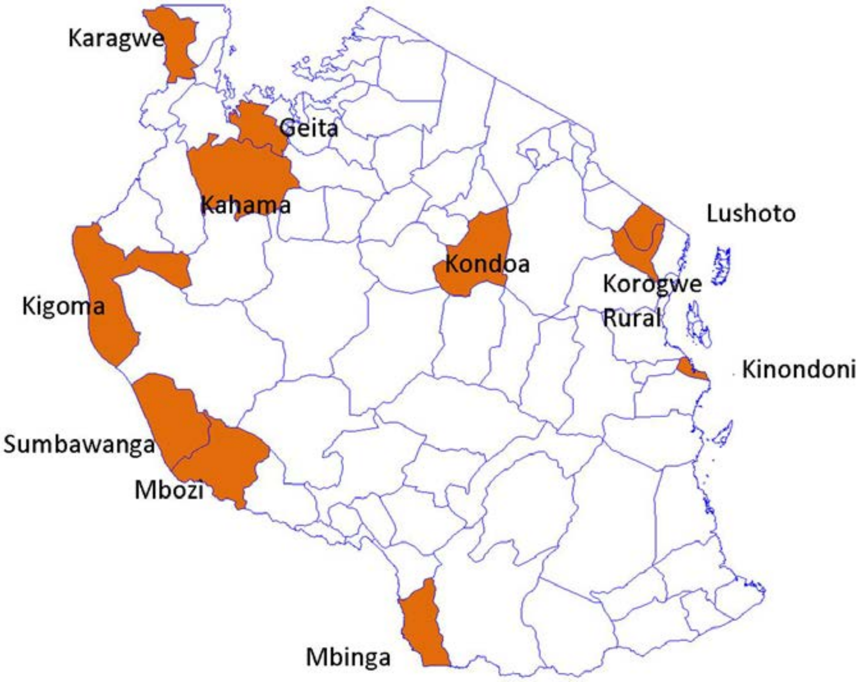
World Bank. (2014). *Service delivery indicators: Tanzania* (Tech. Rep.). The World Bank, Washington D.C.

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

World Bank. (2018). *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

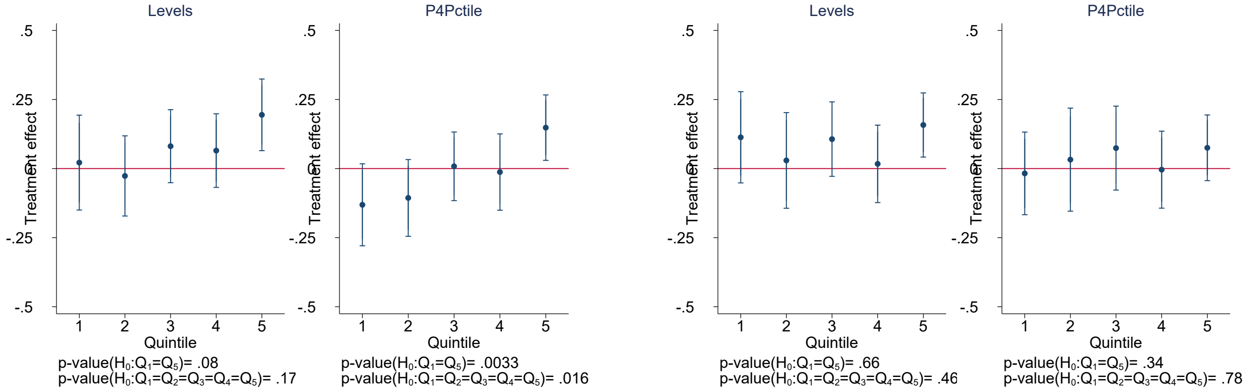
Figures

Figure 1: Districts in Tanzania from which schools are selected



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

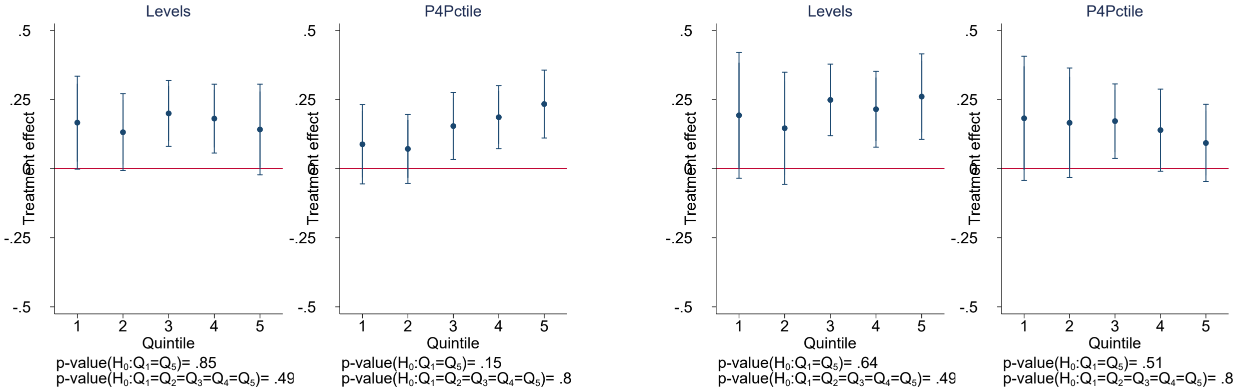
Figure 2: Composite — non-incentivized



(a) Year 1

(b) Year 2

Figure 3: Composite — incentivized



(a) Year 1

(b) Year 2

Tables

Table 1: Skills tested in the Levels schools

Kiswahili	**English**	**Math**
Grade 1		
Letters (TZS 1,992)	Letters (TZS 5,838)	Counting (TZS 513)
Words (TZS 1,619)	Words (TZS 14,749)	Numbers (TZS 750)
Sentences (TZS 2,057)	Sentences (TZS 58,267)	Inequalities (TZS 649)
		Addition (TZS 748)
		Subtraction (TZS 821)
Grade 2		
Words (TZS 1,192)	Words (TZS 5,071)	Inequalities (TZS 803)
Sentences (TZS 1,297)	Sentences (TZS 12,076)	Addition (TZS 1,136)
Paragraphs (TZS 2,214)	Paragraphs (TZS 61,938)	Subtraction (TZS 1,374)
		Multiplication (TZS 1,732)
Grade 3		
		Addition (TZS 694)
Story (TZS 1,709)	Story (TZS 36,250)	Subtraction (TZS 900)
Comprehension (TZS 1,530)	Comprehension (TSZ 22,63)	Multiplication (TZS 3,660)
		Division (TZS 1,820)

This table shows the skills tested in each subject and grade. In parentheses are the pay teachers received in the first year for each student that masters each skill. English payments are higher since the overall pass rate is much lower. In 2016, English instruction was removed from the curriculum in grades 1 and 2 and therefore dropped from the skills tests. See Section 2.3 for details.

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

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

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

Table 3: Effect on test scores

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: Non-incentivized						
Levels (α_1)	.044	.051	.057	.071*	.098*	.096**
	(.044)	(.05)	(.048)	(.039)	(.053)	(.046)
P4Pctile (α_2)	-.0099	-.044	-.029	.077**	.0045	.044
	(.038)	(.04)	(.039)	(.037)	(.05)	(.044)
N. of obs.	4,781	4,781	4,781	4,869	4,869	4,869
$\alpha_3 = \alpha_2 - \alpha_1$	-.054	-.095**	-.085*	.0064	-.094*	-.052
p-value ($H_0 : \alpha_3 = 0$)	.2	.047	.053	.89	.078	.29
Panel B: Incentivized						
Levels (β_1)	.11**	.13***	.17***	.14***	.18***	.22***
	(.047)	(.048)	(.064)	(.045)	(.046)	(.059)
P4Pctile (β_2)	.066*	.017	.059	.093**	.085*	.13**
	(.039)	(.043)	(.054)	(.04)	(.045)	(.056)
N. of obs.	48,077	48,077	48,077	59,680	59,680	59,680
$\beta_3 = \beta_2 - \beta_1$	-.047	-.11**	-.11*	-.044	-.093**	-.096*
p-value ($H_0 : \beta_3 = 0$)	0.30	0.026	0.070	0.31	0.045	0.097
Panel C: Incentivized – Non-incentivized						
$\beta_1 - \alpha_1$	.06	.069	.1	.06	.07	.11
p-value($\beta_1 - \alpha_1 = 0$)	.16	.13	.045	.13	.13	.025
$\beta_2 - \alpha_2$	.073	.056	.082	.016	.076	.077
p-value($\beta_2 - \alpha_2 = 0$)	.09	.2	.12	.68	.078	.11
$\beta_3 - \alpha_3$	.013	-.012	-.021	-.044	.0059	-.037
p-value($\beta_3 - \alpha_3 = 0$)	.76	.8	.7	.28	.91	.49

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

Table 4: Effect on grade repetition

	(1) Year 1	(2) Year 2
Levels (α_1)	-.0097	-.033**
	(.02)	(.016)
P4Pctile (α_2)	.025	.0014
	(.017)	(.014)
N. of obs.	4,781	4,869
Mean control	.13	.14
$\alpha_3 = \alpha_2 - \alpha_1$	.035*	.034**
p-value ($H_0 : \alpha_3 = 0$)	.062	.041

Results from estimating Equation 1 for whether a student is in a lower grade than expected at the end of the first year (Column 1) and at the end of the second year (Column 2). Control variables include student characteristics (age, gender, grade and lag test scores) and school characteristics (PTR, Infrastructure PCA index, a PCA index of how close the school is to different facilities, and an indicator for whether the school is single shift or not). Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Spillovers to other grades and subjects

Panel A: Grade 4

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Levels (α_1)	.16**	.055	.099*	.059	.042	.049
	(.068)	(.046)	(.05)	(.066)	(.057)	(.053)
P4Pctile (α_2)	-.026	-.027	-.027	-.0018	.00071	-.0035
	(.059)	(.049)	(.05)	(.063)	(.052)	(.051)
N. of obs.	1,513	1,513	1,513	1,482	1,482	1,482
$\alpha_3 = \alpha_2 - \alpha_1$	-.18***	-.082*	-.13**	-.061	-.041	-.052
p-value ($H_0 : \alpha_3 = 0$)	.0085	.087	.017	.33	.44	.32

Panel B: Science (Grades 1-3)

	Year 1	Year 2
Levels (α_1)	.069	.083
	(.063)	(.06)
P4Pctile (α_2)	-.0023	.079
	(.05)	(.057)
N. of obs.	4,781	4,869
$\alpha_3 = \alpha_2 - \alpha_1$	-.072	-.0042
p-value ($H_0 : \alpha_3 = 0$)	.25	.94

Results from estimating Equation 1 for grade 4 students (Panel B) and for grade 3 students in science (Panel A). Control variables include student characteristics (age, gender, grade and lag test scores) and school characteristics (PTR, Infrastructure PCA index, a PCA index of how close the school is to different facilities, and an indicator for whether the school is single shift or not). Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Treatment effects on teacher behavior

Panel A: Spot checks

	(1) In school	(2) In classroom
Levels (α_1)	-0.0065	0.015
	(0.038)	(0.039)
P4Pctile (α_2)	-0.0085	0.000017
	(0.032)	(0.036)
N. of obs.	360	360
Mean control	.7	.36
$\alpha_3 = \alpha_2 - \alpha_1$	-.002	-.015
p-value ($H_0 : \alpha_3 = 0$)	.96	.69

Panel B: Student reports

	(1) Extra help	(2) Homework	(3) Call by name	(4) Hit
Levels (α_1)	0.0080	0.017	0.080**	-0.030
	(0.0097)	(0.015)	(0.037)	(0.035)
P4Pctile (α_2)	-0.0022	-0.014	0.047	-0.061*
	(0.0091)	(0.015)	(0.032)	(0.032)
N. of obs.	18,563	18,563	9,557	9,557
Mean control	.062	.12	.5	.37
$\alpha_3 = \alpha_2 - \alpha_1$	-.01	-.032*	-.032	-.031
p-value ($H_0 : \alpha_3 = 0$)	.3	.065	.34	.35

Panel A presents teacher-level data on teacher absenteeism (Column 1), and time-on-task (Column 2). Panel B presents student-level data on teacher behavior (as reported by students) on extra help (Column 1), homework assignment (Column 2), calling by name (Column 3), and hitting/pinching/slapping students (Column 4). We pool the data for both years, except for calling students by name and corporal punishment, which was only collected in the second year. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: 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
$\alpha_3 = \alpha_2 - \alpha_1$	-.059	-.022**	.077*	-.016
p-value ($H_0 : \alpha_3 = 0$)	.2	.037	.082	.28

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

Table 8: Heterogeneity by teacher characteristics

	(1)	(2)	(3)	(4)	(5)
	Male	Age	IRT	HT Rating	Self Rating
Levels*Covariate	-0.011	-0.00036	0.026	0.10***	0.059**
	(0.057)	(0.0012)	(0.033)	(0.023)	(0.029)
P4Pctile*Covariate	0.0049	-0.00017	0.0053	0.057**	0.082***
	(0.054)	(0.0012)	(0.036)	(0.026)	(0.029)
N. of obs.	19,300	19,300	19,300	9,738	19,300

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

Table 9: Teachers' earning expectations

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

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

Table 10: Goal-setting

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

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

FOR ONLINE PUBLICATION

A Randomization Details

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

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

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

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

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

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

Table A.1: Treatment allocation

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

B Theoretical Framework

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

B.1 Basic Setup

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

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

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

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

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

The first order conditions for this problem are:

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

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

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

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

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

B.1.1 Pay for Percentile

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

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

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

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

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

In a symmetric equilibrium, then

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

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

B.1.2 Levels

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

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

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

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

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

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:

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

for all l.

B.1.3 Numerical Simulation Set-up

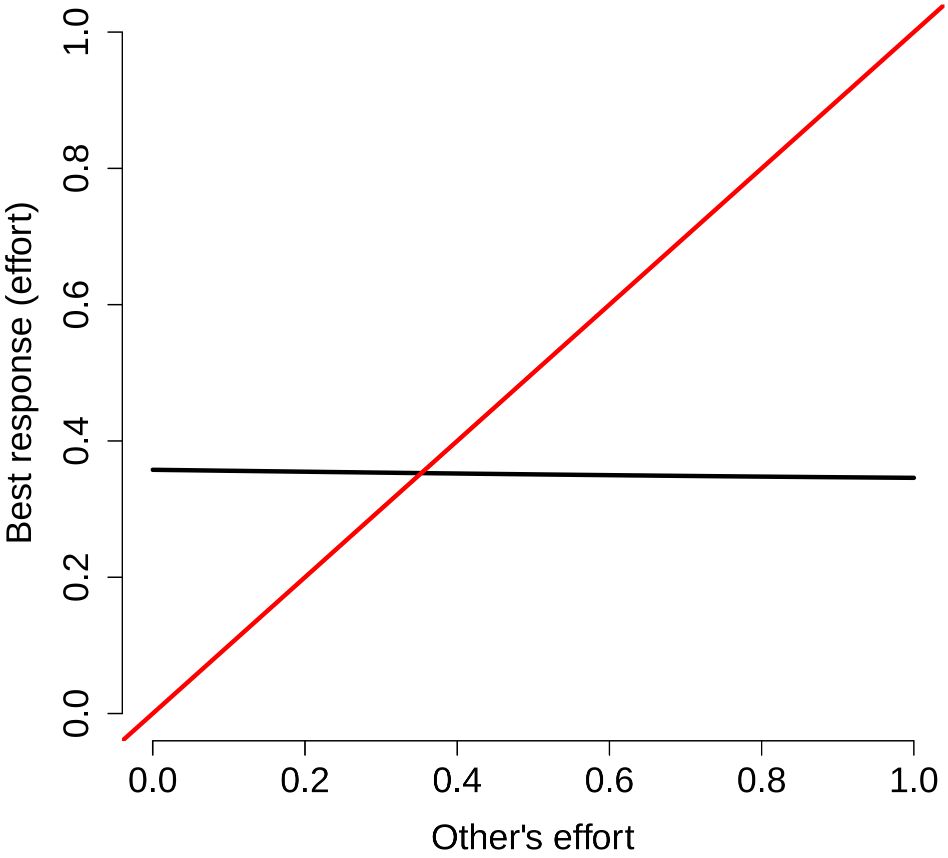
We simulate the equilibrium responses by teachers to both types of incentives in order to better understand teacher behavioral responses to the two treatments in our study. We assume that the teacher's cost function is quadratic (i.e., $c(e) = e^2$), and the shock to student learning follows a standard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

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

B.1.4 Levels Equilibrium

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

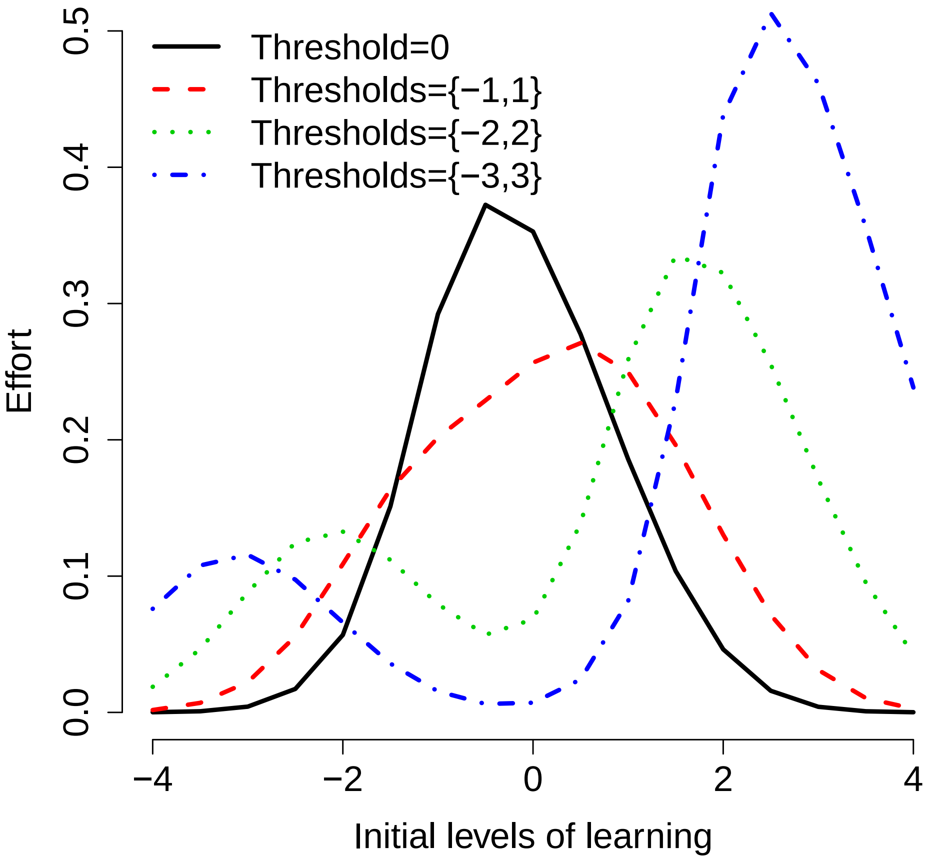
Figure B.1: Teacher $i's$ Best Response curve to other teacher's effort level



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

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

Figure B.2: Threshold Distance and Teacher effort



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

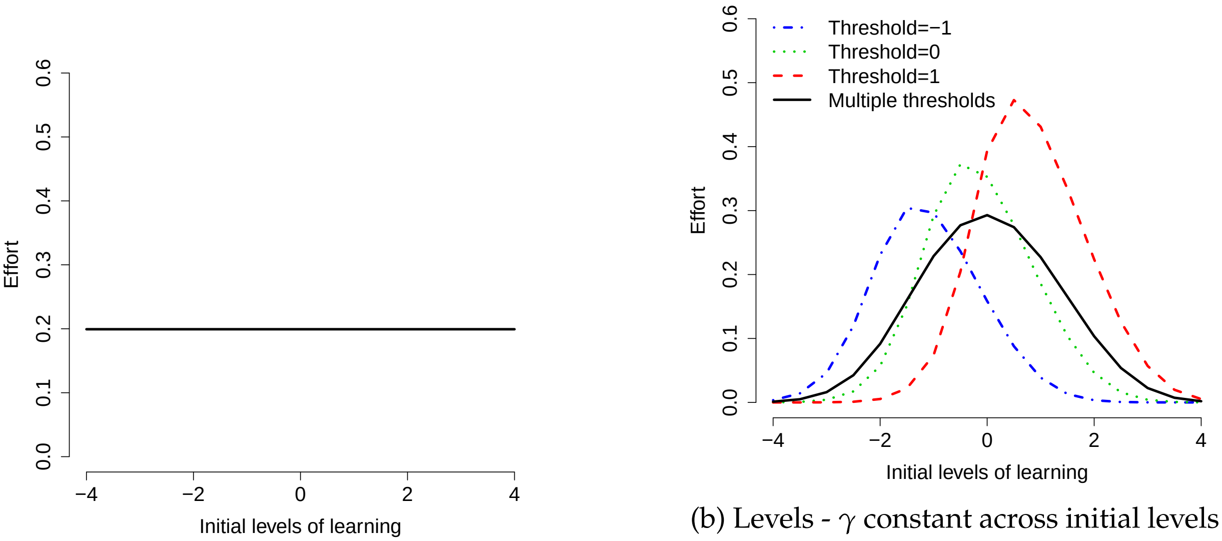
As the equilibrium behavior for teachers under Pay for Percentile was described in detail in Barlevy and Neal (2012), we refer our readers to consult their findings for additional insights.

B.1.5 A Comparison of Optimal Teacher Effort

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

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

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



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

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

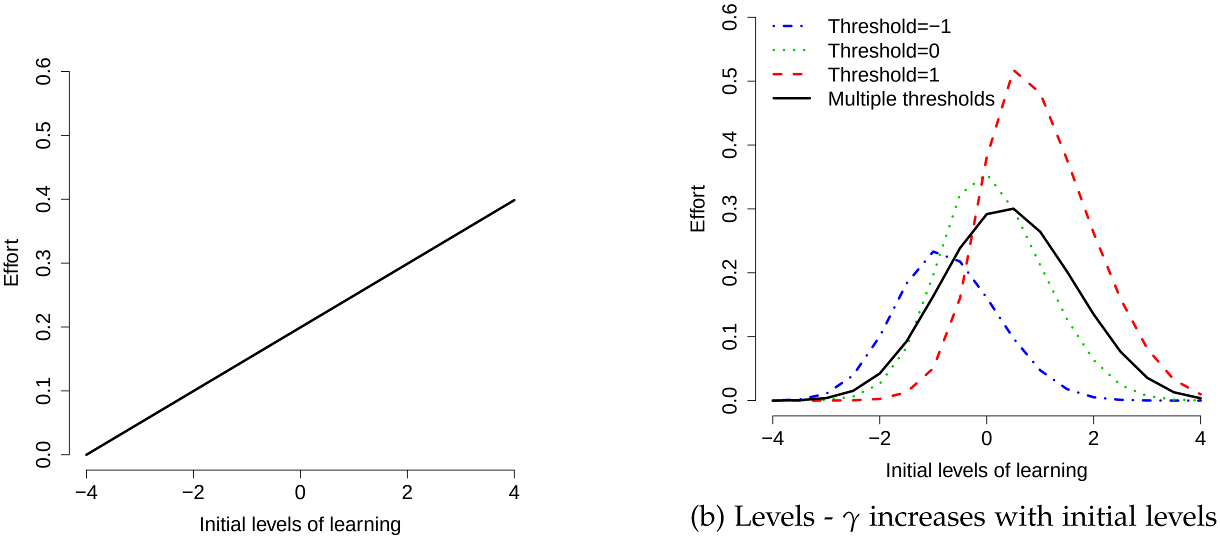
We relax the assumption of constant productivity of teacher effort and allow it to vary with initial learning levels of students. For simplicity, we specify a linear relationship between teacher productivity (γ^l) and student learning levels (a^l) such that $\gamma^l = 1 + 0.25a^l_{(t-1)}$.[38] Figures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

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

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

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

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



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

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

B.2 Robustness of Simulation Results

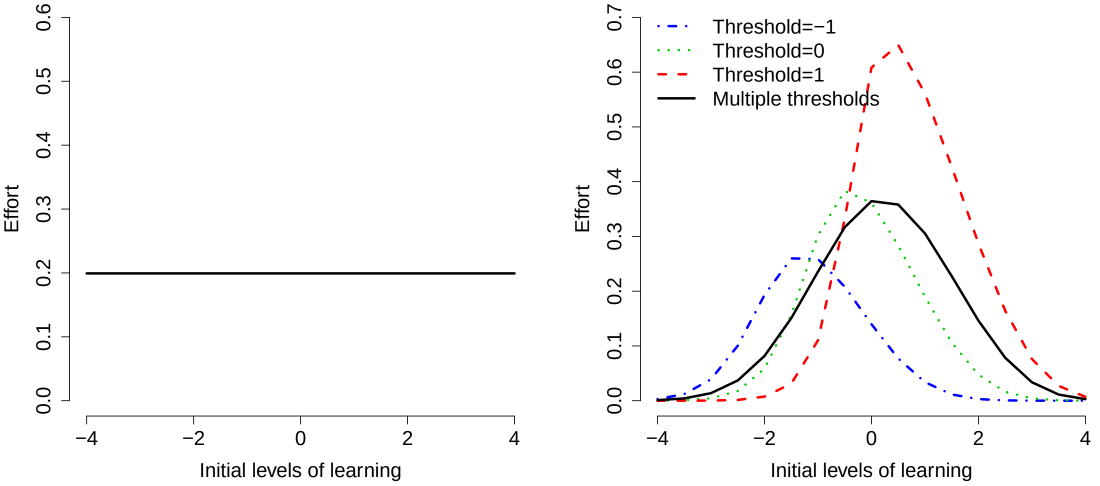
In this section we vary one of the central assumptions in our numerical simulations of the effort exerted by teachers in equilibrium discussed in Section B.1.5. In particular, we change the assumption that students are uniformly distributed across baseline test scores (recall that we had assumed student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

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

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

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

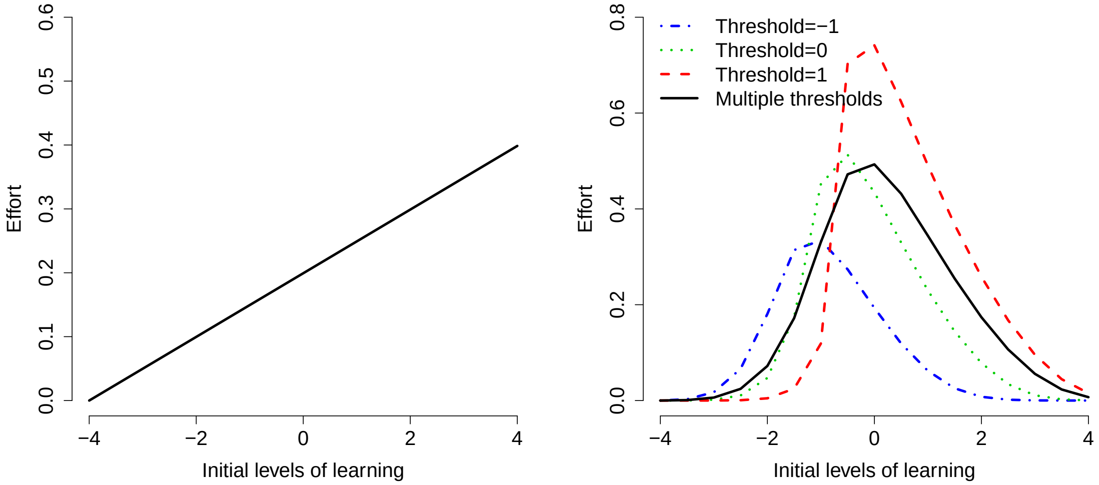
Figure B.5: Incentive design and optimal effort with constant productivity of teacher effort



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

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

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



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

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

C Test Design

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

C.1 Non-Incentivized test

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

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

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

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

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

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

C.2 Incentivized test

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

C.3 Comparability of tests

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

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

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

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

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

D Additional Tables

D.1 Properly seeded contests

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

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

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

D.2 Results for English

Table D.2: Effect on English test scores

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

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

D.3 Balance in Teacher Turnover

Table D.3: Teacher turnover

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

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

D.4 Pass Rates

Table D.4: Pass rates across all skill levels

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table D.9: Lee bounds for the incentivized test

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

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

D.6 National Assessments

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

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

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

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

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

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

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

D.7 Classroom observations

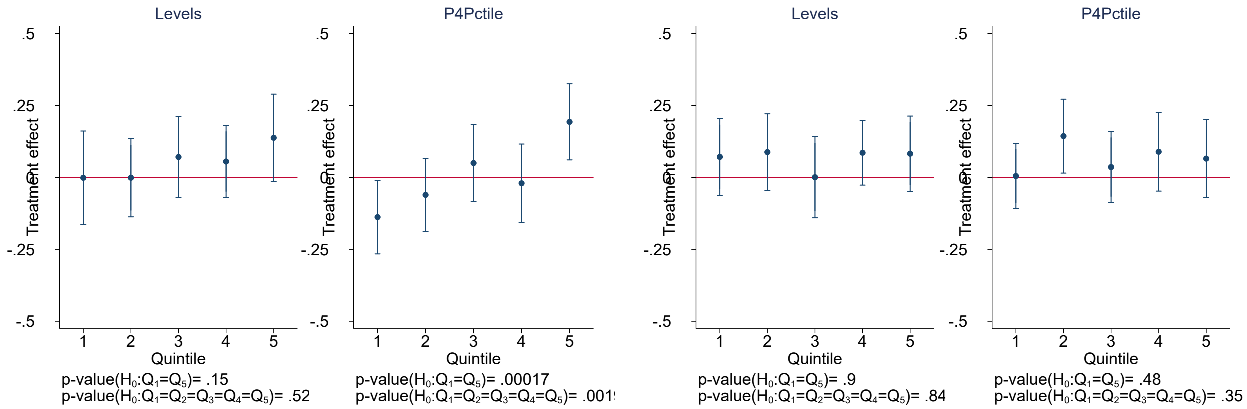
Table D.11: Classroom observations

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

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

D.8 Additional Heterogeneity in Treatment Effects

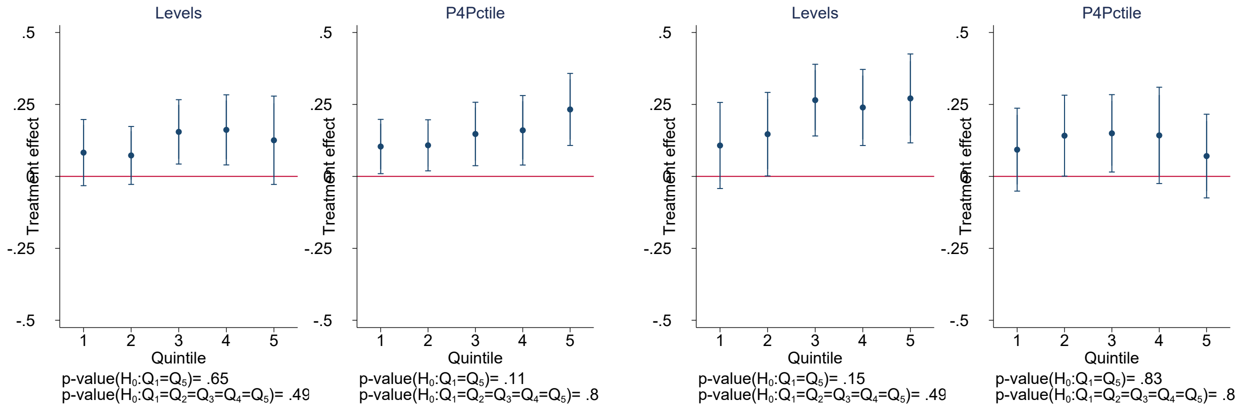
Figure D.1: Math — non-incentivized



(a) Year 1

(b) Year 2

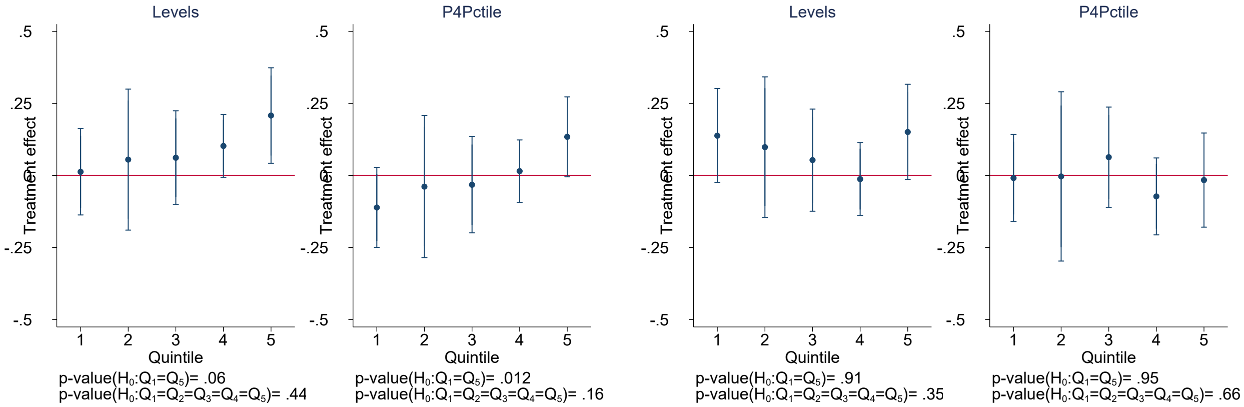
Figure D.2: Math — incentivized



(a) Year 1

(b) Year 2

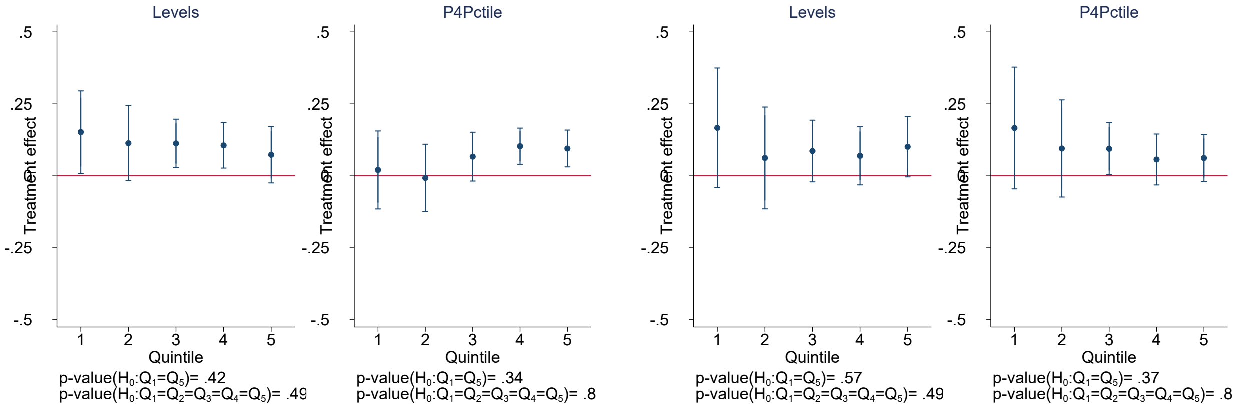
Figure D.3: Kiswahili — non-incentivized



(a) Year 1

(b) Year 2

Figure D.4: Kiswahili — incentivized



(a) Year 1

(b) Year 2

Table D.12: Heterogeneity by student characteristics

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

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

Table D.13: Heterogeneity by school characteristics

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

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

D.9 Teacher Understanding

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

Table D.14: Heterogeneity by teacher's understanding

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

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