Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania

Isaac Mbiti[1,2,5,6], Mauricio Romero[3,2,*], and Youdi Schipper[4]

Abstract:

We use a nationally representative field experiment in Tanzania to compare two teacher performance pay systems in public primary schools: a Pay for Percentile system (a rank-order tournament) and a "Levels" system that features multiple proficiency thresholds. Pay for Percentile can (under certain conditions) induce socially optimal effort among teachers, while Levels systems can encourage teachers to focus on certain students. Despite the theoretical advantage of the tournament system, we find that both systems improved student test scores across the distribution of initial learning levels after two years. However, the Levels system is easier to implement and is more cost-effective.

Keywords: teacher performance pay, pay for percentile, incentive design, Tanzania

Classification: C93, H52, I21, M52, O15

We dedicate this paper to the memory of our colleague Joseph Mmbando.

*__Correspondence address:__ Av. Camino a Santa Teresa 930, CDMX, Mexico; Correspondence e-mail: mtromero@itam.mx

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 the editor (Sule Alan), three anonymous referees, Allison Bigelow, Austin Dempewolff, Mitch Downey, David Evans, David Figlio, John Friedman, Delia Furtado, Guthrie Gray-Lobe, Scott Imberman, Ronak Jain, Joseph Mmbando, Molly Lipscomb Johnson, Terence Johnson, Michael Kremer, Derek Neal, Ömer Özak, Bobby Pakzad-Hurson, Wayne Aaron Sandholtz, Enrique Seira, Daniela Scur, Jay Shimshack, Bryce Millet Steinberg, Tavneet Suri, Rachel Walet, and seminar/conference participants at UC San Diego, Universidad del Rosario, NEUDC, PacDev, RISE, SOLE, SREE, and E-con of Education 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. We received IRB approval from Innovations for Poverty Action, UC San Diego, and University of Virginia. The Tanzania Commission for Science and Technology (COSTECH) also reviewed and approved the protocol. A randomised controlled trials registry entry and the pre-analysis plan are available at: https://www.socialscienceregistry.org/trials/1009. Replication data is available at Mbiti *et al.* (2022).

1 Introduction

Education systems in developing countries are typically characterised by weak accountability structures, which are often associated with low learning levels among students, limited observed levels of teacher effort, and inadequate institutional oversight and support for teachers (Chaudhury *et al.*, 2006; Mbiti, 2016; Bold *et al.*, 2017; World Bank, 2018). Because teachers play a central role in the education production function (Hanushek and Rivkin, 2012; Chetty *et al.*, 2014b,a), and a large share of national education budgets are devoted to their compensation, policymakers and researchers are increasingly interested in interventions that increase teacher effectiveness through more robust accountability measures. Teacher performance pay programs are seen as a potential policy response to improve accountability because they strengthen the links between teacher remuneration and student learning outcomes (World Bank, 2018; Bruns *et al.*, 2011).[1] Yet, the specific manner in which these programs link student performance to teacher pay varies greatly, ranging from simple proficiency threshold (or "bright line") designs to more complex value-added designs and rank-order tournaments (Imberman, 2015; Breeding *et al.*, 2021). However, there is limited evidence on how to best structure teacher incentives, especially in developing country contexts, which are less likely to have the requisite data management capacity to implement these schemes at scale (Breeding *et al.*, 2021).

Incentive schemes based on proficiency thresholds are commonly used in education systems and have well-known weaknesses and strengths. Research from the teacher incentive literature and the broader school accountability literature suggests they tend to favour teachers who serve students from wealthier backgrounds, encourage teachers to focus on marginal students — which could exacerbate inequality in learning outcomes — and present challenges for policymakers who are tasked with selecting the appropriate (learning) thresholds that trigger rewards (or punishments) (Figlio and Loeb, 2011; Neal, 2011; Macartney *et al.*, 2021). Despite these potential drawbacks, the widespread adoption of these systems suggests they may have some advantages compared to other incentive designs. In particular, threshold

[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 increased from 7.9% in 2004 to 11.3% in 2012 (an increase of over 43%) (Imberman, 2015). Less developed countries such as Brazil, Chile, and Pakistan have also implemented performance pay programs, often as large pilots (Alger, 2014; Ferraz and Bruns, 2012; Barrera-Osorio and Raju, 2017; Contreras and Rau, 2012).

(or "bright-line") designs may be well-suited for situations where the thresholds correspond to important objectives, such as curriculum goals. They also provide teachers with clear and salient targets. This can enable teachers to better react to the incentives and provides them with a clearer understanding of their performance (Brehm *et al.*, 2017) — this feedback loop may be critical in contexts where teacher capacity is relatively limited (see Bold *et al.* (2017) for a review of teacher capacity in sub-Saharan Africa).[2] Further, their transparency (and, thus, perceived fairness) could foster acceptance of the system among teachers (Fehr *et al.*, 2007). A practical (and cost-saving) advantage is that these schemes require students to be tested only once and do not require panel data management to link beginning and end of year student performance. This arguably makes them better suited to developing country contexts where there is an urgent need to strengthen the public sector's data management capacity and promote a culture of using data in decision-making and implementation (World Bank, 2017a).

Despite the popularity of incentive systems based on proficiency thresholds, insights from economic theory suggest that sophisticated teacher incentive designs, such as those based on rank-order tournaments, may induce greater — and potentially socially optimal — levels of effort among teachers than those based on proficiency thresholds (Lavy, 2009; Neal, 2011; Barlevy and Neal, 2012; Loyalka *et al.*, 2019).[3] In addition, empirical evidence from a recent meta-analysis suggests teacher incentive schemes in the US that featured rank-ordered tournaments outperform other designs (Pham *et al.*, 2021). Rank-order tournaments are also harder to game, and since the rewards are based on ordinal measures, they can more easily accommodate changes in exam formats and other system-wide changes (Neal, 2011). Yet, the theoretical advantages of tournaments may not materialise in practice if teachers find it difficult to determine how to react to such schemes (Charness and Kuhn, 2011).

We conducted a randomised experiment to examine the effectiveness (and cost-effectiveness) of two individual-level teacher incentive schemes in a nationally representative set of 180 Tanzanian public schools. We randomly assigned 60 schools to a "Levels" scheme with multiple proficiency thresholds

[2]In health-care settings, performance-based incentives with clear targets have been shown to have a positive effect on performance partly because these targets clarify what health workers' responsibilities and tasks are (Miller and Babiarz, 2014; Renmans *et al.*, 2016).

[3]The conditions under which Pay for Percentile, the rank-order tournaments we study, yields socially optimal levels of effort are outlined by Barlevy and Neal (2012). A key assumption is that the social planner maximises total returns from learning minus total cost. Another critical assumption is that the production function of human capital is linear in teacher effort and separable between the student's initial learning levels and other factors affecting learning.

that correspond to important curricular milestones.[4] We randomly assigned 60 schools to a "Pay for Percentile" (a rank-order tournament) scheme based on Barlevy and Neal (2012). The per-student bonus budget was equalised (ex-ante) across grades, subjects, and treatment arms to facilitate comparisons. The average teacher bonus was approximately 3.5% of the annual net salary (roughly half a month's pay).[5] We randomly assigned 60 schools to a control group. In all three groups, teachers were provided with baseline student reports so they were aware of each student's initial skill (or proficiency) level. Our main outcome is student performance on externally administered tests in math, Kiswahili, and English in first, second, and third grades.[6] Following Mbiti *et al.* (2019), we evaluate the incentive programs using data from both the incentivised (or "high-stakes") test administered to all students to determine teacher bonuses and a non-incentivised (or "low-stakes") test administered to a sample of students for research purposes.[7]

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. We included several thresholds to mitigate concerns that incentive programs using single-proficiency thresholds encourage teachers to focus on students close to the passing threshold. The skills thresholds were salient milestones based on the national curriculum, ranging from basic (e.g., number recognition) to more complex skills (e.g., multiplication) to allow teachers to earn rewards from a wide range of students. As reward payments for each skill were inversely proportional to the number of students that passed the skill, harder-to-pass skills were rewarded more.[8]

In the 60 schools assigned to the Pay for Percentile arm, students were first tested and assigned to one of several "baseline ability groups" based on their test scores. Teachers' rewards were proportional to their students' rankings within each group. The system does not penalise teachers who serve disadvantaged students because it explicitly accounts for the differences in initial student performance across teachers.

[4]To the best of our knowledge, this is the first documented implementation of proficiency-based teacher incentives with multiple (curricular-based) thresholds.
[5]Similar incentive sizes were used in Fryer (2013); Glewwe *et al.* (2010); Mbiti *et al.* (2019); Muralidharan and Sundararaman (2011); Lavy (2002); Ladd (1999); Vigdor (2008). See Leigh (2012) for details.
[6]English was dropped from the national curriculum in first and second grades during the experiment. Therefore, we focus on math and Kiswahili test scores. The analysis of English scores is in the Appendix.
[7]Both types of tests were conducted in control schools. However, the "incentivised" test results did not trigger payments in these schools.
[8]Since payments were determined ex-post based on pass rates, teachers faced uncertainty about the exact bonus sizes. However, an individual teacher's effort has a negligible effect on the aggregate pass rate and teachers likely have sufficient ex-ante information (e.g., through experience) to have reasonable predictions about the pass rates. We formalise this intuition in Appendix D.

Given the well-documented concerns about teachers misunderstanding incentive designs (Goodman and Turner, 2013; Fryer, 2013), we developed information packets that used culturally appropriate scripts and examples, and budgeted extra time to explain the design details.

We report two main findings. First, both incentives schemes improve learning outcomes compared to the control group, especially when we examine the results from the incentivised tests. Focusing on the results at the end of the second year of the program, the composite test scores of students in the Levels treatment were 0.22σ higher (p-value < 0.01) compared to the control. For students in the Pay for Percentile treatment, composite test scores were 0.13σ higher (p-value 0.027) compared to the control. For both treatments, gains were lower on the non-incentivised tests, but the magnitudes remained meaningful. For non-incentivised tests, composite tests scores were 0.095σ (p-value 0.036) and 0.041σ (p-value 0.33) higher in Levels schools and Pay for Percentile schools, respectively, when compared to control schools.[9] These learning gains in incentivised subjects were not at the expense of learning in other subjects: we do not find any evidence of negative treatment effects on science (which was non-incentivised).

Second, despite the theoretical predictions and empirical evidence on the effectiveness of tournament schemes relative to alternative designs, we find that composite test scores for the Levels incentive system increased (at least) as much as those in the Pay for Percentile system. At the end of the second year, the estimated treatment effect on the incentivised composite test score in Levels schools was 0.096σ higher (p-value 0.097) than the estimates for Pay for Percentile schools. The treatment effect on the non-incentivised composite test score shows a similar pattern, although the difference is smaller (0.053σ) and statistically insignificant (p-value 0.27).[10] Factoring in the administrative and implementation costs, our analysis shows that the Levels scheme is more cost-effective than the Pay for Percentile scheme. Although theory suggests that Pay for Percentile might produce more equitable learning gains relative to the Levels system, we find similar learning gains in the second year across all five quintiles of the student baseline test score

[9] As the test content was similar across tests, the differences in treatment effects are likely due to differences in student test-taking effort (Levitt *et al.*, 2016; Gneezy *et al.*, 2019). See Section 3.2 and Appendix E for details on the design and implementation of both tests.

[10] Year 2 results are arguably more informative because teachers learn how to better respond to the incentives over time.

distribution (using composite test scores) in both treatment arms. This suggests that multiple thresholds can mitigate the inequality concerns associated with proficiency systems.[11]

We use a comprehensive set of survey data from school administrators, teachers, and students and data from classroom observations to shed light on theoretically relevant mechanisms. Because measuring teacher effort and behaviour using these methods is challenging due to Hawthorne and/or John Henry effects, as well as social desirability bias (Muralidharan and Sundararaman, 2010; Muralidharan, 2017), we try to mitigate these concerns using additional measures such as external (to the classroom) observations of teacher behaviour. In addition, our enumerators examine a sample of student notebooks to document teacher feedback to students. Although we find no changes in teacher absenteeism in either treatment, data from the external classroom observations suggest that teachers in both treatment groups increased teaching time and reduced time off-task, although these effects were imprecisely estimated. Data from student self-reports suggest that the classroom dynamics were more conducive to learning. Teachers in both groups were more likely to call students by their names and were less likely to use corporal punishment, although these estimates were also noisy.[12] We also find that teachers in both systems understood the incentive designs. The high levels of teacher comprehension were partly due to our implementation and communication efforts. Further, teachers in both treatments had high expectations about their earning potential. Teachers expected to earn almost twice the actual average payment in both groups. However, teachers in the Pay for Percentile schools reported they expected to receive 18% lower bonus payments, on average, compared to their Levels counterparts. To the extent that expectations mirror effort, these patterns in expectations may also partly explain the increases in learning outcomes in both treatment groups and the instances where we find smaller treatment effects in the Pay for Percentile treatment relative to the Levels treatment.

[11] Theory predicts that if the productivity of teacher effort is constant across initial student learning levels, teachers will focus less on students in the tails of the ability distribution in proficiency systems. In contrast, they would focus on all students under a Pay for Percentile system (Barlevy and Neal, 2012). However, this difference between the designs becomes less pronounced when the productivity of teacher effort increases with students' initial learning levels. See Appendix D for more details.

[12] Muralidharan (2017) discusses the challenges of measuring teacher effort in the field, especially with limited research budgets. Many articles studying teacher incentives do not find any measurable response in teacher effort, even when they find treatment effects in student learning (e.g., Muralidharan and Sundararaman (2011); Loyalka *et al.* (2019)). This is perhaps unsurprising given that teachers can adjust effort on many margins that are difficult to measure (e.g., pedagogy, time-on-task, homework, socio-emotional support) in response to incentives linked to student learning. Advances and cost reductions in technology have made video recordings of classes more feasible. See Brown and Andrabi (2020) for an example.

Our study contributes to the debate about how to best structure and design teacher incentives. Breeding *et al.* (2021) find that most teacher incentive schemes yield little to no student learning improvements. Further, they suggest these mediocre results are often the consequence of design choices that mute the incentives. However, there is a limited set of adequately powered experimental studies comparing different teacher incentive designs. These comparisons include individual versus group incentives (Muralidharan and Sundararaman, 2011); bonuses based on average class test scores, value-added, and Pay for Percentile, all under a common rank-order tournament structure (Loyalka *et al.*, 2019); in-kind rewards versus public recognition (Barrera-Osorio *et al.*, 2022); and bonuses that featured a loss-aversion framing compared to a traditional bonus scheme (Fryer *et al.*, Forthcoming).[13] Generally, previous studies focused on comparisons featuring more intricate and (theoretically) effective incentives. In contrast, we compare a modified version of a commonly used proficiency design (Levels) that is theoretically less effective and a system generally considered to encourage much greater and potentially socially optimal levels of effort (Pay for Percentile). Contrary to the theoretical predictions, a multiple threshold system tied to foundational literacy and numeracy objectives (Levels) is more cost-effective at improving learning outcomes for all students in early grades compared to a more sophisticated, cost-equivalent (in terms of bonuses) rank-order tournament (or Pay for Percentile) scheme.[14] Our results reinforce the importance of testing theoretical predictions in real-world settings as practical constraints and other unforeseen circumstances can cause individuals to deviate from their predicted behaviour. They also highlight the importance of the practical limitations of tournaments outlined in Charness and Kuhn (2011) and shed light on the trade-offs faced by education authorities who have to consider the (cost-) effectiveness and feasibility of implementing different teacher incentive designs, often with limited information about the education production function.[15]

[13] A related literature compares the effectiveness of different incentive designs for healthcare providers (Singh and Masters, 2018; Mohanan *et al.*, Forthcoming).

[14] We are only aware of four studies in developing country contexts that specifically evaluate Pay for Percentile schemes. Loyalka *et al.* (2019) find that Pay for Percentile incentives increased test scores among math teachers in Chinese schools and that is more effective than rank-order tournaments based on test score levels or value-added measures — this is the only study we are aware of that compares Pay for Percentile to alternative incentive structures. Gilligan *et al.* (2019) find that Pay for Percentile has no impact on student learning in Ugandan schools, except for top students in schools with textbooks. Leaver *et al.* (2021) study the extensive (recruitment) and intensive (effort) effect of Pay for Percentile in the context of Rwandan primary teachers. On the intensive margin (the margin we study), Pay for Percentile increases teacher effort and improves student learning outcomes. Brown and Andrabi (2020) find that pay for percentile induces positive sorting of teachers in a network of private schools in Pakistan. Finally, we are only aware of one study in a developed country studying the impact of Pay for Percentile (Fryer *et al.*, Forthcoming).

[15] Our time frame does not allow us to examine how teachers and administrators would respond when they gained more experience with the incentive schemes (e.g., if the incentive schemes were permanently adopted by the government.)

2 Experimental Design

2.1 Context

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

Limited accountability within the education system is one driver of poor educational quality. Quality assurance systems (e.g., school inspectors) typically focus on superficial issues, rather than issues that may affect learning (Mbiti, 2016). Teacher absence rates further reflect the accountability vacuum. Data from unannounced spot checks shows that 14% of teachers were absent from school and only 47% of the teachers at the school were in the classroom (World Bank, 2015). Adding up school absence, classroom absence, and time spent on non-teaching activities, almost 50% of planned instructional time is lost each day (World Bank, 2015).

To address education quality concerns, Tanzanian teachers' unions have been actively lobbying for better pay. Yet, the correlation between teacher compensation and student learning is extremely low (Kane *et al.*, 2008; Bettinger and Long, 2010; Woessmann, 2011; de Ree *et al.*, 2018). Moreover, teacher salaries in 2016 were relatively high — approximately 500,000 TZS per month ($\sim$ US$250) or over 3 times GDP per capita

(World Bank, 2017b) — comprising about 60% of the education budget.[16] Despite Tanzanian teachers' relatively attractive wages, the teachers' union called a strike in 2012 to demand a 100% increase in pay (Reuters, 2012; PRI, 2013).[17]

2.2 Interventions and Implementation

The interventions in this study were developed in close collaboration with Twaweza, an East African civil society organisation 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).[18]

The KiuFunza program targets teachers in grades 1, 2, and 3 who are responsible for teaching Kiswahili, English, and math (arithmetic). A budget of US$150,000 per year for teacher and headteacher 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. In partnership with EDI (a Tanzanian research firm), Twaweza and a set of local district partners implemented all interventions. Headteachers were offered a bonus of 20% of the combined bonus of all incentivised teachers in their school.[19]

Within each intervention arm, Twaweza distributed information describing the program in early 2015 and 2016: first to teachers and headteachers, and 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-familiarise 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

[16]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, 2017b).
[17]In recent years, other teacher strikes to demand pay increases have occurred in South Africa, Kenya, Guinea, Malawi, Swaziland, Uganda, Benin, and Ghana.
[18]The first set of interventions under this program were launched in 2013, lasted until 2014, and were evaluated by Mbiti *et al.* (2019).
[19]Twaweza included headteachers 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 headteachers.

payments, it was considered "high-stakes" (from the teachers' perspective). Our non-incentivised research test was conducted on a different day but within a few weeks from the incentivised test. Both sets of tests were based on the Tanzanian curriculum. They were developed by Tanzanian education professionals, using formats of the Uwezo learning assessment framework.[20] We provide additional details about the design and implementation of both types of tests in Appendix E.

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. A necessary condition for Pay for Percentile to induce optimal effort is that teachers believe they compete in properly seeded (or fair) contests. To achieve this, the Pay for Percentile scheme 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. Teachers are then rewarded based on 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.

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 (see Appendix C.1 for details on the number and size of the groups). Students without test scores in second and third grade were grouped in an "unknown" ability group.[21] 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 school's historical average test scores. Thus, all first-grade students within a school were assigned to the same group.

[20] Uwezo learning assessments have been routinely conducted in Kenya, Tanzania, and Uganda since 2010.
[21] Roughly 20% of students are grouped into the "unknown" ability group. This includes newly enrolled students and students who were enrolled but were not tested at baseline for some reason.

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 scores. 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 subjects and all grades. Students without an endline test score were given a zero score and were ranked at the bottom of their ability group. Thus, there were no incentives to exclude academically weaker students.

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

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

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 ∼US$0.20 for both groups. Thus, in this example, if a student was in the top 1% of their ability group, their teacher would earn $99 * 0.2$ or US$19.8. 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 enrolment 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 under certain conditions, 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, teachers may find the Pay for Percentile system difficult to grasp. It presents each teacher with a series of tournaments (for each ability group in each subject that they teach); therefore, the bonus payoff is relatively hard to predict, even if the design guarantees a fair system. Furthermore, competing against teachers from schools across the country introduces uncertainty that 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 headteachers; however, they have well-known limitations. For example, they cannot 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 and Schanzenbach, 2010). To mitigate this concern, our Levels design features multiple thresholds ranging from basic skills to more advanced ones 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.[22] Miller and Babiarz (2014) argue that incentive designs based on "bright-line" performance thresholds (and goals) can be effective in helping service providers — in this case, teachers — to focus on achieving these goals. They also argue that bright-line designs are well suited to helping providers focus on achieving important outcomes.[23]

In Levels schools, teachers are paid in proportion to the number of skills that their students in grades 1-3 master in mathematics, Kiswahili, and English at the end of the school year. 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. The bonus per pass for each skill equals the skill budget divided by the number of students passing the skill. For example, suppose the budget allocated

[22] As discussed in Appendix D, a key practical challenge is ensuring that the thresholds are sufficiently close together to prevent teachers from ignoring students who fall between two thresholds. Appendix C.2 shows the passing thresholds are indeed spread across the ability distribution.

[23] In the health sector, Miller and Babiarz (2014) 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 these criteria.

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 per pass, that is, 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.[24]

Table 1: Skills tested in the Levels schools

Kiswahili	**Math**
Grade 1	
Letters (TZS 1,992 or ∼US$ 0.95)	Counting (TZS 513 or ∼US$ 0.24)
Words (TZS 1,619 or ∼US$ 0.77)	Numbers (TZS 750 or ∼US$ 0.36)
Sentences (TZS 2,057 or ∼US$ 0.98)	Inequalities (TZS 649 or ∼US$ 0.31)
	Addition (TZS 748 or ∼US$ 0.36)
	Subtraction (TZS 821 or ∼US$ 0.39)
Grade 2	
Words (TZS 1,192 or ∼US$ 0.57)	Inequalities (TZS 803 or ∼US$ 0.38)
Sentences (TZS 1,297 or ∼US$ 0.62)	Addition (TZS 1,136 or ∼US$ 0.54)
Paragraphs (TZS 2,214 or ∼US$ 1.05)	Subtraction (TZS 1,374 or ∼US$ 0.65)
	Multiplication (TZS 1,732 or ∼US$ 0.82)
Grade 3	
	Addition (TZS 694 or ∼US$ 0.33)
Story (TZS 1,709 or ∼US$ 0.81)	Subtraction (TZS 900 or ∼US$ 0.43)
Comprehension (TZS 1,530 or ∼US$ 0.73)	Multiplication (TZS 3,660 or ∼US$ 1.74)
	Division (TZS 1,820 or ∼US$ 0.86)

Note: This table shows the skills tested in each subject and grade. In parentheses are teachers' payments for each student who masters each skill in the first year.

[24]Enrolment at each school is on average 1.6% of total enrolment across Levels schools. Thus, the total number of a teacher's students passing the threshold has a negligible effect on the overall pass rate across schools. Hence, we can rule out teachers strategically choosing how many students to push over a threshold to maximise earnings.

2.2.3 Teacher understanding of the incentive designs

Teacher incentive programs may be ineffective if teachers cannot understand the program details and, therefore, do not optimally allocate their effort (Goodman and Turner, 2013; Loyalka *et al.*, 2019). These concerns are potentially more important in developing country contexts where public institutions may be less able to disseminate the details of an incentive program to teachers effectively.

During baseline and midline school visits, teams reinforced teachers' familiarity with the programs' main features. We developed culturally appropriate materials to enhance teachers' understanding of the incentive schemes, including Q&A formats, examples, and illustrations. 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 be unfair. A fairer system would group runners into separate races based on their speed.[25]

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 ensure that teachers understood the design details. Because we asked different questions about each incentive scheme during each survey round (baseline, midline, and endline), we cannot compare the trends in understanding over time or across treatments. However, despite the lack of 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 Levels schools knew that the amount of money paid per skill obtained by their students depended on the total number of students that passed 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 them a larger bonus than a student in the highest ability group ranked low among their peers.

[25]We worked closely with Twaweza's communications unit to develop our dissemination strategy and communications. The communications unit is experienced and highly specialised in developing materials to inform and educate the general public in Tanzania. Appendix F provides a copy of the material used to explain the interventions.

2.3 Conceptual Framework and Theoretical Predictions

We develop a simple model of student learning to highlight the differences in teacher effort under the two incentive schemes. For brevity, we first outline the main features of the model and then discuss the predictions. We discuss the details of the model in Appendix D.

We assume students' learning depends on their baseline learning level and their teacher's effort, which is costly to exert, and an idiosyncratic random shock. Individual students have different baseline learning levels, and all teachers are assumed to be equally skilled. Because our study compares the performance of the two incentive systems, in the model we assume teachers only respond to extrinsic (monetary) rewards for tractability. We impose three parametric assumptions to obtain sharper predictions about teacher effort under each incentive scheme. First, we assume that the cost of teacher effort is a quadratic function. Second, we assume that student baseline learning levels are uniformly distributed and verify that we obtain similar predictions if we were to impose a normal distribution instead. Finally, we assume that idiosyncratic random shocks to student learning are normally distributed. We then use simulations to study teachers' utility-maximising efforts under the different incentive schemes. In our framework, the utility-maximising effort is achieved when the marginal benefit (in terms of extra monetary rewards) equals the marginal cost of effort. We also study how teachers direct their effort toward different types of students, a process also referred to as "triage" (Loyalka *et al.*, 2019; Gilligan *et al.*, 2019).

We obtain two main predictions. First, our simulations predict that total teacher effort (and hence student learning) will be greater under Pay for Percentile than Levels. This result is partly driven by the fact that under the Levels system student baseline levels will affect teacher rewards, while they will not play an important role in the Pay for Percentile system if students are grouped into the separate rank-order tournaments appropriately. Both systems are predicted to improve teacher effort and student learning relative to the status quo in control schools if the rewards are sufficiently large to compensate for the costs of effort. This is consistent with the theoretical results of Barlevy and Neal (2012) and the empirical results of Loyalka *et al.* (2019).

Second, teacher "triage" depends on the correlation between the productivity of teacher effort and student baseline ability. Suppose the productivity of teacher effort is constant for all types of students. In that case, teachers under a Pay for Percentile scheme will (optimally) exert equal effort across the entire distribution of students. By contrast, teachers under the Levels system would (optimally) focus on students in the middle of the distribution and exert little effort at the tails. However, if (for example) the productivity of teacher effort is positively correlated with student baseline levels, then under Pay for Percentile teachers would optimally exert more effort towards better-prepared students. Under Levels, teachers would continue to optimally exert more effort towards students in the middle of the distribution and less effort at the tails.

Despite these predictions, there are several considerations that may influence how teachers respond to the incentives in practice. First, the monetary benefits must be sufficiently high to elicit behavioural responses from teachers. Second, both systems have to be appropriately designed. For instance, the groups (or contests) in the Pay for Percentile system must consist of students of similar ability, and the thresholds in the Levels systems cannot be too "far apart" or else teachers might ignore students who are in between these thresholds. Third, the incentives will be less effective if teachers cannot understand the incentives or the cognitive costs of understanding the rules are too high. Fourth, the incentives will also be less effective if teachers do not trust the incentive scheme. More transparent systems can potentially improve teacher understanding, their beliefs about the relationship between effort and rewards, and perhaps engender more trust in the scheme. Thus, although theory predicts that Pay for Percentile will improve learning more than Levels, the greater transparency of Levels may improve teachers' relative response to that system relative to Pay for Percentile. Further, teachers under Levels might be better able to respond due to the alignment of the Levels thresholds with important curriculum milestones.

2.4 A Note on Curriculum Reform and 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% 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 grades 1 and 2 in 2015 to allow teachers to focus on numeracy and literacy in Kiswahili. 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 grades 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 curriculum reform's implementation, it is unclear how to interpret the results for English. In addition, these estimates are less policy-relevant after the reform. Therefore, we only present mathematics and Kiswahili results in the main text to facilitate the analysis. In addition to this change, the curriculum reform prescribed that multiplication would be dropped from the grade 2 curriculum.

3 Data and Empirical Specification

3.1 Sample Selection

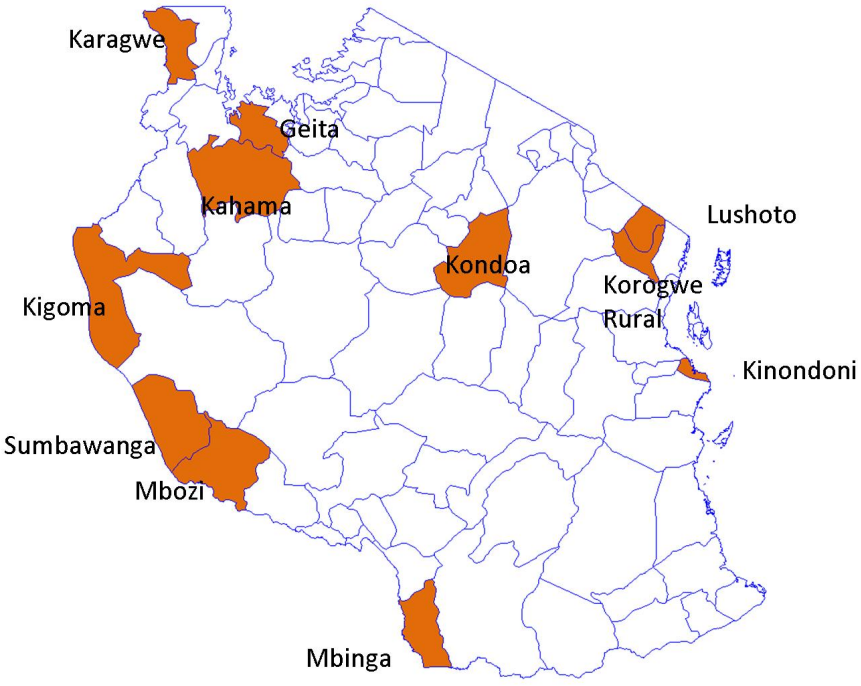
We evaluated the teacher incentive programs using a randomised design. The study was carried out in a nationally representative set of 180 Tanzanian public schools. These schools were part of a previous field experiment — studied by Mbiti *et al.* (2019) — where all students in grades 1, 2, and 3 were tested at the end of 2014. These tests provided the baseline student-level test score information required to implement the Pay for Percentile treatment.[26] As mentioned above, a necessary condition for the Pay for Percentile to

[26]These tests were administered before schools were assigned to the treatment groups discussed in this paper. Online appendices B and E provide more details.

deliver optimal levels of effort is that teachers believe they compete in fair contests. Thus, having reliable information about students' initial learning levels was key.[27]

The 180 schools in the sample are distributed across ten districts. The districts were randomly selected to participate in the experiment (see Figure 1), which provides external validity to our results across Tanzania (Muralidharan and Niehaus, 2017).[28] Within each district, we randomly allocated schools to one of our three experimental groups. Thus, six schools were assigned to the Levels treatment in each district, 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 field experiment and by an index of the overall learning level of students in each school. Further details are provided in Appendix B.

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).

[27] 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 would be fair, suggesting teachers think the contests are fair.

[28] 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.

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 headteacher, including facilities, management practices, and headteacher 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 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. Due to budget constraints, students in grade 4 in the first year were not tracked into grade 5. 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-incentivised student tests with the results from the incentivised tests used to determine teacher bonus payments and conducted in all schools, including control schools. Most articles studying teacher performance pay use incentivised tests to measure the overall treatment effects. However, it is unclear whether incentivised or non-incentivised tests are better for measuring treatment effects. We, therefore, present results from both tests for completeness.[29]

The incentivised and the non-incentivised tests covered very similar content (subject order, question type, phrasing, and difficulty level). The non-incentivised test had more questions for each subject to avoid bottom- and top-coding and 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

[29] As argued by Mbiti *et al.* (2019): "The confirmation that test-taking effort is a salient component of measured test scores by Levitt *et al.* (2016) and Gneezy *et al.* (2019) 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 'incentivised' testing procedure may be a better measure of true human capital."

different. Non-incentivised tests were administered during a regular school day by survey enumerators. In contrast, the incentivised 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 headteachers and teachers managing the flow of students. In addition, most schools used the incentivised test as the official end-of-year test. Several 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 was taken at baseline. Second, each student was assigned one test randomly selected out of ten test versions to prevent copying during the test. Finally, Twaweza teams handled, administered, and electronically scored tests without teacher involvement. Section E provides more details on the design and implementation of both tests.

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. In addition, the distribution of test scores is balanced across groups (Figure A.1 shows the CDFs of test scores are similar across groups). We were able to track 88% of students in the non-incentivised test sample at the end of the second year, with no differential attrition. On the incentivised tests, attendance in Levels and Pay for Percentile schools was higher in the second year (Table A.1). Thus, we present Lee (2009) bounds for the treatment effects on incentivised tests.

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 table 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 randomisation. The poverty index is the first component of a Principal Component Analysis (PCA) 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 (PCA) 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.3 Empirical Specification

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

$$Z_{isdt} = \delta_0 + \delta_1 Levels_s + \delta_2 P4Pctile_s + \delta_3 Z_{isd,t=0} + X_i\delta_4 + X_s\delta_5 + \gamma_d + \gamma_g + \varepsilon_{isdt}, \quad (2)$$

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 normalise 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.[30]

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

4 Results

In this section, we first explore how both incentive systems affected student test scores and grade repetition. We then examine whether the incentives increase observable teacher effort or change teacher behaviour. We then turn to heterogeneity by student and teacher characteristics. Finally, we explore possible mechanisms that could explain our results on test scores. Replication files are available at Mbiti *et al.* (2022).

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

4.1 Test Scores

We present the estimated treatment effects of the incentive programs on student learning using data from both the non-incentivised test (Table 3, Panel A) and the incentivised test (Table 3, Panel B). As discussed earlier, we focus our main analysis on math and Kiswahili due to the curriculum changes. We provide estimates of the intervention on English test scores in Table A.2 in Appendix A. To address multiple testing concerns, we also present estimates for a composite index of learning computed using an Item Response Theory model (Columns 3 and 6 in Table 3).

In the first year, both incentive schemes resulted in small but imprecisely estimated changes in test scores on the non-incentivised test. Focusing on the composite learning index (Panel A, Column 3), test scores increased by about 0.057σ (p-value 0.23) in Levels schools relative to the control group. Pay for Percentile schools scored 0-$.027\sigma$ (p-value 0.48) below control schools. In the second year of the program, the estimated treatment effects on the non-incentivised test are generally larger than the first-year estimates (Panel A, Columns 4-6). Test scores on the composite index increased by 0.095σ (p-value 0.036) in Levels schools and 0.041σ (p-value 0.33) in Pay for Percentile schools.[31]

Most of the existing literature on teacher incentives relies on data from incentivised tests that are used to determine teacher rewards (Muralidharan and Sundararaman, 2011; Fryer, 2013; Neal and Schanzenbach, 2010). Following this practice, we also present the treatment effects of our interventions using incentivised exams (Panel B). Generally, the estimated treatment effects are larger compared to those estimated using the non-incentivised test (Panel A). In the first year of the program our composite measure of learning was 0.17σ higher (p-value $0< 0.01$) in Levels schools relative to the control group and 0.059σ higher in Pay for Percentile schools, but this was not statistically significant (p-value 0.28, see Column 3). In the second year, learning was 0.22σ higher (p-value < 0.01) in Levels schools and 0.13σ higher (p-value 0.027) in Pay for Percentile schools.[32]

[31] Unlike other settings (e.g., Macartney (2016)), there are no dynamic incentives in either treatment. In the Levels scheme, payments do not depend on previous student performance. In the Pay for Percentile scheme, while the initial seeding for each student depends on past performance, teachers typically teach a single grade. Thus, since they get a new set of students each year, behaviour today does not impact payments in the future.

[32] The treatment effects on threshold specific pass rates are shown in Tables A.3- A.6 in Appendix A.

The estimated treatment effects (on the incentivised test) for Levels schools are comparable with those found in previous experiments in India and Mexico (Muralidharan and Sundararaman, 2011; Behrman *et al.*, 2015). The estimated effects for the Pay for Percentile design are lower than those found in Loyalka *et al.* (2019) but larger than those in Gilligan *et al.* (2019).[33] The results suggest that the Levels design performs at least as well as the Pay for Percentile design. Despite the theoretical predictions, we find no evidence that the Pay for Percentile system leads to larger increases in learning relative to Levels. 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 mostly outperforms Pay for Percentile) and statistically significant in three out of four cases.

The larger treatment effects found in the incentivised 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 described in Section 3.2 and Appendix E, our implementation team tightly controlled the administration of the incentivised test, mitigating concerns about cheating. Assuming that test-taking effort drives all the differences between our incentivised and non-incentivised results, student effort can increase test scores between 0.0051σ and 0.11σ (see Panel C). This is generally in line with the findings of Levitt *et al.* (2016) and Gneezy *et al.* (2019).[34]

Given the reward structure, teachers in both treatment arms were motivated to ensure that their students took the incentivised 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 increased student participation in the incentivised test by 5 percentage points. Their counterparts in Pay for Percentile schools increased participation by 3 percentage points (see Table A.1 in Appendix A). Following Lee (2009), we compute bounds on the treatment effects by trimming the excess test-takers from the left and right tails of the incentivised test distribution. Focusing on the year-two results for

[33] For the full sample Gilligan *et al.* (2019) find that Pay for Percentile incentives have a small (0.01σ) and statistically insignificant effect on learning. However, there is substantial heterogeneity in treatment effects. Pay for Percentile incentives improve learning outcomes in schools with books by 0.11σ on the grade-relevant sub-test. In schools without books, there is no significant treatment effect on learning.

[34] The difference between the treatment effects in the incentivised and non-incentivised test across treatments arms is not statistically significant (i.e., testing $\gamma_1 = \gamma_2$ and $\gamma_1 = \gamma_2 = \gamma_3$). We can match some students across both tests. Although matching is not random (see Table A.8), the results are qualitatively similar if we focus on this sample (see Table A.9).

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 A.7 in Appendix A).

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 can appropriately group most students by ability. Our results are generally robust to this sample restriction (see Table A.10 in Appendix A).

Finally, the results are robust to different controls. Our results are qualitatively similar whether we use a parsimonious specification that only includes randomisation strata (see Table A.11) or specifications that use controls selected by a post-double lasso procedure (see Tables A.12-A.13).

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-incentivised test						
Levels (α_1)	.039	.045	.057	.068*	.096*	.095**
	(.047)	(.048)	(.047)	(.04)	(.052)	(.045)
P4Pctile (α_2)	-.015	-.033	-.027	.072**	.0018	.041
	(.04)	(.039)	(.039)	(.037)	(.05)	(.043)
N. of obs.	4,781	4,781	4,781	4,869	4,869	4,869
$\alpha_3 = \alpha_2 - \alpha_1$	-.053	-.078*	-.084*	.0047	-.094*	-.053
p-value ($H_0 : \alpha_3 = 0$)	.23	.084	.056	.92	.074	.27
Panel B: Incentivised test						
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: Incentivised – Non-incentivised						
$\gamma_1 = \beta_1 - \alpha_1$	.065	.075	.1	.063	.072	.11
p-value($\gamma_1 = 0$)	.13	.1	.046	.12	.12	.025
$\gamma_2 = \beta_2 - \alpha_2$	.075	.044	.078	.019	.077	.077
p-value($\gamma_2 = 0$)	.083	.32	.14	.64	.073	.11
$\gamma_3 = \beta_3 - \alpha_3$	.01	-.031	-.024	-.044	.0058	-.037
p-value($\gamma_3 = 0$)	.81	.51	.65	.28	.91	.49
$\gamma_1 - \gamma_2$	-.01	.031	.024	.044	-.0058	.037
p-value($\gamma_1 - \gamma_2 = 0$)	.81	.51	.65	.28	.91	.49
$\gamma_1 - \gamma_3$	.075	.044	.078	.019	.077	.077
p-value($\gamma_1 - \gamma_3 = 0$)	.083	.32	.14	.64	.073	.11
$\gamma_2 - \gamma_3$	.085	.013	.055	-.026	.083	.041
p-value($\gamma_2 - \gamma_3 = 0$)	.26	.86	.56	.71	.31	.65
p-value($\gamma_1 = \gamma_2 = \gamma_3$)	.17	.25	.12	.28	.13	.062

Results from estimating Equation 2 for different subjects at both follow-ups. Panel A uses data from the non-incentivised test taken by a sample of students. Panel B uses data from the incentivised test taken by all students. Control variables in both panels include student characteristics (gender and grade) and school characteristics (PTR, Infrastructure Principal Component Analysis (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. Table A.11 provides a version without school controls. Tables A.12-A.13 provide results when post double lasso selection is used to select the control variables. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.2 Grade repetition

Cross-country comparisons reveal 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). Thus, lowering repetition rates can improve the fiscal efficiency of schools.

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. As a result, promotion is no longer automatic, and pupils can be forced to repeat based on a decision by the school committee (automatic promotion remains in place after grade 3). School committees use internal data to make decisions (not data collected from the study), but these data and decision-making processes are not standardised.

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 0.048), a 24% reduction. There was a small positive and statistically insignificant effect on grade repetition among students in Pay for Percentile schools. Formal hypothesis tests show that the estimated reduction in Levels schools was significantly lower (p-value 0.034) compared to the estimated change in Pay for Percentile schools.

Table 4: Effect on grade repetition

	(1) Year 1	(2) Year 2
Levels (α_1)	-.0095	-.033**
	(.02)	(.017)
P4Pctile (α_2)	.025	.0025
	(.017)	(.014)
N. of obs.	4,781	4,869
Mean control	.13	.14
$\alpha_3 = \alpha_2 - \alpha_1$	.035*	.035**
p-value ($H_0 : \alpha_3 = 0$)	.062	.034

Results from estimating Equation 2 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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-incentivised 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. 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 and the persistence of any learning gains induced by the program (in the second year of the evaluation).

Overall, we do not see decreases in fourth graders' test scores, 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 0.1σ (p-value 0.045) (Column 3). In Pay for Percentile schools, we find relatively small ($-.024\sigma$) and statistically insignificant

(p-value 0.63) effects on composite test scores. Since we tested fourth-grade students and collected information on them 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 *et al.* (2009), where a student incentive program for girls improved the performance of non-eligible boys who believed they would also benefit from the program.[35]

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 incentivised (see Table 5, Panel B).

[35] We also test for any spillover effects on the seventh-grade primary school national exit exam (PSLE). We do not find evidence that our incentives affected students' performance on those tests. See Table A.14 in Appendix A.

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)	.13**	.044	.1**	.059	.042	.049
	(.062)	(.051)	(.05)	(.062)	(.057)	(.053)
P4Pctile (α_2)	-.03	-.033	-.024	-.0038	.00015	-.0028
	(.054)	(.054)	(.05)	(.06)	(.051)	(.05)
N. of obs.	1,513	1,513	1,513	1,482	1,482	1,482
$\alpha_3 = \alpha_2 - \alpha_1$	-.16**	-.077	-.13**	-.063	-.041	-.052
p-value ($H_0 : \alpha_3 = 0$)	.011	.13	.017	.27	.44	.32
Panel B: Science (Grades 1-3)						
	Year 1	Year 2				
Levels (α_1)	.069	.083				
	(.063)	(.06)				
P4Pctile (α_2)	-.002	.078				
	(.05)	(.057)				
N. of obs.	4,781	4,869				
$\alpha_3 = \alpha_2 - \alpha_1$	-.071	-.0055				
p-value ($H_0 : \alpha_3 = 0$)	.26	.92				

Results from estimating Equation 2 for grade 4 students (Panel A) and for grade 1-3 students in science (Panel B). Control variables include student characteristics (age, gender, grade, and lag test scores) and school characteristics (PTR, Infrastructure PCA index, a PCA index of how close 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.4 Equity concerns

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-incentivised tests) and Figure 3 (for the incentivised tests).[36] This allows us to examine where teachers differentially focus their efforts in the distribution of student baseline test scores. The analysis also allows us to gauge whether the Levels system exacerbates inequality relative to the Pay for Percentile System, as suggested by Macartney *et al.* (2021).

Despite equity concerns, we do not find evidence that the Levels system increases inequality relative to the Pay for Percentile System. In both years, we find similar patterns of heterogeneity in treatment effects by student baseline test scores using data from the incentivised and non-incentivised tests. In the first year

[36]We also explore heterogeneity by additional student characteristics such as gender, as well as school characteristics such as pupil-teacher ratio, and find limited evidence of heterogeneity in those characteristics (see Tables A.15 and A.16 in Appendix A for details).

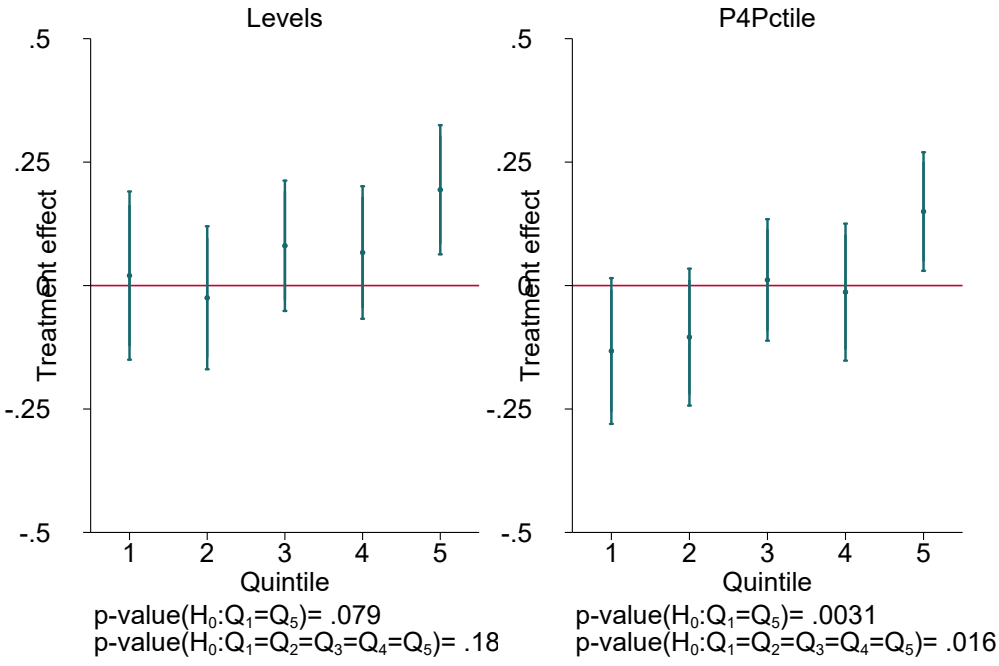
of the program, teachers in both systems focused their attention on the best students. This pattern is more pronounced in Pay for Percentile schools, where we can reject that the estimated learning gains are the same for all quintiles (p-value 0.016). In the second year of the program, the treatment effects were more balanced across the distribution of students, and we fail to reject the hypothesis that the treatment effects for each baseline score quintile are equal.[37] Our first-year results for the Pay for Percentile treatment are in line with Gilligan *et al.* (2019), 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 align with Loyalka *et al.* (2019) who find learning gains across the entire distribution of students.

We also examine if teachers in Levels schools targeted students near passing thresholds. If this were the case, we would expect bunching around the thresholds set by the Levels system. However, the general pattern we observe is that most students are either well above or well below the passing threshold, with no bunching at the threshold (see Figures A.6 and A.7). Combined with our heterogeneity analysis by students' baseline ability, these results allow us to rule out this possibility.

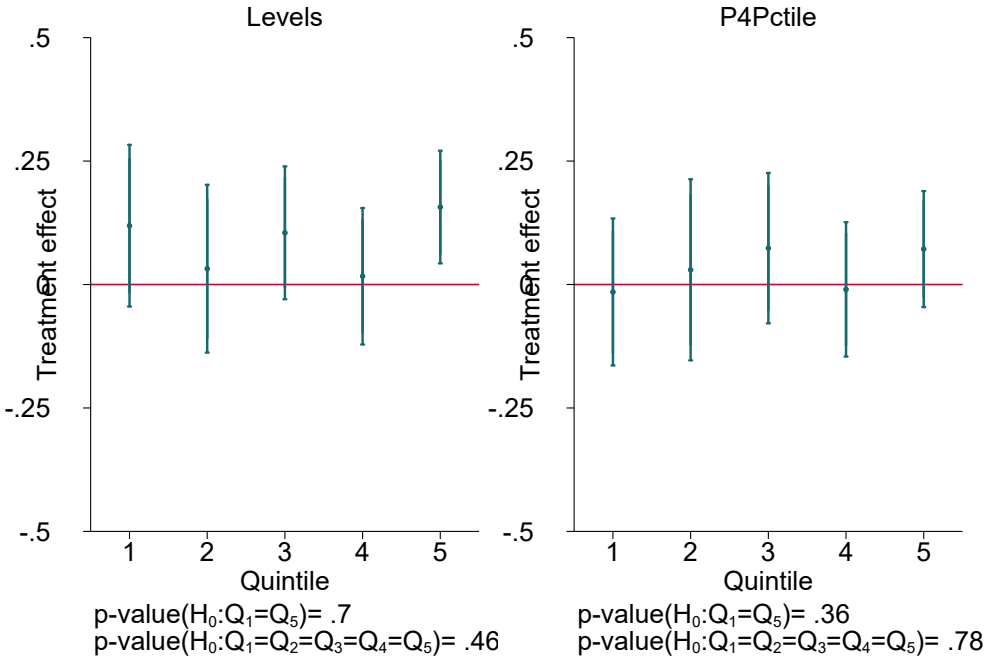
Finally, both treatments reduce the standard deviation and the Gini coefficient of test scores within a classroom (see Appendix A.13). In some cases, the reduction in inequality of test scores is greater for the Levels system, although the difference is not statistically significant nor consistent across years, subjects, or measures of inequality.

[37] Subject specific results are available in Figures A.2 - A.5 in Appendix A.

Figure 2: Heterogeneity in treatment effects by baseline score — non-incentivised test



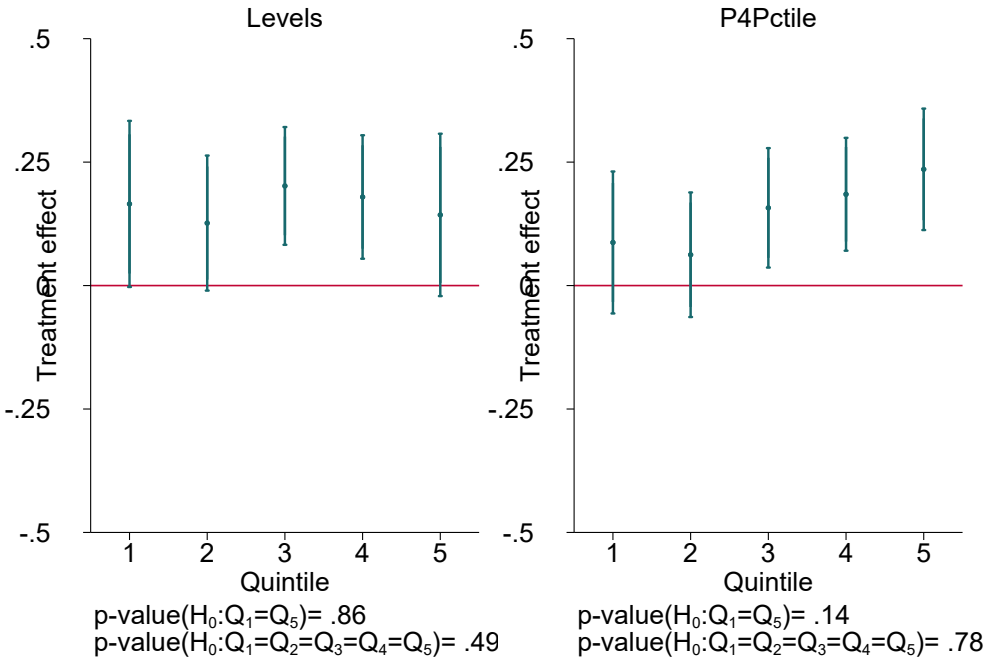
(a) Year 1



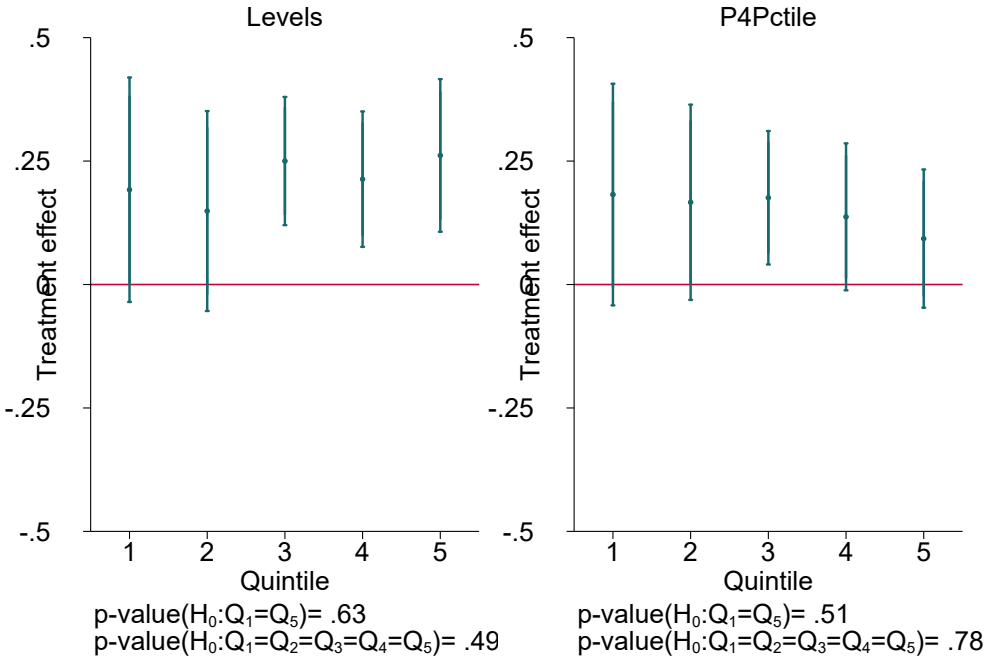
(b) Year 2

Note: These figures show the treatment effects (and their 95% confidence interval) in the composite score of the non-incentivised test (y-axis) at the end of the first year (2a) and the second year (2b) by student's quintile in the baseline score (x-axis). We also show p-values for testing whether the treatment effects for students in the first quintile are the same as the effects for those in the last quintile, and for testing whether the treatment effect is the same across all five quintiles.

Figure 3: Heterogeneity in treatment effects by baseline score — incentivised test



(a) Year 1



(b) Year 2

Note: These figures show the treatment effects (and their 95% confidence interval) in the composite score of the incentivised test (y-axis) at the end of the first year (2a) and the second year (2b) by student's quintile in the baseline score (x-axis). We also show p-values for testing whether the treatment effects for students in the first quintile are the same as the effects for those in the last quintile, and for testing whether the treatment effect is the same across all five quintiles.

4.5 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 and Vesterlund, 2007, 2011). However, we do not find any significant heterogeneous treatment effects by teacher gender (Table 6, Column 1). We also do not find any heterogeneous effects by teacher's age, which proxies for experience.

Although previous studies (e.g., Metzler and Woessmann (2012) and Bietenbeck *et al.* (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, which our survey team measured through math and word association tests (Column 3). More effective teachers, as measured by the headteacher's performance rating, as well as teachers who were more confident in their teaching abilities, responded more to both incentives (Columns 4 and 5).

Table 6: Heterogeneity by teacher characteristics

	(1)	(2)	(3)	(4)	(5)
	Male	Age	IRT	HT Rating	Self Rating
Levels	0.056	0.070	0.066*	0.091**	0.072**
	(0.039)	(0.052)	(0.037)	(0.038)	(0.036)
Gains	-0.0066	-0.000058	0.010	0.035	0.014
	(0.033)	(0.047)	(0.033)	(0.039)	(0.030)
Levels*Covariate	0.0031	-0.00025	0.016	0.099***	0.062**
	(0.056)	(0.0012)	(0.033)	(0.022)	(0.030)
P4Pctile*Covariate	0.035	0.00025	0.0081	0.048*	0.091***
	(0.052)	(0.0012)	(0.036)	(0.027)	(0.029)
Covariate	-0.066	-0.00022	-0.017	-0.037**	-0.077***
	(0.041)	(0.0013)	(0.025)	(0.018)	(0.021)
Covariate mean	.37	39	-.15	-.012	-.074
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), headteacher 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. In addition to the variables shown, all our specifications include both treatment indicators and the covariate. We use three measures of teacher ability to explore the heterogeneity in treatment effects. First, teachers were tested on all three subjects, and we created an index of content knowledge using an IRT model. Second, headteachers were asked to rate teacher performance in seven dimensions, including the ability to ensure that students learn and classroom management skills. Third, to create the self-perception metric, we create a Principal Component Analysis (PCA) 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 for example when students are confused." Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.6 Teacher Effort and Behaviour

Since the treatments were designed to elicit teacher effort and modify their behaviour, we examine teacher responsiveness to the incentives in this section. We measure teacher effort using four different approaches and report the second-year estimates for brevity (Appendix A reports the first-year estimates). First, our survey team measured and collected teacher presence data 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 7, Panel A).[38]

[38]Our results suggest a (small) reduction in school absenteeism, and an increase in classroom presence, but these coefficients are imprecisely estimated. Getting precision on this particular measure would require coordinating multiple surprise visits, which is often impractical due to budget constraints and other logistical challenges (Muralidharan, 2017). Our budget did not allow us to

Second, we use data from student reports to examine additional dimensions of teacher effort (see Panel B in Table 7). We find suggestive evidence that teachers in treatment schools are less likely to hit students and more likely to call them by their names during class time and provide additional help outside the classroom. However, teachers did not increase the amount of homework assigned to students.

Table 7: Treatment effects on teacher behaviour - Year 2

Panel A: Spot checks

	(1) In school	(2) In classroom
Levels (α_1)	-0.025	0.025
	(0.050)	(0.053)
P4Pctile (α_2)	-0.0050	0.023
	(0.044)	(0.044)
N. of obs.	180	180
Mean control	.7	.36
$\alpha_3 = \alpha_2 - \alpha_1$	.02	-.0021
p-value ($H_0 : \alpha_3 = 0$)	.71	.97

Panel B: Student reports

	(1) Extra help	(2) Homework	(3) Call by name	(4) Hit
Levels (α_1)	0.0052	0.0029	0.080**	-0.030
	(0.0096)	(0.018)	(0.037)	(0.035)
P4Pctile (α_2)	0.016*	-0.023	0.047	-0.061*
	(0.0097)	(0.019)	(0.032)	(0.032)
N. of obs.	9,557	9,557	9,557	9,557
Mean control	.062	.12	.5	.37
$\alpha_3 = \alpha_2 - \alpha_1$	.011	-.026	-.032	-.031
p-value ($H_0 : \alpha_3 = 0$)	.29	.24	.34	.35

Panel A presents school-level data on teacher absenteeism (Column 1) and time-on-task (Column 2). Panel B presents student-level data on teacher behaviour (as reported by students), extra help (Column 1), homework assignment (Column 2), calling by name (Column 3), and hitting/pinching/slapping students (Column 4). This table uses data collected at the end of the second school year of the experiment. Table A.19 presents results using data collected at the end of the first school year of the experiment. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As our third measure of teacher effort, we use two sets of classroom observations. We conducted "external classroom observations," where our survey teams observed teacher behaviour by standing outside the classroom for several minutes to prevent disruptions.[39] We also conducted within-classroom observations

collect data in this manner. Instead, these data were collected during announced visits because our main goal was to test students and survey teachers with minimal attrition.

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

following the World Bank Service Delivery Indicator protocols. However, in-class observations are often affected by Hawthorne/John Henry effects, reducing how useful these protocols are (Muralidharan and Sundararaman, 2010). Even though the external observations are less detailed, they are arguably better able to capture broad measures of teacher behaviour because they are not affected by Hawthorne/John Henry effects.

Our findings using the external observations are shown in Table 8. Students in both Levels and Pay for Percentile schools were less likely to be off-task during the classroom observations. However, the reduction is only statistically significant in Pay for Percentile schools. While there is an increase in the likelihood that teachers are teaching (instead of off-task or managing the classroom), this is imprecisely estimated and statistically insignificant for both treatments. The results from the in-class observations (see Appendix Table A.21) also suggest teachers are less likely to be off-task and more likely to be teaching. However, these results are also imprecisely estimated and statistically insignificant for both treatments, perhaps a reflection of Hawthorne/John Henry effects.

Table 8: External classroom observation - Year 2

	(1) Teaching	(2) Classroom management	(3) Teacher off task	(4) Student off task
Levels (α_1)	0.058	-0.028	-0.032	-0.036
	(0.054)	(0.022)	(0.056)	(0.042)
P4Pctile (α_2)	0.024	-0.063***	0.035	-0.073**
	(0.050)	(0.023)	(0.052)	(0.033)
N. of obs.	772	772	772	772
Control mean	.69	.041	.27	.048
$\alpha_3 = \alpha_2 - \alpha_1$	-.033	-.035*	.067	-.038
p-value ($H_0 : \alpha_3 = 0$)	.55	.095	.22	.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 as 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). This table uses data collected at the end of the second school year of the experiment. Table A.20 presents results using data collected at the end of the first school year of the experiment. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As our final measure of teacher effort, we use data from inspections of student notebooks.[40] During these inspections, our enumerators checked whether students had recent assignments and whether they had received feedback from the teacher. These data corroborate the student reports that teachers did not assign more homework in response to the incentives. However, there is some suggestive evidence from student notebooks that homework assignments were longer in Levels schools (see Table A.22 Column 3), but this might reflect multiple testing within this family of outcomes. These data also highlight ample room for improvement among teachers in their grading and feedback practices. Only two-thirds of assignments were graded, and one-fifth of notebooks had written comments or feedback.

In short, despite comprehensive data on teacher absenteeism, time-on-task (measured through classroom observations), student reports on teacher behaviour, and data from student notebooks, we are only able to find suggestive evidence of changes in teacher effort in a narrow set of outcomes. This could be explained by the fact that teachers can adjust effort on many margins (e.g., pedagogy, time-on-task, homework, socio-emotional support) in response to incentives linked to student learning, and these margins are either difficult to measure or unobservable. However, overall, our results suggest minor improvements in teacher effort that are generally similar across both treatment groups.

4.7 Earnings Expectations and Turnover

Before the payout of the bonuses, we collected data on teachers' earnings expectations from the incentives and 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.

Both Pay for Percentile and Levels teachers overestimated their expected earnings. For example, pay for Percentile teachers expected to earn about 430,000 TZS in bonus payments, while Levels teachers expected about 525,000 TZS. However, the average bonus payment was 226,337 TZS in 2016. Thus, on average, teachers in Levels and Pay for Percentile schools expected to earn 2.3 and 1.9 times the actual average

[40] We are not aware of other studies exploiting this potential data source. We believe there is great potential in using student notebooks as a data source to measure student and teacher effort.

payment, respectively. If teacher effort mirrors their expectations, this could provide insights into the factors that drove their behavioural responses.

Teachers in Pay for Percentile schools had lower bonus earnings expectations than their peers in the Levels system. They expected almost 95,000 TZS (US$42) less in bonus payments than teachers in the Levels system, an 18% reduction in bonus expectations relative to the mean expectations of teachers in the Levels system (Column 1).[41] 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.[42] While competitive pressure can be motivating, it can also be demotivating if an individual teacher has low subjective beliefs about their probability of winning 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 in teachers' beliefs about their rankings across the treatments. Teachers were optimistic about their projected earnings: Less than 1% of teachers expected to be among the bottom earners (Column 2) and 7% were worried about earning a low bonus (Column 5). On the other hand, 80% expected to be among the district's top earners (Column 4).[43]

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**	.0058	-.012	.035	-.02
	(37,169)	(.008)	(.04)	(.045)	(.026)
N. of obs.	653	676	676	676	676
Mean Levels	525,641	.0057	.11	.8	.074

This table shows 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[41] As in Brown and Andrabi (2020), expectations do not track closely to actual performance.
[42] The shape of the distribution of expected earnings is similar across treatment arms, with a mean-shift but similar variance. See Figure A.8.
[43] We also explore the extent to which the incentives affected teachers' goal-setting behaviour in Appendix A.18.

One of the potential effects of larger incentives is increasing teacher assignment retention (whether a teacher is still teaching the assigned grade and subject at the end of the year). Teachers can rotate out of their assignments during the year and move to a different grade/subject within the same school or transfer to another school. A teacher who rotates into an incentivised grade/subject combination in a treatment school during the year is eligible for the bonus and is paid a pro-rated bonus. This aspect of the bonus was explained to teachers at baseline.

Teacher assignment retention was slightly higher (around 6 percentage points, from a base of 73%) in the first year in both incentive schools compared to control schools, but the difference is statistically insignificant. In the second year, teacher assignment retention remained 6 percentage points higher in Levels schools (statistically insignificant) compared with control schools. However, assignment retention was 8 percentage points higher (p-value 0.01) in Pay for Percentile schools than control schools by the end of the second year (see Appendix Table A.24). Higher teacher retention, and thus more experience with the incentive scheme, could explain the larger second-year treatment effect on learning for Pay for Percentile schools (relative to the first year).

4.8 Experience with incentive designs

As discussed in Section 3.2, the schools in this study were part of a previous experiment studied by Mbiti *et al.* (2019). Although the randomisation of the treatments we study was stratified by the treatment/control assignment in the previous experiment, some of our current results may be driven by exposure to the previous treatment. For instance, teachers who had previously participated in the treatment studied by Mbiti *et al.* (2019) might be better able to respond to the current treatments due to their experience with a (single) threshold teacher incentives scheme. We explore this possibility in Tables A.25-A.26 in Appendix A. The effects for both treatments are higher for schools that had prior exposure to teacher incentives through the previous experiment. This could reflect learning by doing or greater trust in the system and the implementing team (i.e., teachers are more likely to believe they will get paid for exerting higher effort over time). We cannot reject the null hypothesis that teachers in both systems benefit equally from prior exposure to the incentives.

Due to the churn in teacher grades and school allocations in Tanzania, there is significant variation across schools in the fraction of lower grade teachers who were previously exposed to incentives. We explore this dimension of heterogeneity by interacting the current treatment assignment indicators with a binary variable equal to one if at least 50% of (current) teachers in focal grades and subjects previously participated in an incentive scheme (see Table 10).[44]

This analysis yields five insights — we focus on the results from the high-stakes test as we have the universe of students and thus more power, but the results are qualitatively similar if we focus on the low-stakes test. First, teachers who were exposed to incentives in the past, but not in the current experiment (i.e., are now in Control schools) do not have a positive treatment effect on test scores, suggesting long-term effects of the previous experiment do not drive our results.[45]

Second, with the exception of Pay for Percentile teachers who were not previously exposed to incentives in the past, the treatment effects for all other groups are positive, but not all are statistically significant. In the second year, the treatment effects are positive for all groups and all but one are statistically significant on the incentivised test. The pattern for the previously unexposed Pay for Percentile teachers suggests teachers need additional time to react to a scheme without "bright lines", such as Pay for Percentile.

Third, for teachers who had been exposed to incentives in the past, the treatment effects of the Levels and Pay for Percentile treatments are constant in both years of our study (especially in the high-stakes test), suggesting that after two years of exposure to incentives, there is a "steady-state" in the treatment effects.

Fourth, for teachers who had not been exposed to incentives before, treatment effects increase over time and tend to converge to those of previously exposed teachers. This could either reflect learning by doing

[44] Across schools, the median proportion of teachers who were previously exposed to incentives is 50% in both years of our study. The results are qualitatively similar if we interact the current treatment assignment indicators with a binary variable equal to one if at least 25% of (current) teachers in focal grades and subjects previously participated in an incentive scheme (see Table A.27).

[45] Further, test scores in Control schools who were treated in the previous experiment are not different from other Control schools (see Table A.26). Moreover, if we restrict our analysis to the set of students that were not exposed to the previous treatment (those in Grade 1 in year 1, and those in Grade 1 and 2 in year 2), the results are qualitatively similar (see Table A.28), suggesting that long-term effects from the treatment in the previous experiment are not driving the results in the experiment we analyse in this article.

(with respect to incentives) or the development of trust between individual teachers and the implementing agency.[46]

Fifth, the Levels treatment effect is (at least) as effective as the Pay for Percentile effect regardless of whether teachers were exposed to incentives in the past or not.[47]

[46]Note the treatment effect grows over time in this experiment, as well as in the previous experiment (Mbiti *et al.*, 2019), further suggesting it takes some time to build trust and to get familiarised with performance pay schemes.
[47]Excluding schools that did not receive incentives in the previous experiment from the estimation also yields treatment effects for Levels that are above those of Pay for Percentile schools (see Table A.29). Since these results focus on schools that were previously exposed to incentives, they are more likely to reflect "steady-state" effects.

Table 10: Heterogeneity by whether at least half of the current teachers were previously exposed to incentives

	(1) Year 1	(2) Year 2
Panel A: Non-incentivised test		
Levels × Teachers not incentivized in previous RCT (α_1)	.049	.11**
	(.061)	(.056)
Levels × Teachers incentivized in previous RCT (α_2)	.074	.088
	(.06)	(.063)
P4Pctile × Teachers not incentivized in previous RCT (β_1)	-.063	.0089
	(.056)	(.05)
P4Pctile × Teachers incentivized in previous RCT (β_2)	.015	.062
	(.049)	(.063)
Teachers incentivized in previous RCT	-.054	.024
	(.063)	(.056)
N. of obs.	4,781	4,869
p-value($H_0 : \alpha_1 = \alpha_2$)	.75	.76
p-value($H_0 : \beta_1 = \beta_2$)	.29	.49
p-value($H_0 : \alpha_1 = \beta_1$)	.059	.08
p-value($H_0 : \alpha_2 = \beta_2$)	.3	.68
p-value($H_0 : (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)$)	.48	.35
Panel B: Incentivised test		
Levels × Teachers not incentivized in previous RCT (α_1)	.12	.2***
	(.072)	(.066)
Levels × Teachers incentivized in previous RCT (α_2)	.28***	.29***
	(.077)	(.087)
P4Pctile × Teachers not incentivized in previous RCT (β_1)	-.031	.08
	(.065)	(.068)
P4Pctile × Teachers incentivized in previous RCT (β_2)	.17**	.17**
	(.073)	(.079)
Teachers incentivized in previous RCT	-.087	.088
	(.066)	(.085)
N. of obs.	48,077	59,680
p-value($H_0 : \alpha_1 = \alpha_2$)	.048	.36
p-value($H_0 : \beta_1 = \beta_2$)	.02	.34
p-value($H_0 : \alpha_1 = \beta_1$)	.054	.088
p-value($H_0 : \alpha_2 = \beta_2$)	.15	.11
p-value($H_0 : (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)$)	.59	1

The outcome is the composite test scores across math and Kiswahili. "Teachers incentivised in previous RCT" is equal to one if at least half of the teachers currently teaching Kiswahili or Math were eligible for incentives in the previous experiment (analysed by Mbiti *et al.* (2019)) in a treatment school. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5 Cost-effectiveness

We use accounting records to examine the cost-effectiveness of our interventions, following the framework outlined in Dhaliwal *et al.* (2013). The total annual cost of the teacher incentive programs was US$7.23 per student.[48] This cost estimate includes both the direct costs (value of incentive payments) and the implementation costs (test design and implementation, communications, audit, transfer costs, etc.) of the program. However, in the long run, the cost of the Pay for Percentile scheme is US$1.50 higher (US$8.73 total) due to pre-testing costs to determine ability groups.[49]

We use the treatment effect on the composite index in the incentivised test in the second year to compute cost-effectiveness. We focus on the incentivised test to facilitate comparability with other teacher incentive studies. Since the Pay for Percentile treatment effect is 0.13σ in the second year, the cost-effectiveness of the intervention is 0.0148σ per dollar spent (assuming a linear dose-response relationship, for comparability with other studies, yields a cost-effectiveness of 1.48σ per US$100 spent). On the other hand, the Levels treatment effect is 0.22σ, implying a cost-effectiveness of 0.0304σ per dollar spent (again, assuming a linear dose-response relationship for comparability with other studies yields a cost-effectiveness of 3.04σ per US$100 spent). These estimates suggest that the Levels treatment is as effective as several other interventions in developing countries analysed in the overview by Kremer *et al.* (2013). For instance, the Levels treatment is more cost-effective than a computer-assisted learning program evaluated in India (1.54σ per US$100) and the incentive program on attendance in India (2.28σ per US$100).

6 Conclusion

We use a randomised controlled trial to compare the effectiveness of two different teacher incentive programs to improve early-grade learning in Tanzanian public schools. Specifically, we compare the effectiveness of an innovative multiple-threshold proficiency incentive design relative to a more sophisticated

[48] For reference, average per-student expenditure, including teacher salaries, was ∼98 USD per student in 2014.

[49] 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 primarily fixed costs and relatively small compared to the variable costs, especially at scale.

rank-order tournament-style Pay for Percentile system in terms of their impact on student test scores two years after the start of the program.

We report two sets of findings. First, both programs increase test scores compared to students in the control group. Moreover, the programs did not lead to negative learning spillovers in non-incentivised grades or subjects. Since neither program had any pedagogy or teacher training element, the learning effects are solely attributable to introducing teacher incentives. Both programs were equally effective for teachers with high and low content knowledge, and both programs benefited students across the skill distribution.

Second, despite the theoretical advantage of the Pay for Percentile system, our multiple-threshold proficiency system was at least as effective at increasing test scores and reducing grade repetition as the Pay for Percentile system. Combining these findings on program effectiveness with the lower cost of the Levels program, we conclude that Levels is the more cost-effective incentive program (although both programs are relatively cost-effective).

Our results demonstrate some theoretical and practical considerations education authorities interested in adopting teacher incentive programs must face. 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. In contrast, the multiple-threshold proficiency system used in this study communicates clear student-level targets. These salient targets provide teachers with clear signals about how to allocate their effort in the class. Since developing countries often face 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 or cheating) will increase. Since rank-order tournaments (such as Pay for Percentile) allow education systems to use different tests and test formats, they can minimise these concerns if administrators are willing and able to implement such testing changes. Longer studies conducted at scale will be needed to understand better the long-run advantages and disadvantages of different teacher incentive systems.[50]

Acknowledgments

Twaweza funded this evaluation 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 Asociación 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).

Affiliations

[1]Frank Batten School of Leadership and Public Policy, University of Virginia, 235 McCormick Road, Charlottesville, VA 22904

[2]J-PAL

[3]Centro de Investigación Económica, ITAM, Av. Camino a Santa Teresa 930, CDMX, Mexico

[4]Twaweza, 15 Uganda Avenue, Dar Es Salaam, Tanzania

[5]NBER

[50]While some studies (e.g., Lavy (2020)) study long-term outcomes of being exposed to a teacher performance program, we refer to studies that focus on the steady-state outcomes after the program has been in place for several years (e.g., Neal and Schanzenbach (2010)).

[6]IZA

References

Ahrens, A., Hansen, C.B. and Schaffer, M.E. (2018). 'PDSLASSO: Stata module for post-selection and post-regularization OLS or IV estimation and inference', Statistical Software Components, Boston College Department of Economics.

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

Barlevy, G. and Neal, D. (2012). 'Pay for percentile', *American Economic Review*, vol. 102(5), pp. 1805–31, doi:10.1257/aer.102.5.1805.

Barrera-Osorio, F., Cilliers, J., Cloutier, M.H. and Filmer, D. (2022). 'Heterogenous teacher effects of two incentive schemes: Evidence from a low-income country', *Journal of Development Economics*, vol. 156, p. 102820, ISSN 0304-3878, doi:https://doi.org/10.1016/j.jdeveco.2022.102820.

Barrera-Osorio, F. and Raju, D. (2017). 'Teacher performance pay: Experimental evidence from Pakistan', *Journal of Public Economics*, vol. 148, pp. 75 – 91.

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

Bettinger, E.P. and Long, B.T. (2010). 'Does cheaper mean better? the impact of using adjunct instructors on student outcomes', *The Review of Economics and Statistics*, vol. 92(3), pp. 598–613.

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

Bold, T., Filmer, D., Martin, G., Molina, E., Stacy, B., Rockmore, C., Svensson, J. and Wane, W. (2017). 'Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa', *Journal of Economic Perspectives*, vol. 31(4), pp. 185–204.

Breeding, M.E., Beteille, T. and Evans, D.K. (2021). 'Teacher pay-for-performance : What works? where? and how?', .

Brehm, M., Imberman, S.A. and Lovenheim, M.F. (2017). 'Achievement effects of individual performance incentives in a teacher merit pay tournament', *Labour Economics*, vol. 44, pp. 133–150.

Brown, C. and Andrabi, T. (2020). 'Inducing positive sorting through performance pay: Experimental evidence from pakistani schools', *University of California at Berkeley Working Paper*.

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

Charness, G. and Kuhn, P. (2011). 'Chapter 3 - lab labor: What can labor economists learn from the lab?', in (O. Ashenfelter and D. Card, eds.), *Handbook of Labor Economics*, pp. 229–330, vol. 4, Elsevier, doi:https://doi.org/10.1016/S0169-7218(11)00409-6.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K. and Rogers, F.H. (2006). 'Missing in action: Teacher and health worker absence in developing countries', *Journal of Economic Perspectives*, vol. 20(1), pp. 91–116.

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

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

Contreras, D. and Rau, T. (2012). 'Tournament incentives for teachers: evidence from a scaled-up intervention in Chile', *Economic development and cultural change*, vol. 61(1), pp. 219–246.

Cunningham, R. (2018). 'Unicef think piece series: Curriculum reform', UNICEF Eastern and Southern Africa Regional Office, Nairobi.

Dalton, P.S., Gonzalez, V. and 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. and Rogers, H. (2018). 'Double for nothing? experimental evidence on an unconditional teacher salary increase in Indonesia', *The Quarterly Journal of Economics*, vol. 133(2), pp. 993–1039.

Dhaliwal, I., Duflo, E., Glennerster, R. and 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*, pp. 285–338.

Fehr, E., Klein, A. and Schmidt, K.M. (2007). 'Fairness and contract design', *Econometrica*, vol. 75(1), pp. 121–154, doi:https://doi.org/10.1111/j.1468-0262.2007.00734.x.

Ferraz, C. and Bruns, B. (2012). 'Paying teachers to perform: The impact of bonus pay in Pernambuco, Brazil.', *Society for Research on Educational Effectiveness*.

Figlio, D. and Loeb, S. (2011). 'Chapter 8 - school accountability', in (E. A. Hanushek, S. Machin and L. Woessmann, eds.), *Handbook of the Economics of Education*, pp. 383–421, vol. 3, Elsevier, doi:https://doi.org/10.1016/B978-0-444-53429-3.00008-9.

Fryer, R.G. (2013). 'Teacher incentives and student achievement: Evidence from New York City public schools', *Journal of Labor Economics*, vol. 31(2), pp. 373–407.

Fryer, R.G., Levitt, S.D., List, J. and Sadoff, S. (Forthcoming). 'Enhancing the efficacy of teacher incentives through framing: A field experiment', *American Economic Journal: Economic Policy*.

Gilligan, D.O., Karachiwalla, N., Kasirye, I., Lucas, A.M. and Neal, D. (2019). 'Educator incentives and educational triage in rural primary schools', *Journal of Human Resources*.

Glewwe, P., Ilias, N. and Kremer, M. (2010). 'Teacher incentives', *American Economic Journal: Applied Economics*, vol. 2(3), pp. 205–27, doi:10.1257/app.2.3.205.

Gneezy, U., List, J.A., Livingston, J.A., Qin, X., Sadoff, S. and Xu, Y. (2019). 'Measuring success in education: The role of effort on the test itself', *American Economic Review: Insights*, vol. 1(3), pp. 291–308, doi:10.1257/aeri.20180633.

Gómez-Minambres, J. (2012). 'Motivation through goal setting', *Journal of Economic Psychology*, vol. 33(6), pp. 1223 – 1239, ISSN 0167-4870, doi:https://doi.org/10.1016/j.joep.2012.08.010.

Goodman, S.F. and 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*, vol. 31(2), pp. 409 – 420.

Hanushek, E.A. and Rivkin, S.G. (2012). 'The distribution of teacher quality and implications for policy', *Annual Review of Economics*, vol. 4(1), pp. 131–157.

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

Kane, T.J., Rockoff, J.E. and Staiger, D.O. (2008). 'What does certification tell us about teacher effectiveness? evidence from New York City', *Economics of Education Review*, vol. 27(6), pp. 615–631.

Koch, A. and Nafziger, J. (2011). 'Self-regulation through goal setting', *The Scandinavian Journal of Economics*, vol. 113(1), pp. 212–227, doi:10.1111/j.1467-9442.2010.01641.x.

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

Kremer, M., Miguel, E. and Thornton, R. (2009). 'Incentives to learn', *The Review of Economics and statistics*, vol. 3(91), pp. 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*, vol. 18(1), pp. 1–16.

Lavy, V. (2002). 'Evaluating the effect of teachers' group performance incentives on pupil achievement', *Journal of Political Economy*, vol. 110(6), pp. pp. 1286–1317, ISSN 00223808.

Lavy, V. (2009). 'Performance pay and teachers' effort, productivity, and grading ethics', *American Economic Review*, vol. 99(5), pp. 1979–2011, doi:10.1257/aer.99.5.1979.

Lavy, V. (2020). 'Teachers' pay for performance in the long-run: The dynamic pattern of treatment effects on students' educational and labour market outcomes in adulthood', *The Review of Economic Studies*, ISSN 0034-6527, doi:10.1093/restud/rdaa002, rdaa002.

Leaver, C., Ozier, O., Serneels, P. and Zeitlin, A. (2021). 'Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from rwandan primary schools', *American Economic Review*, vol. 111(7), pp. 2213–46, doi:10.1257/aer.20191972.

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

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

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

Loyalka, P., Sylvia, S., Liu, C., Chu, J. and Shi, Y. (2019). 'Pay by design: Teacher performance pay design and the distribution of student achievement', *Journal of Labor Economics*, vol. 37(3), pp. 621–662, doi:10.1086/702625.

Macartney, H. (2016). 'The dynamic effects of educational accountability', *Journal of Labor Economics*, vol. 34(1), pp. 1–28.

Macartney, H., McMillan, R. and Petronijevic, U. (2021). 'A quantitative framework for analyzing the distributional effects of incentive schemes', National Bureau of Economic Research, doi:10.3386/w28816.

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

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

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C. and Rajani, R. (2019). 'Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania', *The Quarterly Journal of Economics*, vol. 134(3), pp. 1627–1673, ISSN 0033-5533, doi:10.1093/qje/qjz010.

Mbiti, I., Romero, M. and Schipper, Y. (2022). 'Replication data for "Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania"', doi:10.5281/zenodo.7411312.

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

Miller, G. and Babiarz, K. (2014). 'Pay-for-performance incentives in low- and middle-income country health programs', in (A. J. Culyer, ed.), *Encyclopedia of Health Economics*, pp. 457 – 466, San Diego: Elsevier.

Mohanan, M., Donato, K., Miller, G., Truskinovsky, Y. and Vera-Hernández, M. (Forthcoming). 'Different strokes for different folks: Experimental evidence on the effectiveness of input and output incentive contracts for health care providers with varying skills', *American Economic Journal: Applied Economics*.

Muralidharan, K. (2017). 'Chapter 3 - field experiments in education in developing countries', in (A. V. Banerjee and E. Duflo, eds.), *Handbook of Economic Field Experiments*, pp. 323–385, vol. 2 of *Handbook of Economic Field Experiments*, North-Holland, doi:https://doi.org/10.1016/bs.hefe.2016.09.004.

Muralidharan, K. and Niehaus, P. (2017). 'Experimentation at scale', *Journal of Economic Perspectives*, vol. 31(4), pp. 103–24.

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

Muralidharan, K. and Sundararaman, V. (2011). 'Teacher performance pay: Experimental evidence from India', *Journal of Political Economy*, vol. 119(1), pp. pp. 39–77.

Neal, D. (2011). 'Chapter 6 - the design of performance pay in education', in (E. A. Hanushek, S. Machin and L. Woessmann, eds.), *Handbook of The Economics of Education*, pp. 495–550, vol. 4 of *Handbook of the Economics of Education*, Elsevier, doi:https://doi.org/10.1016/B978-0-444-53444-6.00006-7.

Neal, D. and Schanzenbach, D.W. (2010). 'Left behind by design: Proficiency counts and test-based accountability', *Review of Economics and Statistics*, vol. 92(2), pp. 263–283, ISSN 0034-6535.

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

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

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

Pham, L.D., Nguyen, T.D. and Springer, M.G. (2021). 'Teacher merit pay: A meta-analysis', *American Educational Research Journal*, vol. 58(3), pp. 527–566.

PRI (2013). 'Tanzanian teachers learning education doesn't pay', .

Renmans, D., Holvoet, N., Orach, C.G. and Criel, B. (2016). 'Opening the 'black box' of performance-based financing in low- and lower middle-income countries: a review of the literature', *Health Policy and Planning*, vol. 31(9), pp. 1297–1309, ISSN 0268-1080, doi:10.1093/heapol/czw045.

Reuters (2012). 'Tanzanian teachers in strike over pay', .

Singh, P. and 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*, ISSN 0304-3878, doi:https://doi.org/10.1016/j.jdeveco.2018.10.003.

Uwezo (2012). 'Are our children learning? annual learning assessment report 2011', Uwezo, accessed on 05-12-2014.

Uwezo (2013). 'Are our children learning? numeracy and literacy across East Africa', Uwezo, Nairobi, accessed on 05-12-2014.

van der Linden, W.J. and Hambleton, R.K. (2013). *Handbook of modern item response theory*, Springer Science & Business Media.

Vigdor, J. (2008). 'Teacher salary bonuses in North Carolina', .

Woessmann, L. (2011). 'Cross-country evidence on teacher performance pay', *Economics of Education Review*, vol. 30(3), pp. 404–418.

World Bank (2015). 'Service delivery indicators: Tanzania', The World Bank, Washington D.C.

World Bank (2017a). 'Data for development: An evaluation of world bank support for data and statistical capacity', .

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

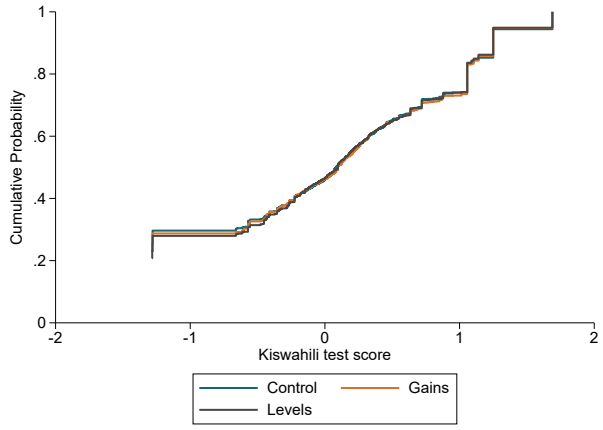
World Bank (2018). 'World development report 2018: Learning to realize education's promise', doi:10.1596/978-1-4648-1096-1.

A Additional Figures and Tables

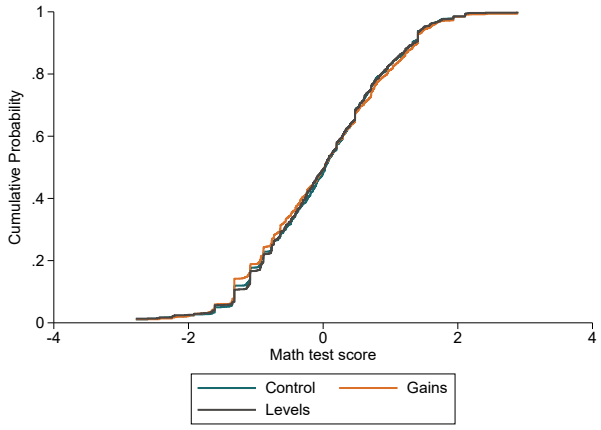
A.1 Test score distribution at baseline

Figure A.1: CDF of test scores at baseline across experimental groups

(a) Kiswahili



(b) Math



A.2 Effects on test takers (in the incentivised test)

Table A.1: Number of test takers, incentivised 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 enrolment in each grade) of the incentivised exam. The unit of observation is the school-grade level. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.3 Test score results for English

Table A.2: Effect on English test scores for grade 3

	(1) Year 1	(2) Year 2
Panel A: Non-incentivised test		
	English	English
Levels (α_1)	.012	.11
	(.086)	(.085)
P4Pctile (α_2)	-.049	.19**
	(.076)	(.081)
N. of obs.	1,532	1,533
$\alpha_3 = \alpha_2 - \alpha_1$	-.061	.08
p-value ($H_0 : \alpha_3 = 0$)	.43	.29
Panel B: Incentivised test		
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: Incentivised test – Non-incentivised test		
$\gamma_1 = \beta_1 - \alpha_1$	.14	.15
p-value($\gamma_1 = 0$)	.12	.14
$\gamma_2 = \beta_2 - \alpha_2$	.2	.04
p-value($\gamma_2 = 0$)	.017	.66
$\gamma_3 = \beta_3 - \alpha_3$	.053	-.11
p-value($\gamma_3 = 0$)	.54	.28
$\gamma_1 - \gamma_2$	-.053	.11
p-value($\gamma_1 - \gamma_2 = 0$)	.54	.28
$\gamma_1 - \gamma_3$	.2	.04
p-value($\gamma_1 - \gamma_3 = 0$)	.017	.66
$\gamma_2 - \gamma_3$	.25	-.074
p-value($\gamma_2 - \gamma_3 = 0$)	.076	.65
p-value($\gamma_1 = \gamma_2 = \gamma_3$)	.054	.33

Results from estimating Equation 2 for different subjects at both follow-ups. Panel A uses data from the non-incentivised 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 incentivised 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.4 Pass rates

Table A.3: 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 skill as evidenced by performance on the incentivised test. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: 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 skill as evidenced by performance on the incentivised test. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: 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 skill as evidenced by performance on the incentivised test. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: 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 skill as evidenced by performance on the incentivised test. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.5 Lee Bounds for the incentivised test

Table A.7: Lee bounds for the incentivised 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 standardised 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). Specifically, after we trim the data, we present the minimum of the lowest values within the confidence intervals (from trimming the left/right tail) and the maximum of the highest values within the confidence intervals. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.6 Incentivised and non-incentivised tests matched sample

Table A.8: Difference between students matched across the incentivised and the non-incentivised test

	(1) Non-matched	(2) Matched	(3) Difference
Panel A: Year 1			
Age	9.29	8.32	0.98***
	(1.68)	(1.33)	(0.05)
Gender	0.50	0.48	0.02*
	(0.50)	(0.50)	(0.01)
Swahili	0.10	-0.07	0.17***
	(1.02)	(0.96)	(0.04)
English	0.09	-0.08	0.17***
	(1.08)	(0.94)	(0.04)
Math	0.13	-0.15	0.28***
	(1.05)	(0.95)	(0.04)
Levels	0.31	0.37	-0.07***
	(0.46)	(0.48)	(0.02)
P4Pctile	0.31	0.36	-0.05***
	(0.46)	(0.48)	(0.02)
Panel B: Year 2			
Age	8.60	8.01	0.60***
	(1.37)	(1.23)	(0.05)
Gender	0.51	0.49	0.02
	(0.50)	(0.50)	(0.01)
Swahili	-0.05	-0.32	0.26***
	(0.99)	(0.90)	(0.04)
English	-0.04	-0.31	0.27***
	(1.01)	(0.79)	(0.04)
Math	-0.07	-0.42	0.34***
	(0.99)	(0.86)	(0.04)
Levels	0.31	0.36	-0.05**
	(0.46)	(0.48)	(0.02)
P4Pctile	0.32	0.35	-0.03
	(0.47)	(0.48)	(0.02)

This table presents the mean and standard error of the mean (in parentheses) for several characteristics of students that we were able to match across the incentivised and the non-incentivised test. The match was manually via a fuzzy merge based on student's names. Column 3 shows the difference between the two groups (and the standard error of the difference). Standard errors are clustered at the school level for the test of equality. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.9: Effect on test scores for the students we are able to match across the incentivised and the non-incentivised test

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: Non-incentivised test						
Levels (α_1)	.068	.05	.077	.068	.039	.06
	(.055)	(.057)	(.054)	(.049)	(.057)	(.051)
P4Pctile (α_2)	.034	.0079	.019	.078*	-.012	.033
	(.044)	(.046)	(.042)	(.046)	(.053)	(.048)
N. of obs.	2,841	2,841	2,841	2,784	2,784	2,784
$\alpha_3 = \alpha_2 - \alpha_1$	-.034	-.042	-.058	.01	-.051	-.027
p-value ($H_0 : \alpha_3 = 0$)	.51	.43	.24	.85	.31	.57
Panel B: Incentivised test						
Levels (β_1)	.25***	.21***	.33***	.15**	.14**	.2**
	(.069)	(.07)	(.091)	(.065)	(.068)	(.087)
P4Pctile (β_2)	.18***	.11*	.21***	.079	.097	.12
	(.06)	(.059)	(.079)	(.061)	(.063)	(.081)
N. of obs.	2,841	2,841	2,841	2,951	2,951	2,951
$\beta_3 = \beta_2 - \beta_1$	-.074	-.092	-.12	-.073	-.041	-.08
p-value ($H_0 : \beta_3 = 0$)	0.26	0.16	0.17	0.29	0.52	0.36
Panel C: Incentivised test – Non-incentivised test						
$\gamma_1 = \beta_1 - \alpha_1$	.17	.15	.23	.08	.091	.13
p-value($\gamma_1 = 0$)	.0028	.0047	.00069	.086	.055	.025
$\gamma_2 = \beta_2 - \alpha_2$	.13	.096	.17	-.0011	.1	.083
p-value($\gamma_2 = 0$)	.0074	.036	.0047	.98	.04	.16
$\gamma_3 = \beta_3 - \alpha_3$	-.04	-.05	-.06	-.082	.011	-.051
p-value($\gamma_3 = 0$)	.45	.28	.35	.083	.84	.43
$\gamma_1 - \gamma_2$	.04	.05	.06	.082	-.011	.051
p-value($\gamma_1 - \gamma_2 = 0$)	.45	.28	.35	.083	.84	.43
$\gamma_1 - \gamma_3$	.13	.096	.17	-.0011	.1	.083
p-value($\gamma_1 - \gamma_3 = 0$)	.0074	.036	.0047	.98	.04	.16
$\gamma_2 - \gamma_3$	.093	.046	.11	-.083	.11	.032
p-value($\gamma_2 - \gamma_3 = 0$)	.28	.55	.29	.32	.22	.77
p-value($\gamma_1 = \gamma_2 = \gamma_3$)	.0047	.015	.0014	.13	.058	.07

Results from estimating Equation 2 for different subjects at both follow-ups. Panel A uses data from the non-incentivised 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 incentivised 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.7 Test score results for properly seeded contests

Table A.10: 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-incentivised test						
Levels (α_1)	.052	.031	.057	.11**	.13**	.14***
	(.047)	(.053)	(.05)	(.05)	(.054)	(.05)
P4Pctile (α_2)	-.008	-.047	-.028	.11**	.089*	.11**
	(.044)	(.049)	(.047)	(.045)	(.051)	(.047)
N. of obs.	3,120	3,120	3,120	3,163	3,163	3,163
$\alpha_3 = \alpha_2 - \alpha_1$	-.06	-.078	-.085*	-.0026	-.039	-.032
p-value ($H_0 : \alpha_3 = 0$)	.18	.12	.07	.96	.46	.52
Panel B: Incentivised test						
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: Incentivised – Non-incentivised						
$\gamma_1 = \beta_1 - \alpha_1$	.07	.08	.11	.053	.0028	.065
p-value($\gamma_1 = 0$)	.16	.12	.061	.27	.96	.29
$\gamma_2 = \beta_2 - \alpha_2$	.081	.073	.099	-.02	-.027	-.0045
p-value($\gamma_2 = 0$)	.11	.14	.095	.66	.56	.93
$\gamma_3 = \beta_3 - \alpha_3$	.012	-.0066	-.0097	-.074	-.03	-.069
p-value($\gamma_3 = 0$)	.82	.9	.87	.13	.63	.29
$\gamma_1 - \gamma_2$	-.012	.0066	.0097	.074	.03	.069
p-value($\gamma_1 - \gamma_2 = 0$)	.82	.9	.87	.13	.63	.29
$\gamma_1 - \gamma_3$	.081	.073	.099	-.02	-.027	-.0045
p-value($\gamma_1 - \gamma_3 = 0$)	.11	.14	.095	.66	.56	.93
$\gamma_2 - \gamma_3$	.093	.067	.09	-.094	-.056	-.074
p-value($\gamma_2 - \gamma_3 = 0$)	.3	.45	.39	.26	.54	.47
p-value($\gamma_1 = \gamma_2 = \gamma_3$)	.21	.21	.12	.3	.82	.5

Results from estimating Equation 2 for different subjects at both follow-ups. Panel A uses data from the non-incentivised 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 incentivised 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.8 Test score results without any controls

Table A.11: Effect on test scores – no controls

	(1)	(2)	(3)	(4)	(5)	(6)
	Year 1			Year 2		
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: Non-incentivised test						
Levels (α_1)	.021	.027	.0086	.059	.094*	.088
	(.048)	(.048)	(.083)	(.04)	(.05)	(.083)
P4Pctile (α_2)	-.028	-.046	-.043	.065*	-.0053	.17**
	(.041)	(.04)	(.076)	(.037)	(.048)	(.079)
N. of obs.	4,781	4,781	1,532	4,869	4,869	1,533
$\alpha_3 = \alpha_2 - \alpha_1$	-.049	-.073	-.052	.0066	-.099*	.087
p-value ($H_0 : \alpha_3 = 0$)	.28	.12	.51	.89	.062	.27
Panel B: Incentivised test						
Levels (β_1)	.1**	.11**	.15**	.12***	.15***	.19***
	(.047)	(.051)	(.066)	(.046)	(.047)	(.06)
P4Pctile (β_2)	.059	.0081	.048	.085**	.071	.11*
	(.042)	(.044)	(.057)	(.042)	(.048)	(.06)
N. of obs.	48,077	48,077	48,077	59,680	59,680	59,680
$\beta_3 = \beta_2 - \beta_1$	-.043	-.099*	-.1	-.036	-.076	-.079
p-value ($H_0 : \beta_3 = 0$)	0.36	0.090	0.15	0.42	0.13	0.21
Panel C: Incentivised test – Non-incentivised test						
$\gamma_1 = \beta_1 - \alpha_1$	.082	.08	.11	.063	.054	.1
p-value($\gamma_1 = 0$)	.077	.11	.052	.13	.26	.058
$\gamma_2 = \beta_2 - \alpha_2$	.087	.053	.09	.021	.076	.077
p-value($\gamma_2 = 0$)	.047	.22	.093	.6	.085	.12
$\gamma_3 = \beta_3 - \alpha_3$	.0059	-.027	-.021	-.042	.023	-.023
p-value($\gamma_3 = 0$)	.9	.61	.73	.32	.67	.67
$\gamma_1 - \gamma_2$	-.0059	.027	.021	.042	-.023	.023
p-value($\gamma_1 - \gamma_2 = 0$)	.9	.61	.73	.32	.67	.67
$\gamma_1 - \gamma_3$	.087	.053	.09	.021	.076	.077
p-value($\gamma_1 - \gamma_3 = 0$)	.047	.22	.093	.6	.085	.12
$\gamma_2 - \gamma_3$	.093	.027	.068	-.022	.099	.053
p-value($\gamma_2 - \gamma_3 = 0$)	.23	.75	.49	.76	.24	.56
p-value($\gamma_1 = \gamma_2 = \gamma_3$)	.088	.22	.092	.32	.19	.12

Results from estimating Equation 2 for different subjects at both follow-ups. Panel A uses data from the non-incentivised test taken by a sample of students. Control variables include student characteristics (age, gender, grade, and lag test scores). Panel B uses data from the incentivised test taken by all students. Control variables include student characteristics (gender and grade). Panel C tests the difference between the treatment estimates in panels A and B. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.9 Test score results with controls selected via Lasso

Table A.12: Effect on non-incentivised test scores with controls selected via Lasso

	(1)	(2)	(3)	(4)	(5)	(6)
	Year 1			Year 2		
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: All controls						
Levels (α_1)	.039	.045	.057	.068*	.096*	.095**
	(.047)	(.048)	(.047)	(.04)	(.052)	(.045)
P4Pctile (α_2)	-.015	-.033	-.027	.072**	.0018	.041
	(.04)	(.039)	(.039)	(.037)	(.05)	(.043)
N. of obs.	4,781	4,781	4,781	4,869	4,869	4,869
$\alpha_3 = \alpha_2 - \alpha_1$	-.053	-.078*	-.084*	.0047	-.094*	-.053
p-value ($H_0 : \alpha_3 = 0$)	.23	.084	.056	.92	.074	.27
Panel B: Lasso controls						
Levels (α_1)	.021	.027	.037	.05	.095*	.085**
	(.047)	(.047)	(.046)	(.039)	(.049)	(.043)
P4Pctile (α_2)	-.028	-.042	-.04	.064*	-.0034	.033
	(.041)	(.039)	(.04)	(.036)	(.048)	(.042)
N. of obs.	4,781	4,781	4,781	4,869	4,869	4,869
$\alpha_3 = \alpha_2 - \alpha_1$	-.049	-.069	-.077*	.014	-.098*	-.051
p-value ($H_0 : \alpha_3 = 0$)	.27	.13	.074	.77	.058	.29

Results from estimating Equation 2 for different subjects at both follow-ups. Both panels use data from the non-incentivised test taken by a sample of students. Panel A uses all the school, student, and household controls. Panel B uses the controls selected by post double lasso selection, as implemented by (Ahrens *et al.*, 2018). Table A.11 provides a version without school controls. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.13: Effect on incentivised test scores with controls selected via Lasso

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: All controls						
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 B: Lasso controls						
Levels (β_1)	.11**	.13***	.17***	.12***	.14***	.19***
	(.046)	(.049)	(.064)	(.046)	(.044)	(.055)
P4Pctile (β_2)	.058	.014	.051	.085**	.067	.1*
	(.041)	(.044)	(.057)	(.042)	(.047)	(.058)
N. of obs.	48,077	48,077	48,077	59,680	59,680	59,680
$\beta_3 = \beta_2 - \beta_1$	-.048	-.11**	-.12*	-.036	-.078*	-.082
p-value ($H_0 : \beta_3 = 0$)	0.29	0.031	0.075	0.42	0.099	0.16

Results from estimating Equation 2 for different subjects at both follow-ups. Both panels use data from the incentivised test taken by a sample of students. Panel A uses all the school, student, and household controls. Panel B uses the controls selected by post double lasso selection, as implemented by (Ahrens *et al.*, 2018). Table A.11 provides a version without school controls. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.10 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). We then merged them without 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 publicised 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 A.14).

We do find that test scores decreased on the SNFA examination in 2015. However, this is not consistent 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 are available upon request.

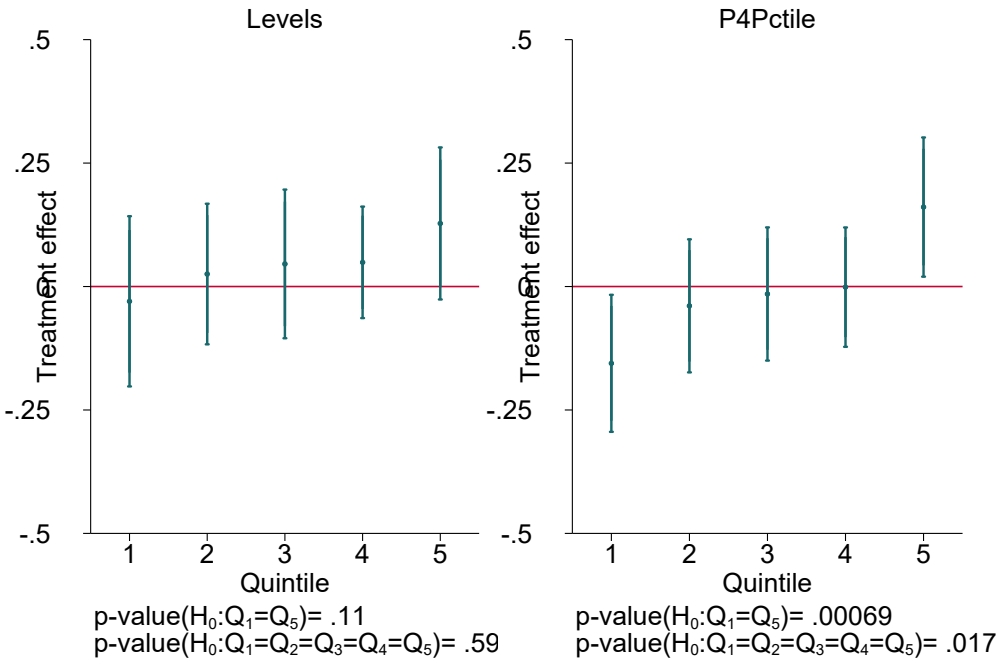
Table A.14: Effect on national assessments (Grade 7 - PSLE)

	(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

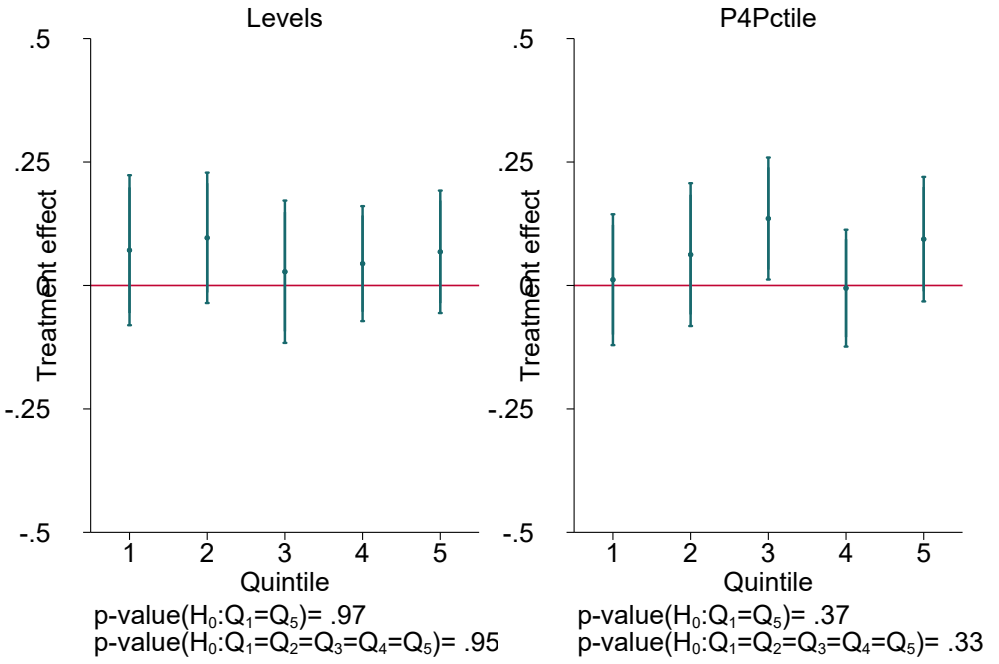
Data comes from Primary Schools Leaving Exam (PSLE) of The National Examination Council of Tanzania (NECTA). Standard errors, clustered at the school level, are in parentheses.

A.11 Additional heterogeneity in treatment effects

Figure A.2: Math — non-incentivised

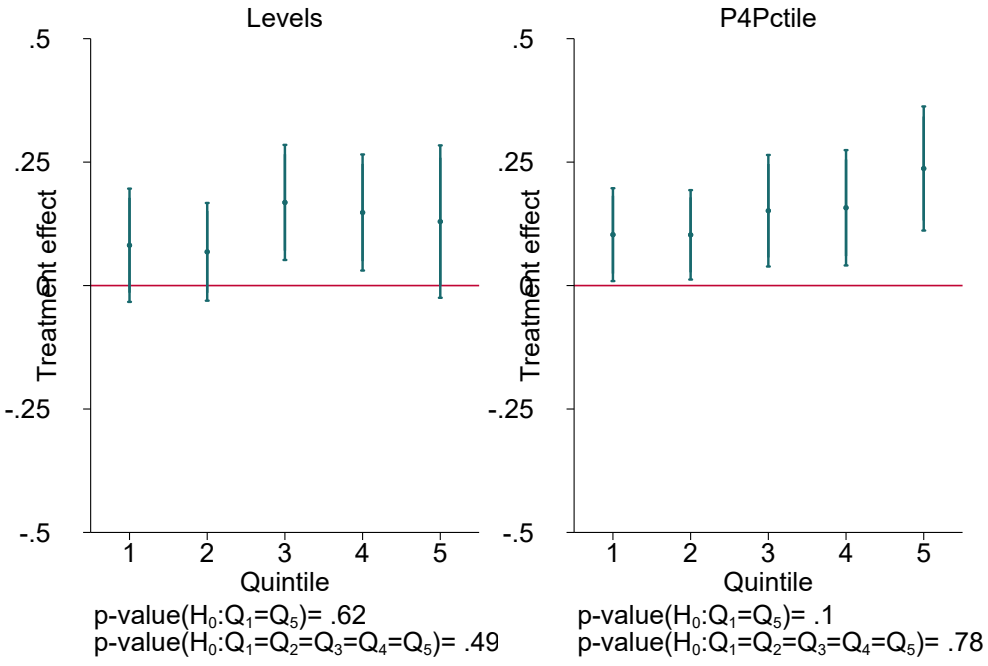


(a) Year 1

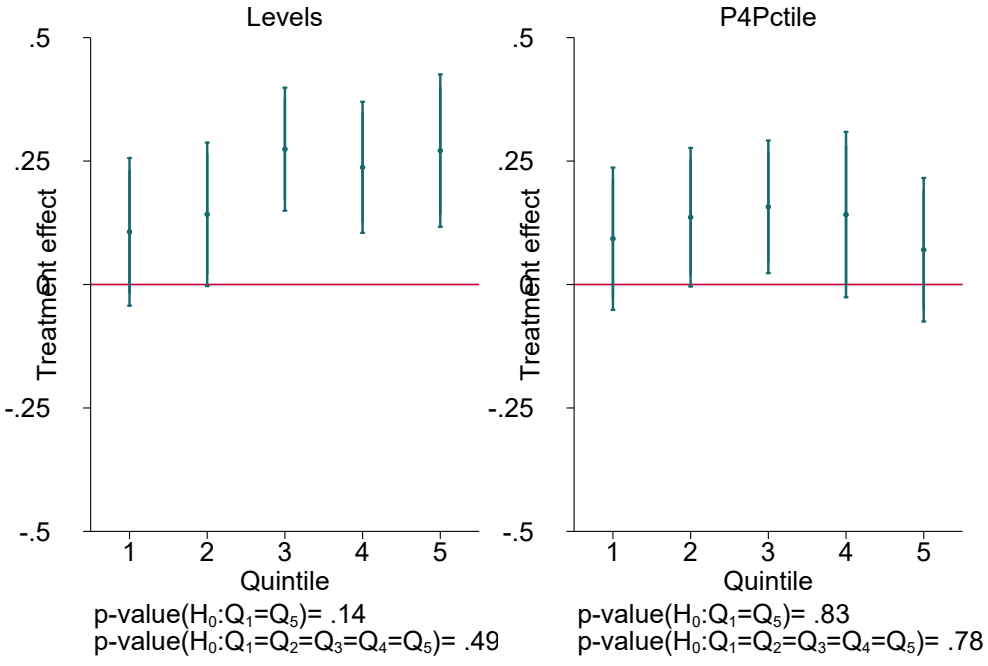


(b) Year 2

Figure A.3: Math — incentivised

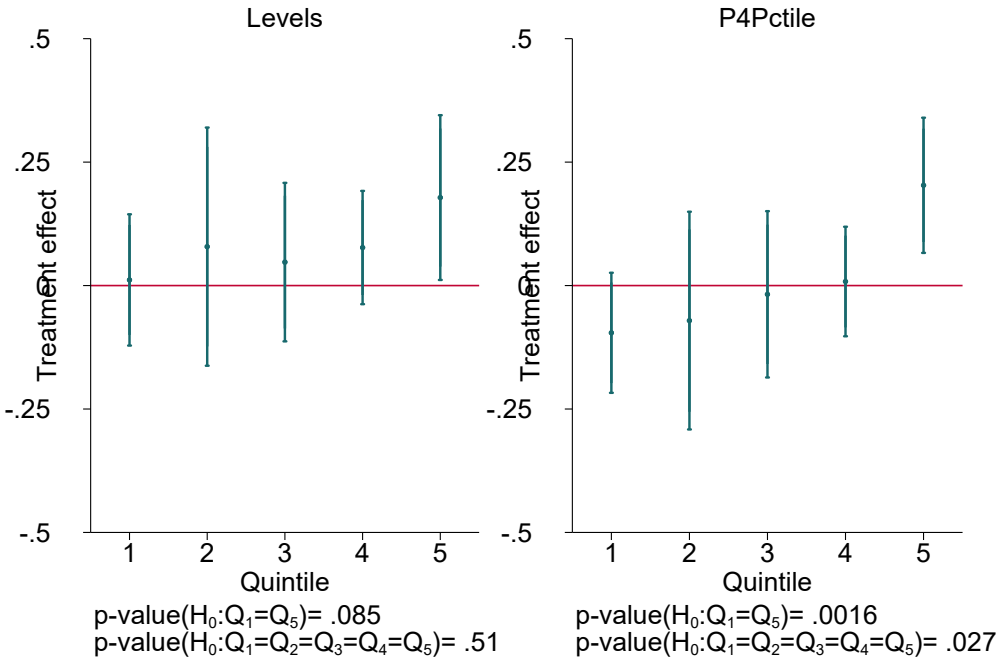


(a) Year 1

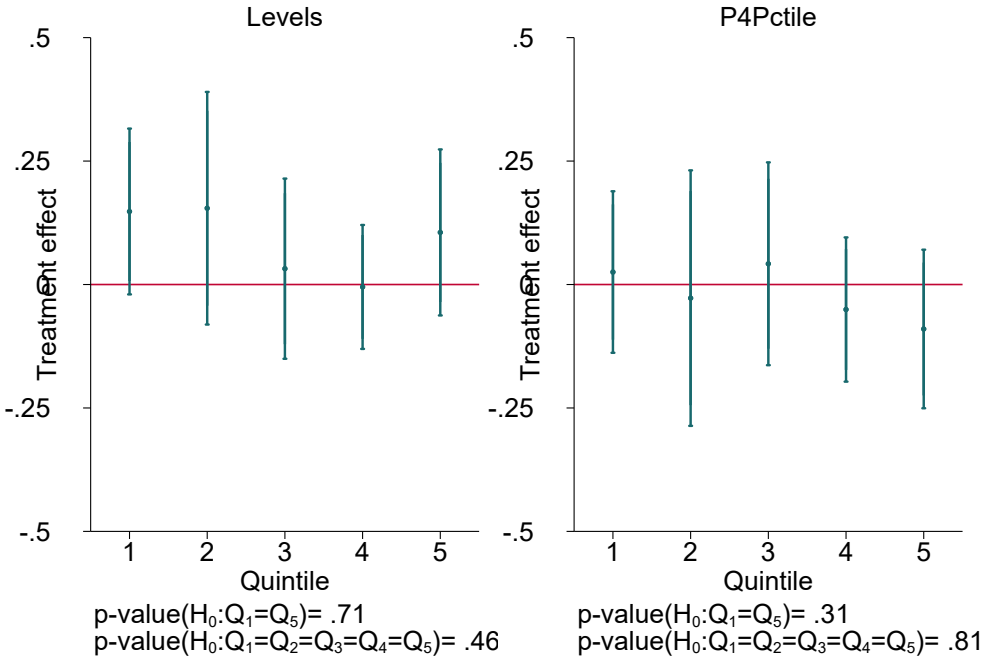


(b) Year 2

Figure A.4: Kiswahili — non-incentivised

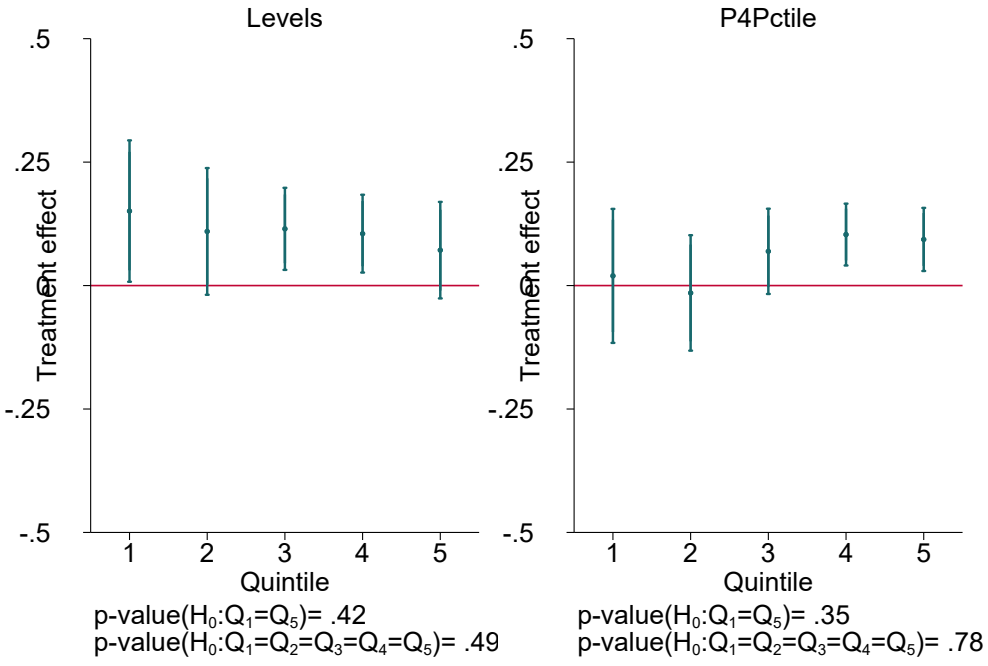


(a) Year 1

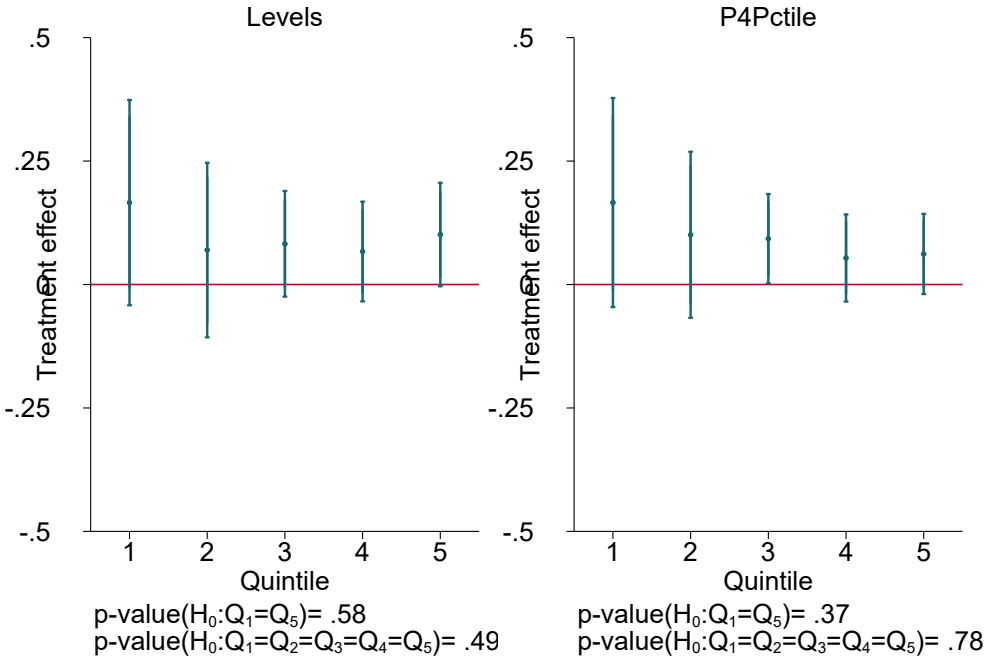


(b) Year 2

Figure A.5: Kiswahili — incentivised



(a) Year 1



(b) Year 2

Table A.15: Heterogeneity by student characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
	Math			Swahili		
	Male	Age	Test(Yr0)	Male	Age	Test(Yr0)
Levels	0.071	-0.047	0.059	0.044	0.35**	0.071*
	(0.050)	(0.14)	(0.036)	(0.052)	(0.16)	(0.039)
Gains	0.021	-0.052	0.043	0.0022	-0.057	-0.0080
	(0.045)	(0.14)	(0.032)	(0.045)	(0.17)	(0.036)
Levels*Covariate (α_2)	-0.025	0.011	0.026	0.036	-0.032*	0.011
	(0.039)	(0.015)	(0.033)	(0.048)	(0.018)	(0.029)
P4Pctile*Covariate (α_1)	0.0095	0.0089	0.063**	-0.024	0.0050	0.029
	(0.042)	(0.016)	(0.027)	(0.049)	(0.019)	(0.030)
Covariate	0.012	0.032**	0.30***	-0.045	0.039***	0.29***
	(0.029)	(0.013)	(0.024)	(0.034)	(0.014)	(0.024)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650
$\alpha_3 = \alpha_2 - \alpha_1$	.035	-.0024	.037	-.06	.037**	.018
p-value ($H_0 : \alpha_3 = 0$)	.4	.88	.23	.23	.048	.56

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

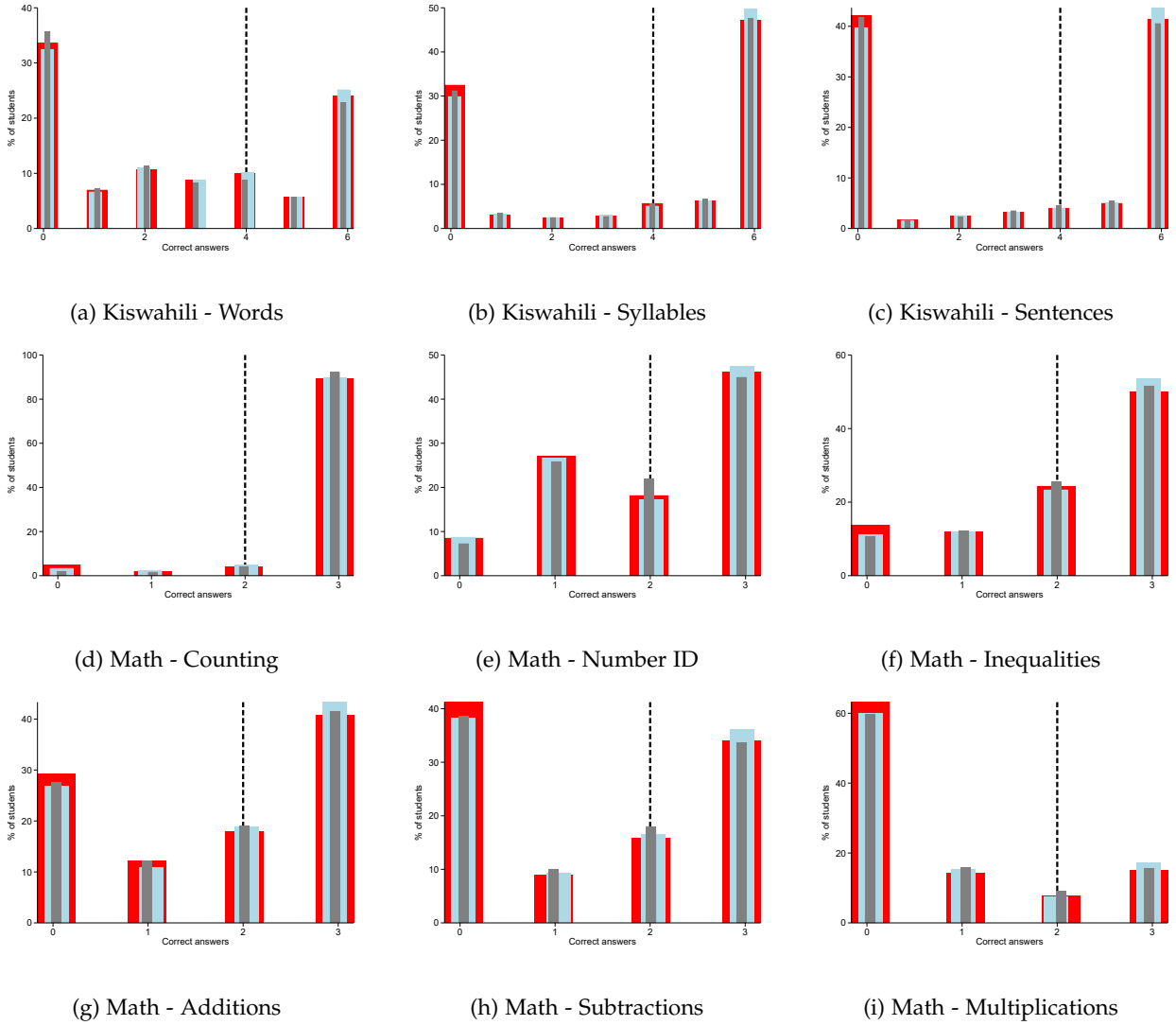
Table A.16: Heterogeneity by school characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
	Math			Swahili		
	Facilities	PTR	Fraction Weak	Facilities	PTR	Fraction Weak
Levels	0.062	0.060	0.14	0.073*	0.12	0.12
	(0.038)	(0.085)	(0.10)	(0.041)	(0.078)	(0.097)
Gains	0.027	0.17**	0.16*	-0.011	0.086	0.14
	(0.032)	(0.073)	(0.085)	(0.037)	(0.082)	(0.090)
Levels*Covariate (α_2)	0.031	-0.00015	-0.16	-0.018	-0.00096	-0.098
	(0.023)	(0.0015)	(0.18)	(0.027)	(0.0014)	(0.17)
P4Pctile*Covariate (α_1)	-0.027	-0.0025**	-0.24	-0.030	-0.0018	-0.29*
	(0.026)	(0.0012)	(0.15)	(0.030)	(0.0013)	(0.16)
Covariate	0.030*	0.00022	-0.14	0.032	0.00055**	-0.22
	(0.018)	(0.00029)	(0.18)	(0.020)	(0.00025)	(0.17)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650
$\alpha_3 = \alpha_2 - \alpha_1$	-.057**	-.0024	-.079	-.012	-.00082	-.19
p-value ($H_0 : \alpha_3 = 0$)	.023	.18	.62	.69	.63	.26

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

A.12 Raw data – high-stakes test

Figure A.6: Year 1 - Raw Data



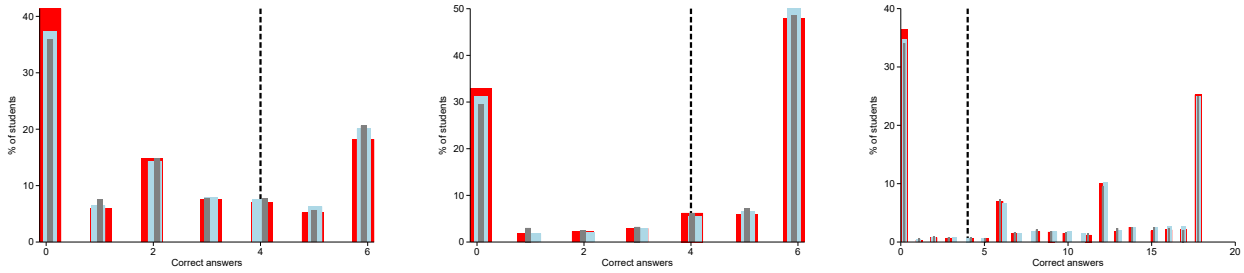
(a) Kiswahili - Words (b) Kiswahili - Syllables (c) Kiswahili - Sentences

(d) Math - Counting (e) Math - Number ID (f) Math - Inequalities

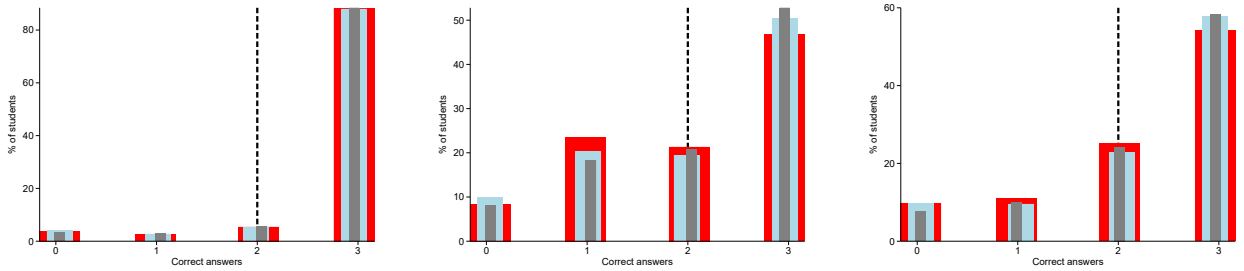
(g) Math - Additions (h) Math - Subtractions (i) Math - Multiplications

Note: This represents the proportion of correct responses across experimental groups for each skill tested in the high-stakes exam at the end of the first year. The red bars show the distribution of correct answers for the control group, the light blue bars the distribution for the Levels group, and the gray bars the distribution for the Pay for Percentile group. The vertical dotted line represents the passing thresholds.

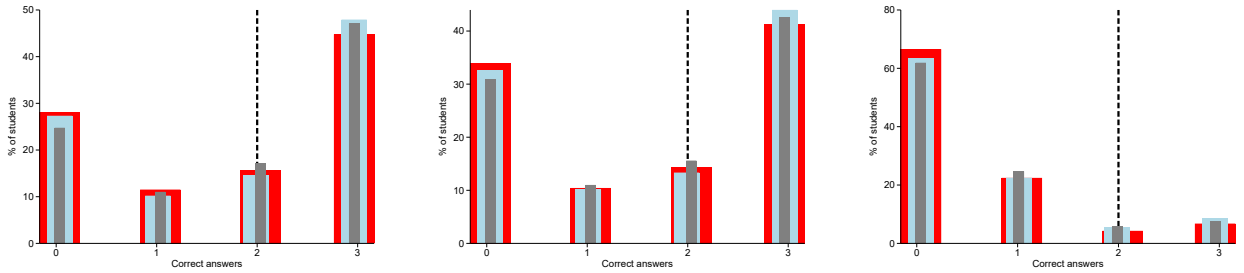
Figure A.7: Year 2 - Raw Data



(a) Kiswahili - Words (b) Kiswahili - Syllables (c) Kiswahili - Sentences



(d) Math - Counting (e) Math - Number ID (f) Math - Inequalities



(g) Math - Additions (h) Math - Subtractions (i) Math - Multiplications

Note: This represents the proportion of correct responses across experimental groups for each skill tested in the high-stakes exam at the end of the second year. The red bars show the distribution of correct answers for the control group, the light blue bars the distribution for the Levels group, and the gray bars the distribution for the Pay for Percentile group. The vertical dotted line represents the passing thresholds.

A.13 Effect on inequality

Table A.17: Effect on standard deviation of test scores within a school-grade

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: Non-incentivised test						
Levels (α_1)	.0032	-.06**	-.032	-.016	-.097**	-.073**
	(.023)	(.025)	(.026)	(.024)	(.038)	(.032)
P4Pctile (α_2)	.007	-.016	-.019	-.038*	-.073**	-.073**
	(.023)	(.027)	(.026)	(.022)	(.035)	(.03)
N. of obs.	5,399	5,399	5,399	5,400	5,400	5,400
Control mean	.74	.74	.73	.74	.75	.73
$\alpha_3 = \alpha_2 - \alpha_1$	.0038	.044*	.013	-.022	.024	-.0008
p-value ($H_0 : \alpha_3 = 0$)	.87	.079	.58	.36	.49	.98
Panel B: Incentivised test						
Levels (β_1)	-.026	-.071***	-.065***	-.022	-.044**	-.037
	(.016)	(.02)	(.023)	(.013)	(.022)	(.022)
P4Pctile (β_2)	-.074***	-.058***	-.096***	-.06***	-.051***	-.071***
	(.015)	(.018)	(.02)	(.011)	(.018)	(.018)
N. of obs.	48,118	48,118	48,118	59,755	59,755	59,755
Control mean	.92	.9	1.2	.92	.89	1.1
$\beta_3 = \beta_2 - \beta_1$	-.047***	.013	-.03	-.038***	-.0063	-.034
p-value ($H_0 : \beta_3 = 0$)	0.0027	0.46	0.13	0.0035	0.76	0.13

Results from estimating Equation 2 using as an outcome the standard deviation of test scores at the grade-school level. Panel A uses data from the non-incentivised test taken by a sample of students. Panel B uses data from the incentivised test taken by all students. Control variables in both panels 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). Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.18: Effect on Gini coefficient on test scores within a school-grade

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: Non-incentivised test						
Levels (α_1)	.002	-.02**	-.0045	-.003	-.023**	-.012*
	(.005)	(.01)	(.0056)	(.0053)	(.011)	(.0064)
P4Pctile (α_2)	.0064	.0015	.0021	-.0067	-.014	-.0088
	(.0046)	(.0091)	(.0055)	(.0051)	(.01)	(.0062)
N. of obs.	4,752	4,695	4,768	4,825	4,769	4,850
Control mean	.12	.19	.13	.14	.16	.12
$\alpha_3 = \alpha_2 - \alpha_1$	.0044	.022**	.0066	-.0037	.0091	.0031
p-value ($H_0 : \alpha_3 = 0$)	.4	.03	.23	.51	.37	.63
Panel B: Incentivised test						
Levels (β_1)	-.011***	-.033***	-.024***	-.0085**	-.02***	-.015**
	(.0036)	(.0088)	(.0065)	(.0037)	(.0077)	(.0061)
P4Pctile (β_2)	-.015***	-.013	-.019***	-.013***	-.014**	-.016***
	(.0032)	(.0083)	(.0058)	(.0031)	(.0066)	(.0051)
N. of obs.	48,042	47,777	48,043	59,662	59,504	59,672
Control mean	.16	.24	.22	.15	.19	.19
$\beta_3 = \beta_2 - \beta_1$	-.0041	.02**	.005	-.0044	.0062	-.0014
p-value ($H_0 : \beta_3 = 0$)	.2	.025	.42	.21	.41	.82

Results from estimating Equation 2 using as an outcome the Gini coefficient of test scores at grade-school level. Panel A uses data from the non-incentivised test taken by a sample of students. Panel B uses data from the incentivised test taken by all students. Control variables in both panels 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). Standard errors, clustered at the school level, are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.14 Teacher behaviour in year 1

Table A.19: Treatment effects on teacher behaviour - Year 1

Panel A: Spot checks		
	(1) In school	(2) In classroom
Levels (α_1)	0.012	0.0061
	(0.053)	(0.057)
P4Pctile (α_2)	-0.012	-0.023
	(0.044)	(0.050)
N. of obs.	180	180
Mean control	.7	.36
$\alpha_3 = \alpha_2 - \alpha_1$	-.024	-.029
p-value ($H_0 : \alpha_3 = 0$)	.65	.6
Panel B: Student reports		
	(1) Extra help	(2) Homework
Levels (α_1)	0.011	0.033
	(0.018)	(0.024)
P4Pctile (α_2)	-0.022	-0.0055
	(0.017)	(0.024)
N. of obs.	9,006	9,006
Mean control	.062	.12
$\alpha_3 = \alpha_2 - \alpha_1$	-.034*	-.038
p-value ($H_0 : \alpha_3 = 0$)	.073	.16

Panel A presents school-level data on teacher absenteeism (Column 1) and time-on-task (Column 2). Panel B presents student-level data on teacher behaviour (as reported by students), extra help (Column 1), homework assignment (Column 2), calling by name (Column 3), and hitting/pinching/slapping students (Column 4). This table uses data collected at the end of the first school year of the experiment. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.20: External classroom observation - Year 1

	(1) Teaching	(2) Classroom management	(3) Teacher off task	(4) Student off task
Levels (α_1)	-0.016	0.013	0.00045	0.0094
	(0.056)	(0.011)	(0.055)	(0.011)
P4Pctile (α_2)	-0.089*	-0.0016	0.083*	0.0050
	(0.050)	(0.012)	(0.050)	(0.011)
N. of obs.	1,308	1,308	1,308	1,308
Control mean	.69	.041	.27	.048
$\alpha_3 = \alpha_2 - \alpha_1$	-.074	-.015	.083	-.0044
p-value ($H_0 : \alpha_3 = 0$)	.23	.26	.16	.66

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). This table uses data collected at the end of the first school year of the experiment. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.15 Classroom observations

Table A.21: In-class classroom observations - Time usage

	(1) Teaching	(2) Grading	(3) Off task
Levels (α_1)	0.0055	0.0061	-0.012
	(0.053)	(0.027)	(0.054)
P4Pctile (α_2)	0.0041	-0.0019	-0.0022
	(0.058)	(0.025)	(0.055)
N. of obs.	772	772	772
Control mean	.54	.1	.13
$\alpha_3 = \alpha_2 - \alpha_1$	-.0014	-.008	.0094
p-value ($H_0 : \alpha_3 = 0$)	.98	.73	.86

The outcomes are the proportion of time during a classroom observation that the teacher spent teaching (Column 1), grading (Column 2), or off-task (Column 3). This table uses data collected at the end of the second school year of the experiment. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.16 Student notebook inspection

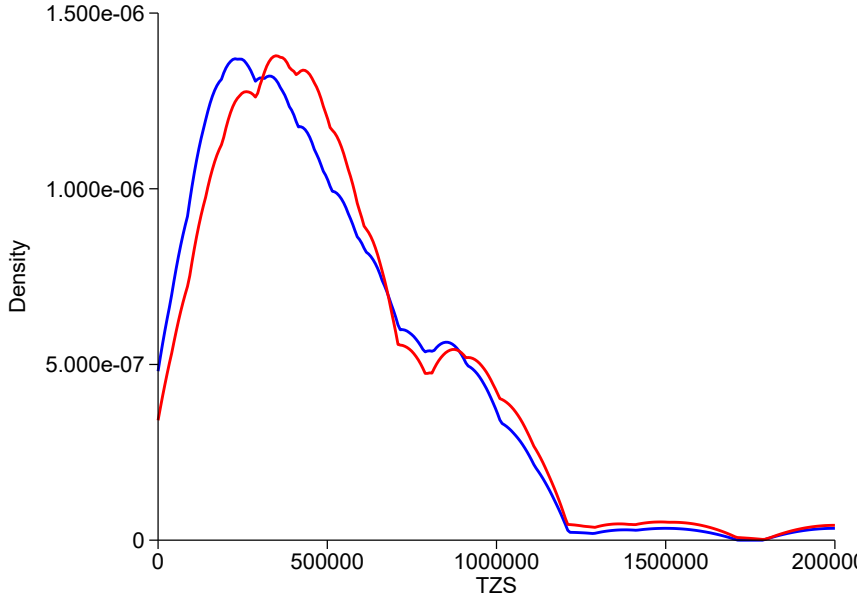
Table A.22: Student notebook inspection

	(1) Notebook	(2) Assignment	(3) Length	(4) Graded	(5) Marks	(6) Correct answer	(7) Pictures	(8) Feedback
Levels (α_1)	-0.022	0.015	0.083***	-0.0057	-0.0074	0.032	-0.0048	0.000013
	(0.025)	(0.033)	(0.025)	(0.037)	(0.037)	(0.028)	(0.0039)	(0.054)
P4Pctile (α_2)	-0.031	-0.0092	-0.0050	0.027	0.0069	0.051*	0.00086	0.053
	(0.021)	(0.029)	(0.020)	(0.030)	(0.031)	(0.027)	(0.0048)	(0.040)
N. of obs.	9,557	8,193	6,784	6,784	6,784	6,784	6,784	6,784
Mean control	.86	.81	.82	.66	.63	.14	.0055	.21
$\alpha_3 = \alpha_2 - \alpha_1$	-.0092	-.024	-.088***	.032	.014	.019	.0057	.053
p-value ($H_0 : \alpha_3 = 0$)	.64	.45	.00062	.3	.66	.5	.22	.31

The outcomes come from inspections our enumerators conducted on student notebooks. The outcome in Column 1 is whether the student had a notebook (=1) or not (=0). The outcome in Column 2 is whether there was a recent (in the past 5 school days) assignment in the notebook (conditional on having the notebook). Column 3 measures the length of the assignment (conditional on having an assignment) in pages (A4 size), Column 4 whether the assignment was graded or not, Column 5 whether there were check-marks written by the teacher in the assignment, Column 6 whether the correct answer was written by the teacher, Column 7 whether the teacher had drawn any pictures (e.g., a smiley face), and Column 8 whether there was any written feedback (e.g., "good work" or "need to work on xx"). This table uses data collected at the end of the second school year of the experiment. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.17 Expected earnings

Figure A.8: Expected earnings across treatments



A.18 Goal setting

We also explore the extent to which the incentives affected teachers' goal-setting behaviour (see Table A.23).[51] 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 incentivised Twaweza test than control group teachers (Column 2). In contrast, teachers in Pay for Percentile schools were 2.5 percentage points (p-value 0.34) more likely to have set goals for the Twaweza test (Column 2). Our surveys also collected information about specific teacher goals on the Twaweza test. Because Twaweza tests were administered in all schools, we collected 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% of my students to pass") for the Twaweza incentivised test, compared to an insignificant increase of about 4 percentage points in Pay for Percentile schools (Column 4). However, these differences are not statistically significant.

[51] Recent papers in behavioural economics provide evidence on general productivity effects of setting goals; for example, Koch and Nafziger (2011); Gómez-Minambres (2012), and Dalton *et al.* (2015).

Table A.23: Goal-setting

	Goals		Twaweza test goals	
	School exam (1)	Twaweza exam (2)	General (3)	Specific (number) (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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.19 Balance in teacher turnover

Table A.24: Teacher turnover

	(1)	(2)
	Still teaching incentivised 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.20 Heterogeneity by previous treatment status

Table A.25: Heterogeneity by whether the school was exposed to teacher incentives in the past

	(1) Year 1	(2) Year 2
Panel A: Non-incentivised test		
Levels × Control in previous RCT (α_1)	.072	-.065
	(.091)	(.064)
Levels × Incentives in previous RCT (α_2)	.06	.16***
	(.054)	(.055)
P4Pctile × Control in previous RCT (β_1)	-.099	-.069
	(.11)	(.081)
P4Pctile × Incentives in previous RCT (β_2)	-.018	.085*
	(.042)	(.05)
N. of obs.	4,781	4,869
p-value($H_0 : \alpha_1 = \alpha_2$)	.91	.0075
p-value($H_0 : \beta_1 = \beta_2$)	.5	.099
p-value($H_0 : \alpha_1 = \beta_1$)	.21	.96
p-value($H_0 : \alpha_2 = \beta_2$)	.096	.18
p-value($H_0 : (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)$)	.51	.55
Panel B: Incentivised test		
Levels × Control in previous RCT (α_1)	-.019	-.052
	(.082)	(.083)
Levels × Incentives in previous RCT (α_2)	.23***	.31***
	(.074)	(.067)
P4Pctile × Control in previous RCT (β_1)	-.084	-.17
	(.13)	(.15)
P4Pctile × Incentives in previous RCT (β_2)	.096	.2***
	(.061)	(.061)
N. of obs.	48,077	59,680
p-value($H_0 : \alpha_1 = \alpha_2$)	.023	.00075
p-value($H_0 : \beta_1 = \beta_2$)	.21	.026
p-value($H_0 : \alpha_1 = \beta_1$)	.67	.51
p-value($H_0 : \alpha_2 = \beta_2$)	.054	.067
p-value($H_0 : (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)$)	.7	.99

The outcome is the composite test scores across math and Kiswahili. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.26: Heterogeneity by whether the school was exposed to teacher incentives in the past and showing the effect of previous treatments

	(1) Year 1	(2) Year 2
Panel A: Non-incentivised test		
Levels × Control in previous RCT (α_1)	.022	-.026
	(.1)	(.07)
Levels × Incentives in previous RCT (α_2)	.012	.13**
	(.046)	(.055)
P4Pctile × Control in previous RCT (β_1)	-.06	-.051
	(.091)	(.089)
P4Pctile × Incentives in previous RCT (β_2)	-.043	.073
	(.045)	(.058)
Incentives in previous RCT	.033	-.019
	(.059)	(.064)
N. of obs.	4,781	4,869
p-value($H_0 : \alpha_1 = \alpha_2$)	.93	.081
p-value($H_0 : \beta_1 = \beta_2$)	.87	.25
p-value($H_0 : \alpha_1 = \beta_1$)	.48	.79
p-value($H_0 : \alpha_2 = \beta_2$)	.26	.31
p-value($H_0 : (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)$)	.83	.78
Panel B: Incentivised test		
Levels × Control in previous RCT (α_1)	-.007	.0051
	(.11)	(.12)
Levels × Incentives in previous RCT (α_2)	.1	.21***
	(.065)	(.067)
P4Pctile × Control in previous RCT (β_1)	-.04	-.23
	(.16)	(.2)
P4Pctile × Incentives in previous RCT (β_2)	.065	.17**
	(.064)	(.067)
Incentives in previous RCT	.041	-.1
	(.087)	(.12)
N. of obs.	48,077	59,680
p-value($H_0 : \alpha_1 = \alpha_2$)	.38	.13
p-value($H_0 : \beta_1 = \beta_2$)	.53	.057
p-value($H_0 : \alpha_1 = \beta_1$)	.84	.2
p-value($H_0 : \alpha_2 = \beta_2$)	.57	.53
p-value($H_0 : (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)$)	.98	.33

The outcome is the composite test scores across math and Kiswahili. This regression only controls partially for the strata of randomization as it only includes fixed effects for two out of the three dimensions for stratification: districts and above/below median baseline scores. The randomization design also stratified by previous RCT treatment status. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.27: Heterogeneity by whether at least one-quarter of the current teachers were previously exposed to incentives

	(1) Year 1	(2) Year 2
Panel A: Non-incentivised test		
Levels × Teachers not incentivized in previous RCT (α_1)	.15**	.03
	(.068)	(.059)
Levels × Teachers incentivized in previous RCT (α_2)	.0022	.15***
	(.054)	(.057)
P4Pctile × Teachers not incentivized in previous RCT (β_1)	-.085	-.13**
	(.072)	(.062)
P4Pctile × Teachers incentivized in previous RCT (β_2)	-.017	.11**
	(.043)	(.054)
Teachers incentivized in previous RCT	-.014	-.12*
	(.069)	(.075)
N. of obs.	4,781	4,869
p-value($H_0 : \alpha_1 = \alpha_2$)	.063	.13
p-value($H_0 : \beta_1 = \beta_2$)	.41	.0042
p-value($H_0 : \alpha_1 = \beta_1$)	.0024	.037
p-value($H_0 : \alpha_2 = \beta_2$)	.7	.54
p-value($H_0 : (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)$)	.015	.19
Panel B: Incentivised test		
Levels × Teachers not incentivized in previous RCT (α_1)	.14*	.15*
	(.076)	(.084)
Levels × Teachers incentivized in previous RCT (α_2)	.21***	.28***
	(.077)	(.07)
P4Pctile × Teachers not incentivized in previous RCT (β_1)	-.057	.0048
	(.08)	(.094)
P4Pctile × Teachers incentivized in previous RCT (β_2)	.11	.19***
	(.067)	(.064)
Teachers incentivized in previous RCT	-.076	-.13
	(.081)	(.1)
N. of obs.	48,077	59,680
p-value($H_0 : \alpha_1 = \alpha_2$)	.43	.18
p-value($H_0 : \beta_1 = \beta_2$)	.09	.091
p-value($H_0 : \alpha_1 = \beta_1$)	.027	.14
p-value($H_0 : \alpha_2 = \beta_2$)	.15	.14
p-value($H_0 : (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)$)	.27	.62

The outcome is the composite test scores across math and Kiswahili. "Teachers incentivised in previous RCT" is equal to one if at least one-quarter of the teachers currently teaching Kiswahili or Math were eligible for incentives in the previous experiment (analysed by Mbiti *et al.* (2019)) in a treatment school. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.21 Test score results for students previously unexposed to incentives

Table A.28: Effect on test scores on students that were not exposed to the teacher incentives in the previous experiment

	(1)	(2)	(3)	(4)	(5)	(6)
		Year 1			Year 2	
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: Non-incentivised test						
Levels (α_1)	-.00021	.087	.073	.012	.11*	.075
	(.071)	(.064)	(.064)	(.045)	(.06)	(.052)
P4Pctile (α_2)	-.034	.0042	-.019	.04	-.031	.0075
	(.057)	(.046)	(.049)	(.041)	(.061)	(.051)
N. of obs.	1,661	1,661	1,661	3,336	3,336	3,336
$\alpha_3 = \alpha_2 - \alpha_1$	-.034	-.083	-.091	.028	-.14**	-.067
p-value ($H_0 : \alpha_3 = 0$)	.62	.18	.14	.59	.029	.24
Panel B: Incentivised test						
Levels (β_1)	.081	.14**	.16**	.13***	.21***	.24***
	(.058)	(.058)	(.076)	(.049)	(.048)	(.061)
P4Pctile (β_2)	.052	-.00041	.037	.097**	.098**	.14**
	(.045)	(.055)	(.066)	(.041)	(.049)	(.06)
N. of obs.	17,871	17,871	17,871	44,222	44,222	44,222
$\beta_3 = \beta_2 - \beta_1$	-.028	-.14**	-.12	-.031	-.12**	-.1*
p-value ($H_0 : \beta_3 = 0$)	0.63	0.022	0.13	0.50	0.011	0.078
Panel C: Incentivised – Non-incentivised						
$\gamma_1 = \beta_1 - \alpha_1$	.072	.045	.073	.11	.095	.15
p-value($\gamma_1 = 0$)	.27	.46	.31	.018	.069	.0066
$\gamma_2 = \beta_2 - \alpha_2$	.08	-.0076	.05	.053	.12	.12
p-value($\gamma_2 = 0$)	.17	.89	.46	.22	.023	.029
$\gamma_3 = \beta_3 - \alpha_3$	.0079	-.053	-.023	-.055	.026	-.032
p-value($\gamma_3 = 0$)	.9	.41	.75	.23	.65	.58
$\gamma_1 - \gamma_2$	-.0079	.053	.023	.055	-.026	.032
p-value($\gamma_1 - \gamma_2 = 0$)	.9	.41	.75	.23	.65	.58
$\gamma_1 - \gamma_3$	.08	-.0076	.05	.053	.12	.12
p-value($\gamma_1 - \gamma_3 = 0$)	.17	.89	.46	.22	.023	.029
$\gamma_2 - \gamma_3$	.088	-.06	.026	-.0023	.15	.088
p-value($\gamma_2 - \gamma_3 = 0$)	.38	.56	.83	.98	.13	.37
p-value($\gamma_1 = \gamma_2 = \gamma_3$)	.34	.68	.57	.059	.049	.014

Results from estimating Equation 2 for different subjects at both follow-ups. The sample only includes Grade 1 students in year 1, and Grade 1 and Grade 2 students in year 2. Panel A uses data from the non-incentivised test taken by a sample of students. Panel B uses data from the incentivised test taken by all students. Control variables in both panels 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. Table A.11 provides a version without school controls. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.22 Test score results for schools with incentives in the previous experiment

Table A.29: Effect on test scores across schools that were exposed to incentives in the past

	(1)	(2)	(3)	(4)	(5)	(6)
	Year 1			Year 2		
	Math	Kiswahili	Combined	Math	Kiswahili	Combined
Panel A: Non-incentivised test						
Levels (α_1)	.035	.043	.054	.11**	.15**	.15***
	(.052)	(.055)	(.053)	(.048)	(.063)	(.053)
P4Pctile (α_2)	-.0007	-.023	-.016	.11**	.049	.086*
	(.042)	(.044)	(.042)	(.043)	(.056)	(.049)
N. of obs.	3,733	3,733	3,733	3,801	3,801	3,801
$\alpha_3 = \alpha_2 - \alpha_1$	-.036	-.066	-.07	-.0041	-.11*	-.065
p-value ($H_0 : \alpha_3 = 0$)	.42	.16	.11	.93	.057	.2
Panel B: Incentivised test						
Levels (β_1)	.13**	.15***	.2***	.2***	.23***	.31***
	(.054)	(.055)	(.072)	(.053)	(.052)	(.068)
P4Pctile (β_2)	.08*	.039	.085	.14***	.13***	.2***
	(.043)	(.048)	(.06)	(.044)	(.049)	(.061)
N. of obs.	40,437	40,437	40,437	49,753	49,753	49,753
$\beta_3 = \beta_2 - \beta_1$	-.053	-.11**	-.12*	-.058	-.098**	-.11*
p-value ($H_0 : \beta_3 = 0$)	0.28	0.029	0.076	0.21	0.045	0.076
Panel C: Incentivised – Non-incentivised						
$\gamma_1 = \beta_1 - \alpha_1$	.088	.096	.13	.083	.066	.14
p-value($\gamma_1 = 0$)	.069	.055	.018	.085	.24	.02
$\gamma_2 = \beta_2 - \alpha_2$	.076	.056	.094	.036	.079	.1
p-value($\gamma_2 = 0$)	.11	.26	.11	.43	.092	.053
$\gamma_3 = \beta_3 - \alpha_3$	-.012	-.04	-.039	-.048	.013	-.036
p-value($\gamma_3 = 0$)	.79	.38	.46	.28	.8	.53
$\gamma_1 - \gamma_2$	.012	.04	.039	.048	-.013	.036
p-value($\gamma_1 - \gamma_2 = 0$)	.79	.38	.46	.28	.8	.53
$\gamma_1 - \gamma_3$	.076	.056	.094	.036	.079	.1
p-value($\gamma_1 - \gamma_3 = 0$)	.11	.26	.11	.43	.092	.053
$\gamma_2 - \gamma_3$	.064	.017	.055	-.012	.093	.066
p-value($\gamma_2 - \gamma_3 = 0$)	.41	.84	.57	.87	.27	.48
p-value($\gamma_1 = \gamma_2 = \gamma_3$)	.15	.16	.059	.22	.22	.042

Results from estimating Equation 2 for different subjects at both follow-ups. The sample excludes schools that were control schools in the previous experiment. Panel A uses data from the non-incentivised test taken by a sample of students. Panel B uses data from the incentivised test taken by all students. Control variables in both panels 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. Table A.11 provides a version without school controls. Standard errors, clustered at the school level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B Randomisation details

This study builds on the sample of 350 schools that participated in the 2013 to 2014 KiuFunza study (see Mbiti *et al.* (2019) 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. To determine teacher awards, incentivised tests were conducted in schools assigned to the incentives or combination treatment (a total of 140 schools). To facilitate the computation of treatment effects on incentivised tests, we also conducted these tests in 40 control schools.

We take the 180 schools where endline "incentivised" tests were conducted in 2014. Specifically, the experimental sample for this experiment includes 70 schools from the incentive arm in the previous experiment, 70 schools from the combination arm, and 40 schools from the control group. We use these tests as the baseline data to implement the teacher incentive schemes in this study. This baseline data is essential for the Pay for Percentile incentive scheme as we have to split students into groups and properly seed each contest.

Each district had seven schools in the "teacher incentives" arm, seven in "combination" (incentives and inputs), and four in the control group. We randomly assign schools from the previous treatment groups into two new treatment groups (Levels or Pay for Percentile) and a control group. However, to study the long-term impacts of teacher incentives (in a companion paper), we assigned a higher proportion of schools in the "teacher incentives" (which involved threshold teacher incentives) to Levels. Similarly, we assign a higher proportion of schools in the control group from the previous experiment 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. We created an overall measure of student learning and categorised schools as above or below the median. Table B.1 summarises the number of schools randomly allocated to each treatment arm based on their assignment in the previous experiment. Each district has 18 schools, six in each new experimental group (Levels, Pay for Percentile, and control).

Because the study was carried out in 10 districts, there are 60 schools in each new treatment group: 30 above the median in baseline learning and 30 below.

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

Table B.1: Treatment allocation

		KiuFunza II (current experiment)			
		Levels	P4Pctile	Control	Total
KiuFunza I (previous experiment)	Teacher incentives	40	20	10	70
	Inputs+Incentives	10	30	30	70
	Control	10	10	20	40
	Total	60	60	60	180

C Additional details on the incentive designs

C.1 Pay for percentile groups

As mentioned above, a necessary condition for the Pay for Percentile to deliver optimal levels of effort is that teachers believe they are competing in fair contests (Barlevy and Neal, 2012). This requires the tournaments to be adequately seeded, which Barlevy and Neal (2012) defines as: "creating comparison sets containing classrooms with common measured resources, equally experienced teachers, and identical distributions of baseline student achievement." In practice, this is impossible, but we attempt to approximate this by creating comparison sets with students with similar achievement levels.

In Grade 1, since none of the students had incoming test scores, we created ability groups based on the school's historical average test scores. As a result, all students in the same school belonged to the same group, even if each group included students from different schools. We created ten groups, each with 6–7 schools — see Table C.1 for the total number of students in Grade 1 in each group in year 1 and Table C.6 for year 2. The same group was used for all subjects.

In Grades 2 and 3, we attempted to create ten equally sized groups each year based on the test scores obtained at the end of the previous year. However, since the incentivised (or "high-stakes") is not as granular, it presents some bottom and top coding. Thus, it was not always possible to create ten equally sized groups; instead, the bottom and the top groups are usually slightly larger. These groups are created using student-level information, and thus students in the same school may not belong to the same group, and each group includes students from different schools.

In the first year, we had nine ability groups in both Kiswahili and Math (see Tables C.2 and C.3) in Grade 2. In Grade 3, we had seven ability groups in Kiswahili and ten groups in Math (see Tables C.4 and C.5). There was an additional group (implicitly) with all the students for whom we had no test-score information from the previous year and thus could not be placed in any group. This last group accounted for approximately 20% of the students tested at the end of the school year in both grades.

In the second year, we had nine ability groups in both Kiswahili and Math (see Tables C.7 and C.8) in Grade 2. In Grade 3, we had seven ability groups in Kiswahili and ten groups in Math (see Tables C.4 and C.5). As before, there was an additional group (implicitly) with all the students for whom we had no test-score information from the previous year and thus could not be placed in any group. This last group accounted for approximately 20% of the students tested at the end of the school year.

Table C.1: Groups — Grade 1 — Year 1

Ability	Students	%
Group 1	599	7.09
Group 2	954	11.29
Group 3	767	9.08
Group 4	935	11.07
Group 5	982	11.62
Group 6	824	9.75
Group 7	979	11.59
Group 8	594	7.03
Group 9	925	10.95
Group 10	889	10.52

Group numbers go from lowest ability (group 1) to highest ability.

Table C.2: Kiswahili groups — Grade 2 — Year 1

Ability	Students	%
Group 1	631	10.03
Group 2	628	9.98
Group 3	631	10.03
Group 4	658	10.46
Group 5	613	9.74
Group 6	614	9.76
Group 7	629	10.00
Group 8	651	10.35
Group 9	1,236	19.65

Group numbers go from lowest ability (group 1) to highest ability.

Table C.3: Math groups — Grade 2 — Year 1

Ability	Students	%
Group 1	639	10.16
Group 2	696	11.06
Group 3	553	8.79
Group 4	629	10.00
Group 5	633	10.06
Group 6	638	10.14
Group 7	648	10.30
Group 8	1,267	20.14
Group 10	588	9.35

Group numbers go from lowest ability (group 1) to highest ability.

Table C.4: Kiswahili groups — Grade 3 — Year 1

Ability	Students	%
Group 1	554	10.14
Group 2	545	9.97
Group 3	541	9.90
Group 4	551	10.08
Group 5	1,232	22.55
Group 7	491	8.99
Group 8	1,550	28.37

Group numbers go from lowest ability (group 1) to highest ability.

Table C.5: Math groups — Grade 3 — Year 1

Ability	Students	%
Group 1	547	10.01
Group 2	547	10.01
Group 3	546	9.99
Group 4	566	10.36
Group 5	535	9.79
Group 6	544	9.96
Group 7	540	9.88
Group 8	563	10.30
Group 9	698	12.77
Group 10	378	6.92

Group numbers go from lowest ability (group 1) to highest ability.

Table C.6: Groups — Grade 1 — Year 2

Ability	Students	%
Group 1	931	9.35
Group 2	957	9.61
Group 3	987	9.91
Group 4	1,089	10.93
Group 5	1,319	13.24
Group 6	1,010	10.14
Group 7	967	9.71
Group 8	753	7.56
Group 9	1,239	12.44
Group 10	709	7.12

Group numbers go from lowest ability (group 1) to highest ability.

Table C.7: Kiswahili groups — Grade 2 — Year 2

Ability	Students	%
Group 1	999	10.03
Group 2	996	10.00
Group 3	997	10.01
Group 4	993	9.97
Group 5	996	10.00
Group 6	996	10.00
Group 7	996	10.00
Group 8	999	10.03
Group 9	1,989	19.97

Group numbers go from lowest ability (group 1) to highest ability.

Table C.8: Math groups — Grade 2 — Year 2

Ability	Students	%
Group 1	1,128	11.32
Group 2	1,104	11.08
Group 3	1,090	10.94
Group 4	1,112	11.16
Group 5	1,102	11.06
Group 6	1,192	11.97
Group 7	2,077	20.85
Group 8	492	4.94
Group 9	664	6.67

Group numbers go from lowest ability (group 1) to highest ability.

Table C.9: Kiswahili groups — Grade 3 — Year 2

Ability	Students	%
Group 1	723	11.12
Group 2	776	11.93
Group 3	669	10.29
Group 4	947	14.56
Group 5	499	7.67
Group 6	804	12.36
Group 7	2,086	32.07

Group numbers go from lowest ability (group 1) to highest ability.

Table C.10: Math groups — Grade 3 — Year 2

Ability	Students	%
Group 1	767	11.79
Group 2	552	8.49
Group 3	664	10.21
Group 4	689	10.59
Group 5	631	9.70
Group 6	609	9.36
Group 7	793	12.19
Group 8	577	8.87
Group 9	644	9.90
Group 10	578	8.89

Group numbers go from lowest ability (group 1) to highest ability.

C.2 Proficiency thresholds (Levels)

In the Levels design, the skills thresholds were salient milestones based on the national curriculum, ranging from basic (e.g., number recognition) to more complex skills (e.g., multiplication) to allow teachers to earn rewards from a wide range of students.

Table C.11 shows what percentile of the (incentivised) test-score distribution corresponds to each threshold. Specifically, we estimate the minimum — 1st percentile to avoid the results being driven by outliers — standardised scores of students who pass the thresholds. While there is some variance across subjects and grades, overall, thresholds are spread across the ability distribution. For example, for Grade 2 in the

second year, the "easiest" threshold for Kiswahili is placed at the 17th percentile, and the "hardest" one is at the 47th percentile.

Table C.11: Where are thresholds located in the test score distribution

Threshold	Percentile year 1	Percentile year 2
Kiswahili – Grade 1		
Letters	49.10	50.72
Words	48.11	48.87
Sentences	57.76	58.59
Math – Grade 1		
Counting	4.94	4.80
Numbers	13.04	9.80
Inequalities	8.01	6.73
Addition	23.86	26.99
Subtraction	29.12	33.93
Kiswahili – Grade 2		
Words	27.20	17.90
Sentences	33.12	25.42
Paragraphs	42.46	47.56
Math – Grade 2		
Inequalities	12.07	6.41
Addition	32.57	22.20
Subtraction	34.20	23.81
Multiplication	37.48	99.90
Kiswahili – Grade 3		
Story	26.60	18.74
Comprehension	39.77	30.67
Math – Grade 3		
Addition	26.23	23.22
Subtraction	37.48	27.12
Multiplication	60.27	49.93
Division	39.43	35.32

Note: This table shows where in the test score distribution lies the students with the lowest score that was able to achieve the skill — approximating the percentile of the test score distribution corresponding to each threshold.

D Theoretical Framework

We present a set of simple models to clarify the potential behavioural responses of teachers and schools in our interventions. We first characterise equilibrium effort levels of teachers in both incentive systems, 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. Our model builds on the framework of Barlevy and Neal (2012).

D.1 Basic Setup

Our simple setup has different types of students (indexed by l). Students may vary by the 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 the endline are 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. We assume $a_{j(t-1)}^l$ is an adequate summary statistic for all previous inputs, including past teacher effort. The productivity of teacher effort (e_j^l) is captured by γ^l and is assumed to be constant across teachers. In other words, we assume teachers are equally capable — Barlevy and Neal (2012) also impose this assumption in their basic setup. 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 utilitarian social planner would choose teacher effort to maximise 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:

$$\gamma^l = c_l'(e_j^l) \tag{D.1}$$

for all l and all j. To keep the model simple, we assume teachers are risk-neutral and abstract from multi-tasking concerns. To keep notation simple, we assume all teachers have identical productivity; however, this can easily be relaxed without altering the results presented below.

D.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 groupings, such that students in the same group are similar to each other. Under this incentive scheme, teachers maximise their expected payoffs, net of costs, from each rank-order tournament. The teacher's maximisation 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

$$(N-1)\pi\gamma^l f^l(0) = c_l'(e^l) \tag{D.2}$$

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 maximise total learning and the payoff is $\pi = \frac{1}{(N-1)f^l(0)}$.

D.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 maximisation 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., teachers ignore the effect of their effort on the overall pass rate). Thus, the first order conditions for the teacher's maximisation problem become:

$$\sum_t C_j^l \gamma^l h^l (T_t - a_{j(t-1)}^l - \gamma^l e_j^l) \frac{\Pi_t}{\sum_l \sum_n C_n^l P\left(v_n^l > T_t - a_{n(t-1)}^l - \gamma^l e_n^l\right)} = c_l'(e_j^l) \tag{D.3}$$

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 characterise our symmetric equilibrium as:

$$\sum_t C_j^l \gamma^l h^l (T_t - a_{j(t-1)}^l - \gamma^l e^l) \frac{\Pi_t}{\sum_l N C_n^l P\left(v^l > T_t - a_{(t-1)}^l - \gamma^l e^l\right)} = c_l'(e^l) \tag{D.4}$$

for all l.

D.1.3 Numerical Simulation Set-up

We simulate the equilibrium responses by teachers to both types of incentives to better understand teacher behavioural 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 classroom. We assume that

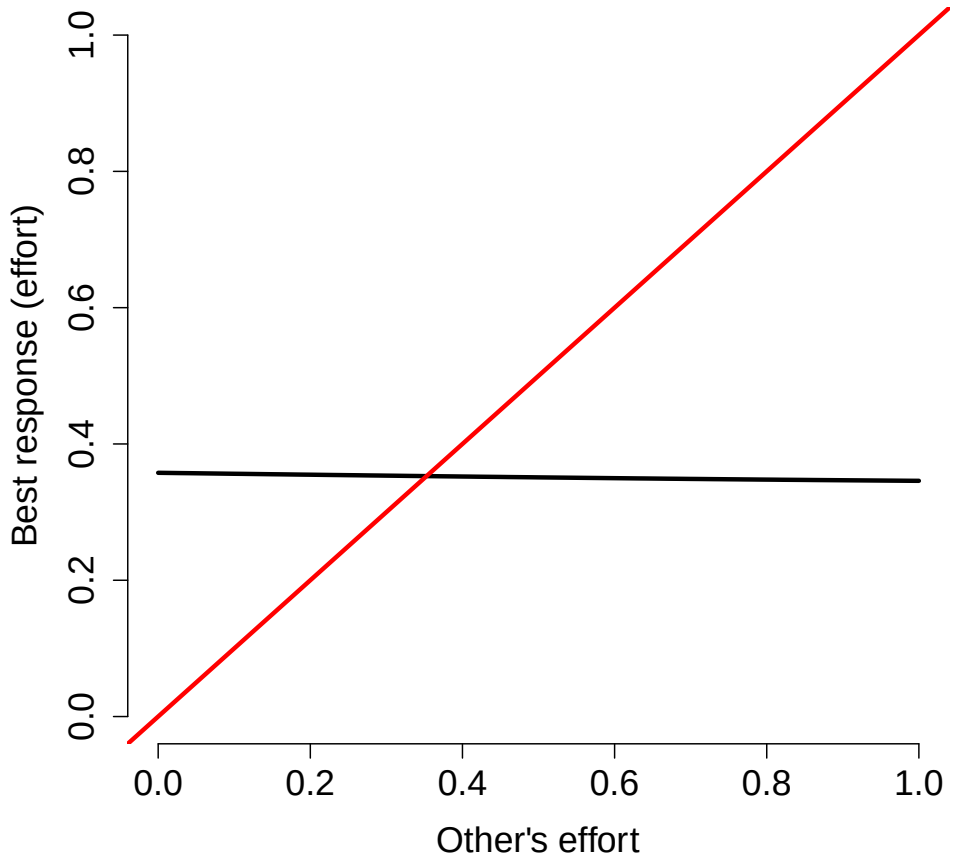
student baseline learning levels are uniformly distributed within each class from -4 to 4, in 0.5 intervals. As a result, each classroom has 17 students, with one student at each (discrete) baseline learning level. In Appendix D.2 we show that our qualitative results are robust to a normal distribution of student baseline learning levels.

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 D.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.

D.1.4 Levels Equilibrium

We first simulate equilibrium behaviour under the Levels scheme in Figure D.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 D.1 shows, in the context of our of simulations, there is only one (rational expectations) equilibrium characterised by Equation D.4.

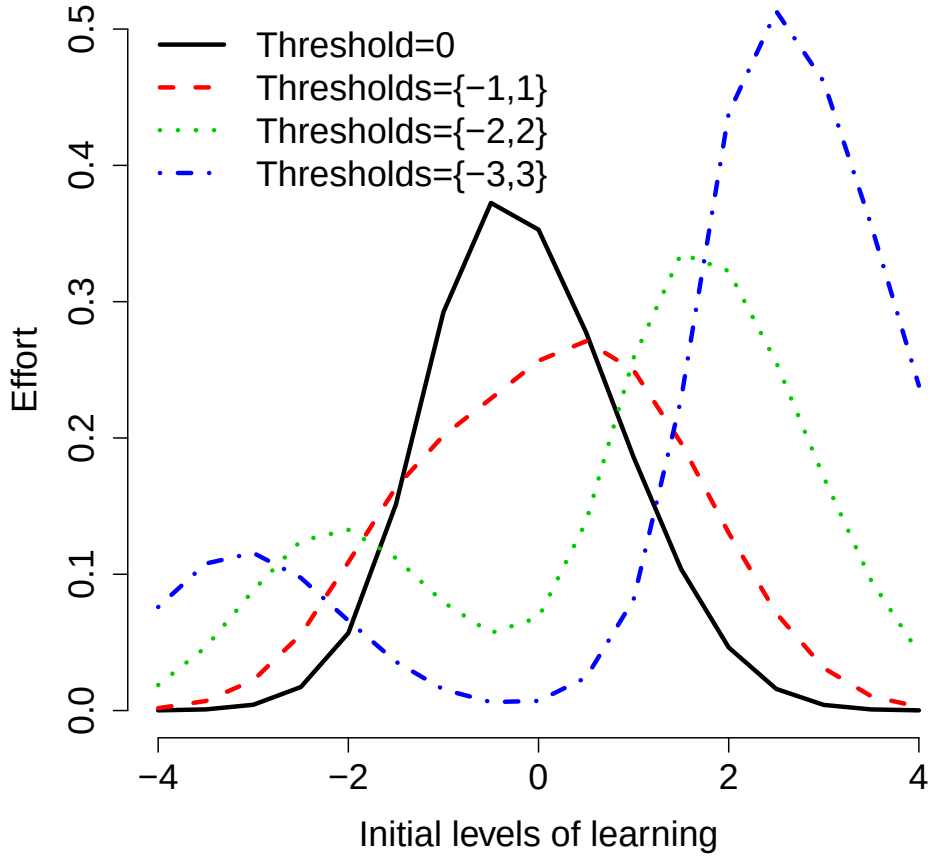
Figure D.1: Teacher $i's$ Best Response curve to other teacher's effort level



Note: An example of a set of best response curves for given initial parameter values. We assume all teachers are giving the same value of effort for all thresholds except one (but the effort may differ across thresholds). In the x-axis, we show the effort exerted by all except i in the threshold of interest. In the y-axis, we plot teacher i effort level in that threshold. 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 an 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 D.2.

Figure D.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).

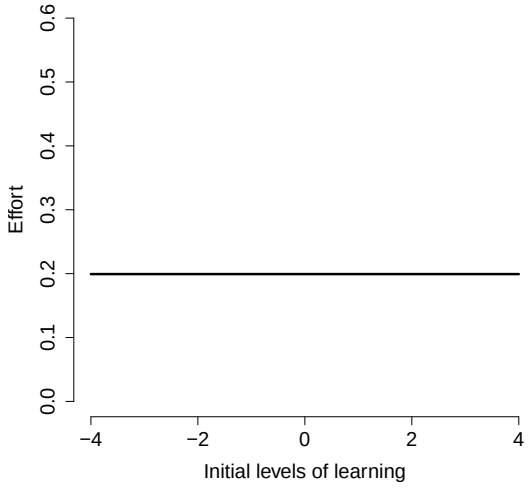
As the equilibrium behaviour for teachers under Pay for Percentile is described in detail in Barlevy and Neal (2012), we refer our readers to consult their findings for additional insights.

D.1.5 A Comparison of Optimal Teacher Effort

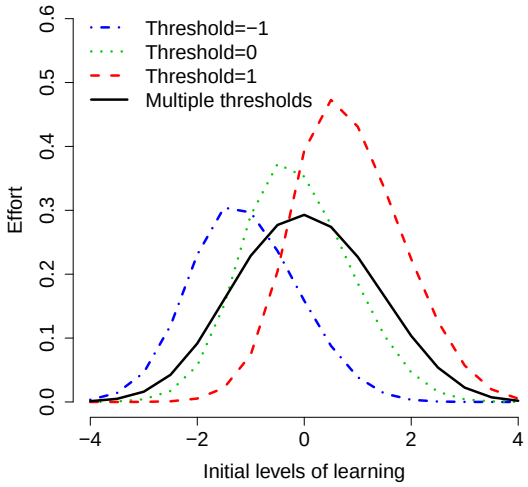
We compute equilibrium teacher responses under two different stylised scenarios (or assumptions about the productivity of teacher effort in the production function) to illustrate how changes in these assumptions can alter equilibrium responses. This exercise aims to highlight the impact of the production 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 D.3a and D.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 D.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. Using numerical integration techniques, we can compute total teacher effort, and thus total learning gains under both incentive systems. Our computations show that total effort is higher under Pay for Percentile than Levels (3.39 vs 1.97).

Figure D.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.



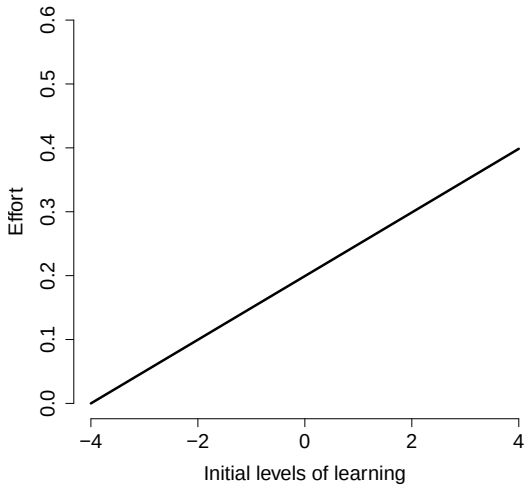
(b) Levels - γ constant across initial levels of learning. The total effort exerted by teachers is 1.55 under the -1 threshold, 1.88 under the 0 threshold, 2.37 under the 1 threshold, and 1.97 under multiple thresholds.

We relax the assumption of constant productivity of teacher effort and allow it to vary with the 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 $\gamma^l = 1 + 0.25a^l_{(t-1)}$. 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.

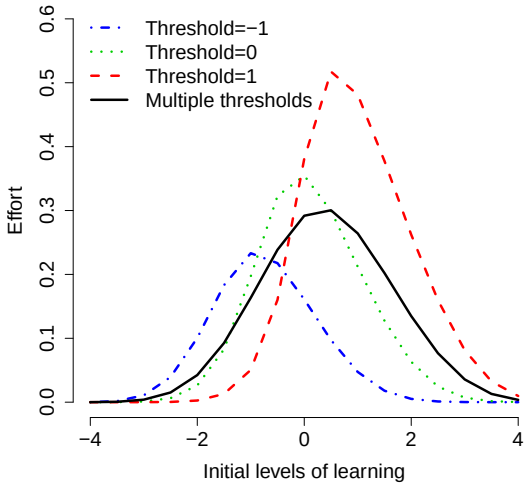
Figures D.4a and D.4b show the numerical solutions of optimal teacher effort for different initial levels of student learning. In the Pay for Percentile system, focusing on better-prepared students increases the likelihood of winning the rank-order contest (among that group of students), while the marginal 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. In this scenario, total teacher effort and, thus, total learning gains is also higher under Pay for Percentile compared to Levels (3.39 vs. 1.88).

Our numerical exercise suggests that testing for equality of treatment effects across the distribution of student baseline test scores in the Pay for Percentile arm allows us to shed light on the role of teacher effort in the education production function.

Figure D.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 multiple thresholds.

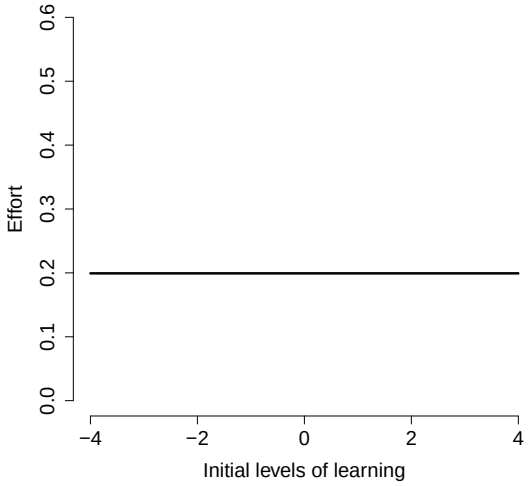
D.2 Robustness of Simulation Results

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 D.1.5. 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 — specifically, we assume a binomial distribution centred around zero. Figures D.5 and D.6 show the optimal effort of teachers across both incentive schemes.

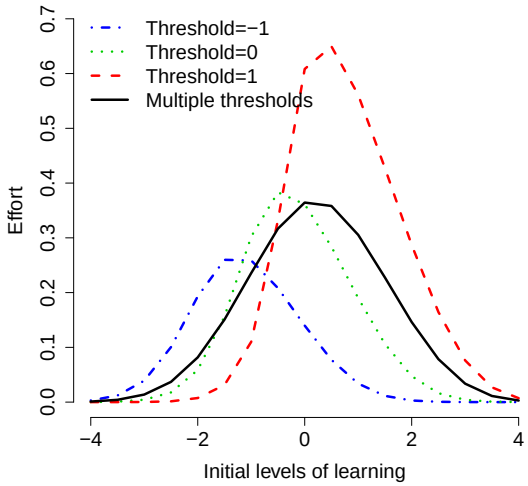
As seen in the figures below, teacher responses are equal in the pay for percentile scheme (P4Pctile) regardless of the distribution of baseline student learning. This result is unsurprising given the equilibrium condition in Equation D.2. On the other hand, for the proficiency scheme (Levels) the optimal

teacher effort changes when the distribution of baseline test scores changes (see Equation D.4). However, qualitatively the result is the same as with a uniform distribution of baseline test scores.

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

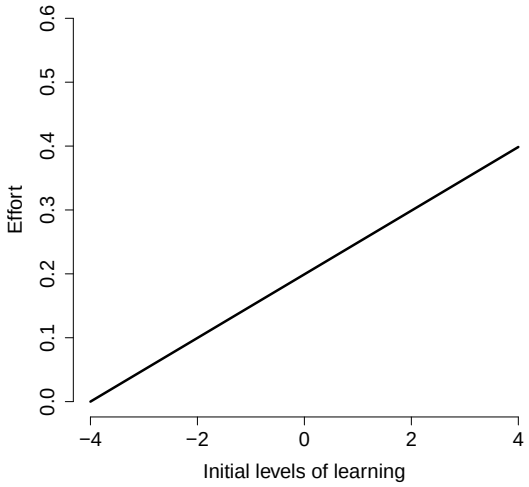


(a) P4Pctile - γ constant across initial levels of learning

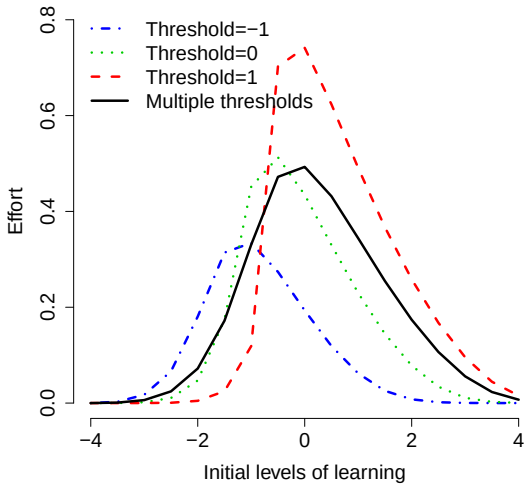


(b) Levels - γ constant across initial levels of learning

Figure D.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

E Test Design

Tanzanian education professionals developed the tests used in this evaluation. 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 (see https://www.twaweza.org/go/uwezo). The test developers developed two types of tests: a non-incentivised (or low-stakes) test that was used for research purposes and an incentivised (or high-stakes) test that Twaweza used to determine teacher bonuses. Both tests followed the testing procedures and protocols established in Mbiti *et al.* (2019).

E.1 Non-Incentivised test

The non-incentivised (or low-stakes) test was administered to a sample of 30 students in each school (10 students each from Grades 1 through 3). An additional 10 students from Grade 4 were also tested to test for spillovers. Sampled students are followed throughout the two-year study, except Grade 4 students who were not followed into Grade 5. These non-incentivised tests were only used for research purposes. To prevent confusion in schools, these non-incentivised tests were conducted by a separate team from the intervention team (or the incentivised tests). Since students in the early grades are still learning to write, written tests are not standard. We, therefore, 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 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. Students were allowed to use pencil and paper for the numeracy and spelling questions. 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 experiment analysed by Mbiti *et al.* (2019), 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 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 we expected them to be enrolled in. Both tests were grade-specific tests designed to measure the main competencies outlined in the curriculum. The content of the tests is summarized in Table E.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 for grade 2, and 9 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 for grade 2, and 24 for grade 3. In the second year, the number of items on the grade 1 math test was reduced. However, we added several easier items to the grade 3 test and left the length of the grade 2 test unchanged.

We standardise 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.

E.2 Incentivised test

The incentivised (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 "incentivised tests" in control schools so that we could compute treatment effects using the incentivised test data.

The incentivised test items were used as the core of the non-incentivised test; the latter is an extended version of the former (see next sub-section). The two test types are very similar: both test core curriculum skills in the same sequence in a one-on-one setting. However, because of the large number of students, test time was more limited in the incentivised test, and the number of test skills was slightly lower than in the non-incentivised test. The incentivised test had a core skills section that was administered in both

Levels and P4Pctile schools. A lower level and a higher-level skills section were added for the P4Pctile tests to allow for a broad spectrum of performance percentiles.

Several 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 was 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 the day of the week and the first letter of the student's name. Finally, tests were handled, administered, and scored by Twaweza without teacher involvement. Several checks were done ex-post by Twaweza to verify there had been no cheating on the high-stakes test.

E.3 Comparability of tests

Both types of tests followed the same test-development framework. As a result, the subject order, question type, and phrasing were similar across both tests. The main difference is that the incentivised test is shorter (about 15 minutes per student) and uses various stopping rules to reduce testing time. The non-incentivised 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 E.1. Indeed, the test information functions (Figures E.1-E.6) reveal that while both sets of exams can identify student's ability with roughly the same error for students near the middle of the distribution — the test information function shows the accuracy of the ability estimates for different abilities. It is inversely related to the measurement error. See van der Linden and Hambleton (2013) for more details. However, the non-incentivised test has more information to identify students with more than two standard deviations above or below the mean.

Although the content between the two types of tests is similar, there are several important differences in the administration of the tests. The non-incentivised tests included an "other subject" module to measure potential spillover effects. Non-incentivised tests were administered by taking sampled students out of their classrooms during a regular school day. In contrast, the incentivised tests were more "official" as all students in Grades 1-3 were tested on a prearranged test day. Students in other grades would sometimes

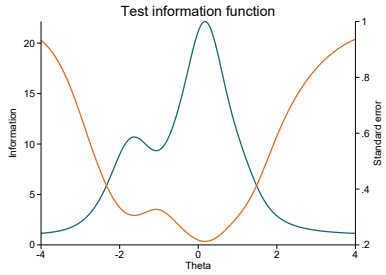
be sent home on the test day to avoid distractions. Extra-curricular activities were also cancelled during the Twaweza test. In addition, most schools used the incentivised test as the end-of-year test. This also likely encouraged students in the control group to exert effort on the test.

Table E.1: Comparison of non-incentivised (low-stakes) and incentivised (high-stakes) test content

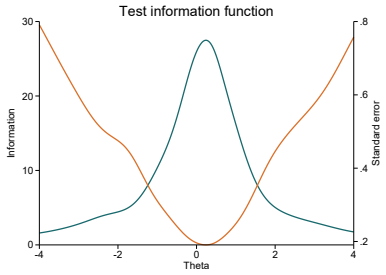
	Non-incentivised (low-stakes)						Incentivised (high-stakes)		
	Year 1			Year 2			Both Years		
	Kiswahili			Kiswahili			Kiswahili		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
Syllables	+	-	-	+	+	+	+	-	-
Words	+	+	-	+	+	+	+	+	-
Sentences	+	+	-	+	+	+	+	+	-
Writing words	+	+	+	-	-	-	-	-	-
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 summarises 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 the low-stakes test is that the high-stakes test is designed to measure proficiency, so the test has various stopping rules to reduce testing time.

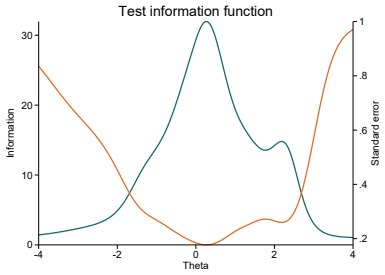
Figure E.1: Test information function - Math - Year 1



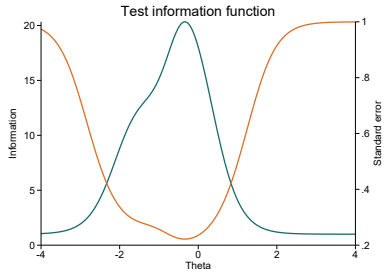
(a) Non-Incentivised - Grd 1



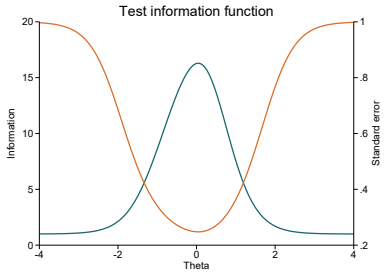
(b) Non-Incentivised - Grd 2



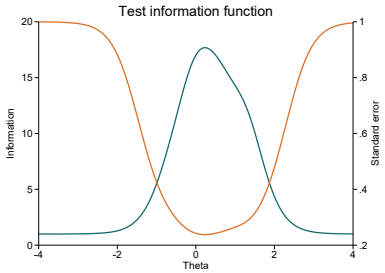
(c) Non-Incentivised - Grd 3



(d) Incentivised - Grd 1

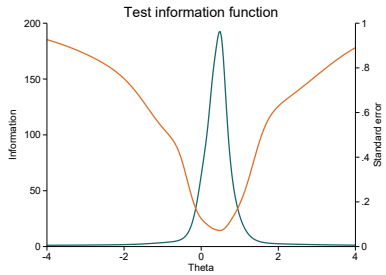


(e) Incentivised - Grd 2

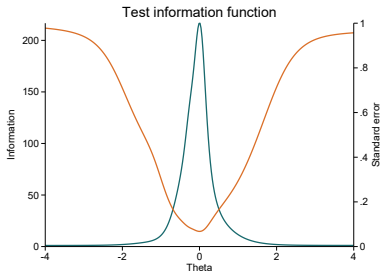


(f) Incentivised - Grd 3

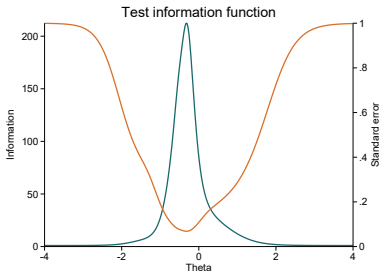
Figure E.2: Test information function - Kiswahili - Year 1



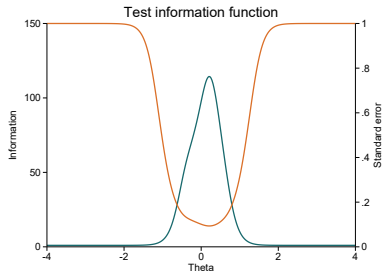
(a) Non-Incentivised - Grd 1



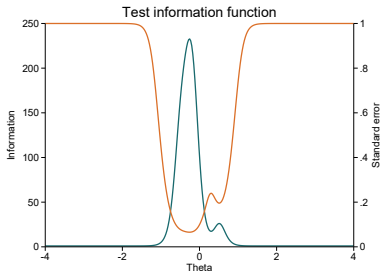
(b) Non-Incentivised - Grd 2



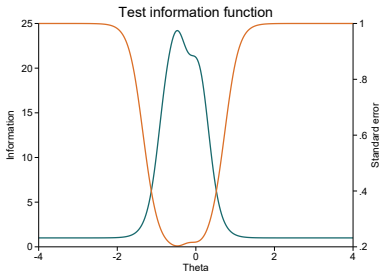
(c) Non-Incentivised - Grd 3



(d) Incentivised - Grd 1

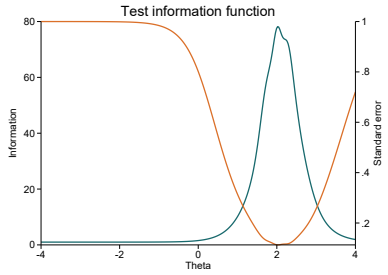


(e) Incentivised - Grd 2

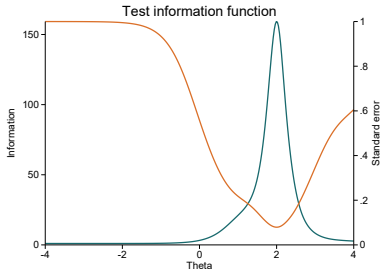


(f) Incentivised - Grd 3

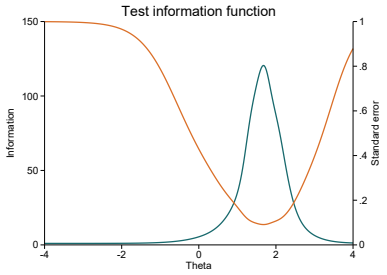
Figure E.3: Test information function - English - Year 1



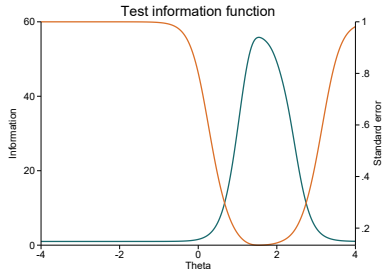
(a) Non-Incentivised - Grd 1



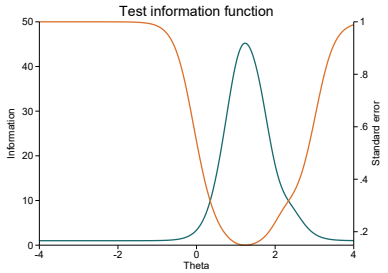
(b) Non-Incentivised - Grd 2



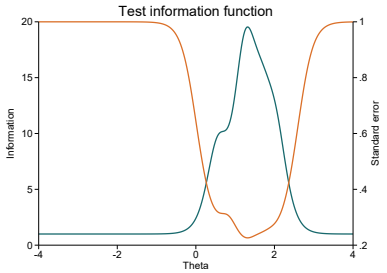
(c) Non-Incentivised - Grd 3



(d) Incentivised - Grd 1

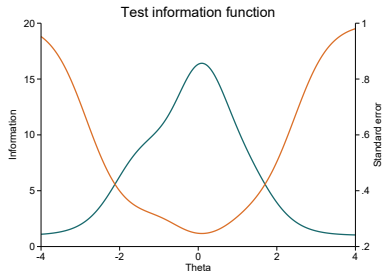


(e) Incentivised - Grd 2

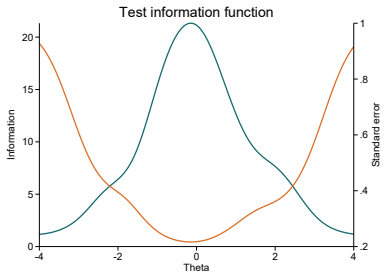


(f) Incentivised - Grd 3

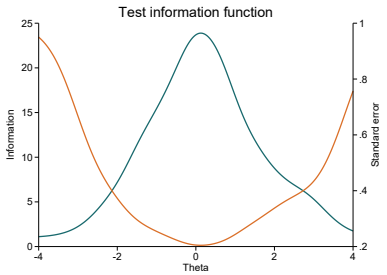
Figure E.4: Test information function - Math - Year 2



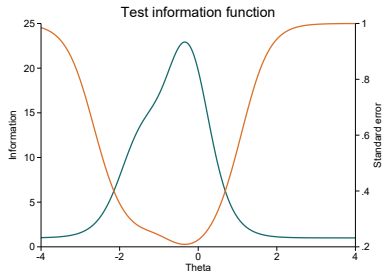
(a) Non-Incentivised - Grd 1



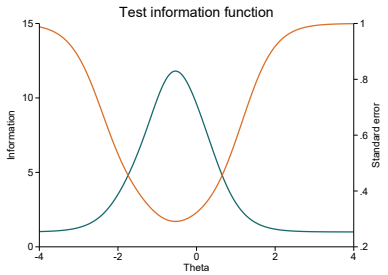
(b) Non-Incentivised - Grd 2



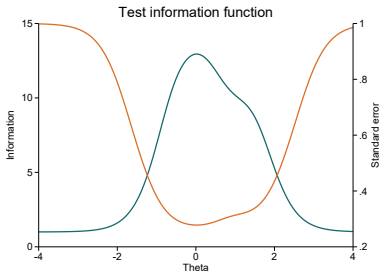
(c) Non-Incentivised - Grd 3



(d) Incentivised - Grd 1

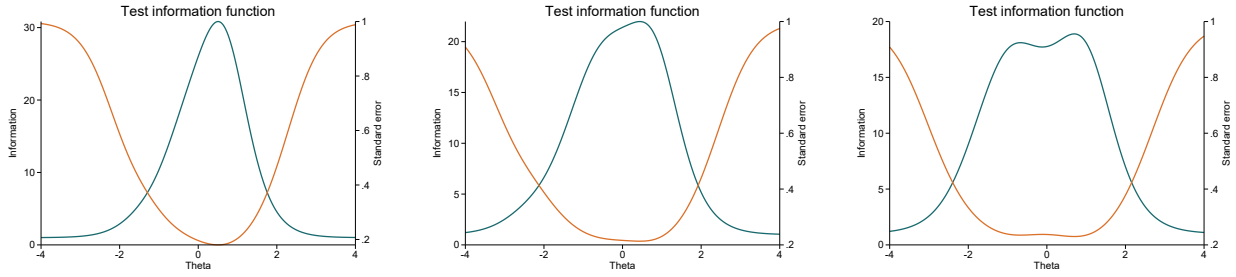


(e) Incentivised - Grd 2

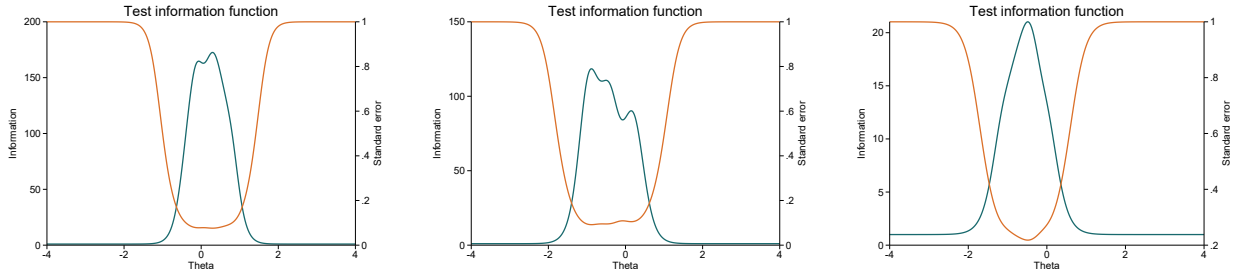


(f) Incentivised - Grd 3

Figure E.5: Test information function - Kiswahili - Year 2

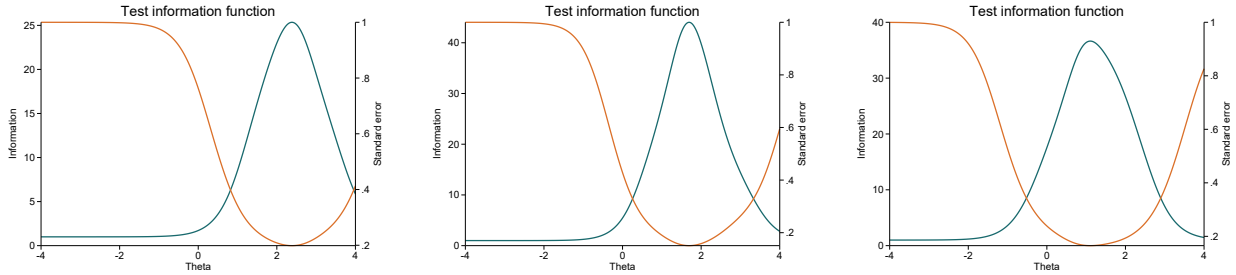


(a) Non-Incentivised - Grd 1 (b) Non-Incentivised - Grd 2 (c) Non-Incentivised - Grd 3

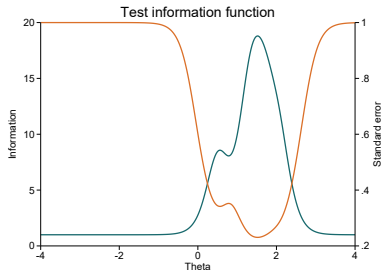


(d) Incentivised - Grd 1 (e) Incentivised - Grd 2 (f) Incentivised - Grd 3

Figure E.6: Test information function - English - Year 2



(a) Non-Incentivised - Grd 1 (b) Non-Incentivised - Grd 2 (c) Non-Incentivised - Grd 3



(d) Incentivised - Grd 1 (e) Incentivised - Grd 2 (f) Incentivised - Grd 3

F Communication materials to explain the interventions to teachers

The translation below, from the original materials in Kiswahili, was provided by Joseph Mmbando.

F.1 Gains

KiuFunza 2015

Improving Learning in Tanzania

Information - Cash on Delivery Competition.

Twaweza, a citizen centered initiative, in cooperation with the Commission for Science and Technology and the Government of Tanzania is working to find ways to improve learning in schools. In order to achieve this, scientific trials known as KiuFunza are being implemented to test various education policies.

KiuFunza was inaugurated by the Minister of State, The Prime Minister's Office, Regional Administration and Local Government (PMO-RALG) Hon. Hawa Ghasia (Mp) in January 2013.

In 2015 these scientific trials will be implemented in 21 Local Government Authorities (LGAs). As part of the trials a new program called KiuFunza Mashindano is implemented in 66 schools in these LGAs.

Table 1 provides information on all participating LGAs. If these trials are successful the government will consider to scale up to cover more LGAs.

Table 1: Number of Mashindano schools in KiuFunza LGAs

Names of Local Government Authorities	Number of Schools in the Competition
Geita and Nyang'hwale	6
Kahama, Ushetu, na Msalala	6
Karagwe and Kyerwa	6
Kigoma and Uvinza	5
Kinondoni	7
Kondoa and Chemba	6
Korogwe	6
Lushoto and Bumbuli	6
Mbozi and Momba	6
Sumbawanga and Kalambo	6
Mbinga and Nyasa	6
Total	**66**

The selection of Local Government Authorities and schools was done through a lottery. Your school was chosen in this lottery to be part of the Mashindano program. This flyer explains what this program is using questions and answers.

1. Last year my school was already part of a Twaweza program: is this the same program?

No, but some features of the program are the same. It has the same goal: to improve reading Kiswahili, English and doing Arithmetic in standards 1, 2 and 3. The program will use a bonus payment for teachers and head teachers in these subjects. The bonus is paid on top of the normal salary. You have been invited to participate because you teach one or more of these topics in these grades.

2. What do I need to do now to participate?

Today you just need to take time to understand the Mashindano program and confirm your participation. Today is an important training day where Twaweza representatives explain how you as a beneficiary of the program can maximize your benefit.

3. What is different this year?

There are three main differences:

1) This year **improvements** in student skills are appreciated (not passing a complete test). Even small improvements in your students' learning may earn you bonus money;
2) You are in a **competition** with other teachers for the bonus;
3) Improvements for all **students at all levels** can earn you a bonus (not just students passing the exam).

4. What do you mean by an improvement in student learning?

KiuFunza Mashindano uses two skill tests for each student: test 1 measures the skill level of the student at the start of the year and test 2 measures the skill level at the end of the year. Test 2 is at the end of 2015; test 1 was done at the end of last year. If level 2 is higher than level 1 there is an improvement in student learning. For example, at test 1 a student could read words but not a paragraph; during 2015 the teacher teaches her to read a paragraph; at test 2 the student can read a paragraph. This is an improvement that will be measured and appreciated with a bonus. For Grade 1 students, we do not have test results for test 1, since they were not in school in 2014. For these students, we take a reliable average measure to replace test 1 for that student.

5. What type of test will you use to calculate student improvement?

The test looks like KiuFunza tests you may have seen before and will measure the same skills. However, we will use it to calculate the difference in results in a precise and detailed way. Take for example a girl student named Farida who is in Grade 3. If last year her Grade 2 score was 50 and this year her score is 70, she will have improved her skills by 20 points according to our test.

6. How does this competition work?

At the end of the year we will have two test levels for each student in the Mashindano schools. Then we will calculate the improvements that each student has made. Some students will have improved by 20, like Farida. Other students will have improved less, others more. We give a few examples in **Table 2** below.

Table 2: Example of Test 1 and 2 Student Test Scores

Student Name	First test	Second test	Improvement = Second Test - First test
Farida	50	70	20
John	50	52	2
Maria	50	65	15
Mohamed	50	80	30

The important takeaway in this example is that KiuFunza will rank teachers according to the improvement that their students have made between test 1 and test 2. In this example, if these are Grade 3 Math scores and all four students had a different Math teacher, the teacher of Mohamed would score best and therefore earn the largest bonus, while John's teacher will earn the smallest bonus.

7. But I have students at different levels in my class!

That does not matter. We will still calculate the improvement for each student and pay their teachers accordingly. To do this transparently we create student ability groups based on test 1 scores. These ability groups are national groups and are not confined to your school only.

Students in each ability group have similar learning levels as others in that group, across all KiuFunza schools. If your student is in the top group, that is, the group with highest ability of all others, then that student is in that group together with other students of the same highest ability from all KiuFunza Districts and schools.

8. But that is unfair: our students may have lower ability than those in other schools?

No, this is exactly why we create ability groups. The competition will be conducted among students in the **same** ability group. We make sure that, no matter where they go to school, kids of the same ability will compete in one group. We measure their improvement, rank improvements from "most improved" to "least improved" after test 2, and pay according to these ranks. The students who have improved most earn their teachers the highest bonus amounts.

9. But what about students that were not tested last year? And students that drop out of school or are not present the day of test 2?

For students without test 1 scores, we will put these students in an ability group together since we cannot accurately measure their starting ability. Students that drop out or are not present on the day of test 2 will earn you no money.

10. How many ability groups are there?

The number of ability groups in a class and subject are shown in **Table 3**.

Table 3: Number of Ability Groups in a Class and Subject

Class	Subject	Number of Groups
Standard 1	Kiswahili	10
	English	10
	Arithmetic	10
Standard 2	Kiswahili	10
	English	11
	Arithmetic	9
Standard 3	Kiswahili	10
	English	11
	Arithmetic	9

At the end of the year, Twaweza will send volunteers to administer the KiuFunza endline test (test 2) to all students of standards 1, 2 and 3 in Kiswahili, English and Arithmetic. After getting the results, Twaweza will rank the students in each ability group based on their test results. The larger their improvement compared with test 1, the higher their rank and the larger the payment to their teacher.

11. How much money is in this competition?

The KiuFunza teachers will compete to earn a bonus from a budget of **Tshs. 127,475,860** in all the subjects and classes as shown in **Table 4**.

Table 4: Total bonus for the Competition per subject and class.

Classes	Subject	Amount (TShs.)
Standard 1	Kiswahili	15,324,600
	English	15,324,600
	Arithmetic	15,324,600
	Total for Standard 1	**45,973,800**
Standard 2	Kiswahili	14,484,200
	English	14,484,200
	Arithmetic	14,484,200
	Total for Standard 2	**43,452,600**
Standard 3	Kiswahili	12,683,200
	English	12,683,200
	Arithmetic	12,683,200
	Total for Standard 3	**38,049,600**
Total for All Classes		**127,476,000**

As you can see from the table above, the total budget is split among grades depending on the number of students in each grade. The funds per grade are then split evenly among the three subjects, Kiswahili, English and math. Under this system all grade, subject combinations are valued equally. For example, Grade 1 English teachers will earn the same total bonus as Grade 1 Kiswahili teachers.

12. But how much can I expect to earn?

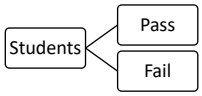
A student who, based on the end of the year test results, ranks at the top of his or her ability group will earn his or her teacher an estimated 3,200 TZS; a student who ranks in the middle will earn his or her teacher an estimated 1,600 TZS, and a student who ranks at the bottom of his or her ability group will earn his or her teacher 0 TZS. If you have a class of 50 students and all of them rank in the middle at test 2 you earn an estimated TZS 80,000. If all 50 students are rank in the top of their ability group you will earn an estimated of TZS 160,000, and if all of your students are at the bottom then you will earn nothing.

13. Why can't you tell me exactly what I will earn?

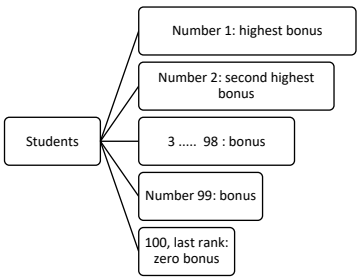
This is because what you will earn depends on your efforts and the efforts of other teachers. If other teachers do more to improve their students' learning, e.g. because they spend more time on instruction, they will achieve a higher rank for their students and thus earn more. But if you do more and your students rank higher, you will earn more.

14. Can you explain again why we should focus on all students?

In previous years of KiuFunza, a teacher was only paid for each of his or her students who passed the KiuFunza tests at the end of the year, successfully completing all of the grade-specific skills as dictated by the curriculum. Teachers who were lucky enough to have very good students simply earned more.



Under this new program, students of all ability levels can earn a bonus payment for his or her teacher. Even if a student is low ability, as long as he or she improves her skills and learns more than other low ability students over the course of the year, a teacher can earn a bonus payment for this student.



15. How can a teacher maximize his or her benefit?

The most successful in this bonus scheme will be teachers who find ways to help students of all ability levels improve their learning as much as possible over the course of the school year. How? Perhaps by teaching longer hours; perhaps by grouping children within a class according to their starting skills; perhaps by using more books; perhaps by asking bright, more advanced students to help weaker students. KiuFunza does not tell teachers how to do this; we trust you know best.

16. What will Head Teachers earn?

As in previous years, Head Teachers will benefit whenever teachers benefit. For every TZS 5,000 that a teacher earns, the Head Teacher receives TZS 1,000. The idea is that Head Teachers play an important role in making sure that teachers can teach effectively. We appreciate this role and invite Head Teachers and teachers to work as a team and therefore benefit as much as possible.

Questions and detailed information

Below we provide answers to some key questions about this programme of paying bonuses. Community meetings will be held at each school to provide additional information about the intervention. You can also contact your District KiuFunza Coordinator, his name and phone number are available at the school on the posters and at the end of this leaflet.

1. Why is learning to read Swahili, English and Arithmetic so important?

These skills are the foundation of education; and without these skills, students cannot succeed in the upper classes. Every student must have these basic skills to be successful in life.

2. Why do you focus on classes 1, 2 and 3?

Learning in these early classes creates a solid foundation for learning in life. By the end of Grade 3, the learner must know how to read, write and count.

3. Why did you choose only a few schools?

Since we are testing a new idea it is best to try it in a few schools in the first few districts. We have experimented for two years 2013 and 2014 and have decided to continue in 2015 by improving our experiment. If the experiment is successful this idea can be implemented by the Government throughout the country.

4. What does this bonus mean?

The teachers of classes 1, 2 and 3 will be paid a bonus after their pupils have done a test that will be given at the end of the year and prove that they can read Kiswahili, read English and do maths. This bonus money is extra after their salaries are paid. Our expectation is that teachers will increase accountability and make students learn these skills.

5. What kind of format will the exams have?

The KiuFunza tests are based on the curriculum standards set by government for Grades 1, 2 and 3. There will be three exams - reading Swahili, reading English and doing Arithmetic.

6. How do you come up with the tests?

The tests are based entirely on the national curriculum for each subject and each grade. We have a panel of experts, including people from the Curriculum Institute to come up with questions at the right level.

7. What are the exams like?

These exams are quite different than normal school exams. Twaweza researchers will come to your school at the end of the year. Each child will be tested individually face-to-face with the researcher. The test is verbal so answers are spoken by the child but there will be paper available for them to write on if this helps them. We work hard to make sure that the child is not frightened and that the atmosphere is generally helpful to them.

8. How will you calculate the payments?

The bonus will be calculated based on how well your pupils have been able to pass compared to their peers in their ability group in the past year exams. Students who learn the most, i.e. are at the top of his or her group at the end of the year, will earn their teacher the most money. Students who learn the least, i.e. are at the bottom of his or her ability group at the end of the year, will earn their teacher no bonus.

9. What method will you use to divide the students into ability groups?

We shall use last year's data, we have the results for all students in classes 2 and 3 so we can categorize them from those results. For the 2nd and 3rd grade students that we did not test last year, since we do not have their results, we will put them in one of their own ability group. For new Grade 1 students, we will use different indicators of school ability in general to use last year's school data to categorize them. Ability groups will focus on the skills of each subject in each class. For example, in 2013 and 2014, the skills in each subject were as follows:

Standard I Kiswahili	Reading syllables, reading words, reading sentences
Standard I English	recognising letters, reading words, reading sentences
Standard I Arithmetic	counting, recognising numbers, which number is bigger, addition, subtraction
Standard 2 Kiswahili	Reading words, reading sentences, reading paragraphs
Standard 2 English	Reading words, reading sentences, reading paragraphs
Standard 2 Arithmetic	which number is bigger, addition, subtraction, multiplication
Standard 3 Kiswahili	Reading a story and answering comprehension questions
Standard 3 English	reading a story, answering comprehension questions
Standard 3 Arithmetic	addition, subtraction, multiplication, division

10. How will the teacher know which groups my students are at the beginning of the year?

A list of each of your students and the ability group will be presented to you at the beginning of the year by KiuFunza Volunteers.

11. How will you know the test 1 result for a new student that has joined the class?

For Grade 2 and 3 students for whom we do not have test 1 scores, we will put these students in an ability group together since we cannot accurately measure their ability.

12. How will you know the test 1 result for students in Grade 1?

Since Grade 1 students have just started school, we do not have a test 1 score for them. Instead, we use a reliable measure to predict their ability based on things we know about the school and other students at the school.

13. Is it best to increase the effort to increase reading Kiswahili, English, and maths skills for low-performing students or to extend these skills among high-performing students?

In this contest, every student is important. Because these competitions are held within the competence groups, it is quite possible that a low-performing student, call him Constance Msisiri, will give you the highest possible bonus award, that a high performing student called Farida Hassani. If Constance Msisiri,, scores more marks that the rest of his ability group taking the first place, while Farida Hassani was the last one in the ability group, then Constance Msisiri, will earn you more bonus shillings than Farida Hassani.

14. What will you do to give students marks after the test?

In each capacity group, students will have different results. Keep in mind that although they started out with the same abilities, we hope that the teacher's efforts will allow them to increase ability, but to different levels as well. Therefore, after the KiuFunza examination, we will see the changes in each group. Thus, students will be challenged and awarded based on those exam results. But since the exams are short, more than one student will get the same result, rank or position in the group.

15. How many shillings should I expect to be paid?
Because we will pay bonuses through a competition between all teachers, we will never know your personal victory against all the other teachers in our experiment. But, on average, and using last year's results as a predictive measure, on average, a teacher with an average class of 60 to 70 students who will pass on average within their ability groups can expect to be paid TZS 104,000. In order to make sure you win over other teachers is important to teach and make sure your students pass more skills that other students in their groups, by ensuring as much as possible that your students pass more skills in reading Swahili, English, and to do the maths.

16. Why don't you tell me how much I will get?
Although we are aware of the amount allocated to pay the bonuses, the amount a teacher will receive depends on the level of learning that his or her students will show in the test. So, we can tell you about the average share but some teachers will earn less than others depending on how the students have increased their abilities. Thus, we can know the actual amounts after the test results.

17. Will the Headmaster of the school be given any bonus?
The head teacher will receive 20% or 1/5 of the total received by all the teachers in their school. Since the head teacher has an important role in overseeing the work of teachers, we want to reward them when teachers do well. At the same it is the teacher in the classroom who truly makes sure that children learn so it is fair that they get the most money. However, head teachers are also awarded for all teachers in their school so in the end may get a large bonus also.

18. Will teachers of difficult subjects, like English, earn a bigger bonus than teachers of other, easier subjects?
No. In each grade, there is an equal amount per subject (Table 4 contains full information).

19. Will teachers be trained to prepare for these exams?
No. Specialized training is not necessary since the test is about the subjects they teach in the curriculum. What teachers need is to use their skills and experience to teach students to read Kiswahili, and English and to do maths.

20. What should a teacher with problems in teaching do?
The teacher may ask for help from the Head Teacher, fellow teachers or retired teachers near the school. She can also ask for help at the nearest Teacher's Center. Twaweza believes that if a teacher decides to get good results he will find ways to achieve that goal. The key is for the teacher to be committed and to be accountable appropriately.

21. Will Twaweza help pay for in-service teacher training?
No. In this program there will be no payments for training or other expenses. The bonus payment is done after the students' results become known at the end of the year.

22. What should the teacher do if the students do not learn despite their best efforts?
This bonus is for those that have passed, not for the amount of effort. KiuFunza believes that if a teacher helps students, uses good teaching techniques, many students will learn the skills they need and pass the test.

23. What if the teacher is transferred to another school or classroom?

At the beginning of the year, KiuFunza collects the names of all teachers who teach classes 1, 2 and 3. We have contacted the District Education Officer (DEO) and requested her/him not to make the transfer if not necessary. However, when the transfer of a teacher occurs, KiuFunza will pay the teacher whose information was collected at the beginning of the year. It is the responsibility of the head teacher to report to the District Coordinator on the transfer as soon as possible. It is up to the new teacher to agree with the old teacher, ie the teacher whom KiuFunza recognized, about sharing the bonus between them. This agreement is for the benefit of both teachers, both old and new, and the Head teacher also ensures that many students succeed even if there is a teacher transfer.

24. Some schools employ volunteer teachers who are not paid by the government. Will teachers like these be paid by KiuFunza?
This bonus payments are meant for the teacher for each grade, subject whose names we collected at the beginning of the year. If a volunteer teacher assists the main teacher to teach the grade/subject, this volunteer teacher should agree with the main teacher about the possibility of sharing the bonus payment, but should not be registered directly with Twaweza to receive payment. If this main teacher is a volunteer, this teacher may be registered with Twaweza to receive the bonus payment.

25. Wouldn't it be better to buy more books or build a better school than pay teachers?
Books and classrooms are important and are provided by the Government. But Twaweza believes that a dedicated teacher has a unique role to play in learning and that is why we try to pay this bonus to see if it will make him more diligent and make the students learn.

26. In what way will teachers be paid?
Teachers will be paid after the final exam results of 2015. Teachers have the freedom to choose to be paid via mobile money or bank accounts.

27. What information does the KiuFunza need to enable it to pay this bonus?
At the beginning of the year, KiuFunza needed to know:

- the names of teachers who teach Kiswahili, English and Arithmetic in each of the standards I, II and III;
- The payment method chosen by the teacher;
- the names of all students in classes 1,2 and 3; and also the photograph of students.

In some schools, this information will be partly collected by Economic Development Initiatives (EDI), a research organization working on behalf of Twaweza. They are assisting in collecting this data this year to decrease the amount of interview time needed with teachers.

We urge all teachers to make sure their information is complete and correct. All teachers should list their personal bank or personal phone account information and not that of other people such as their spouses, etc. Teachers who will list either non-personal information will delay their payments and those of their peers.

28. Will Twaweza be offering any other funding or program to address input shortages or other problems at the school?
No. Unlike parts of the KiuFunza program in past years, this year we are only implementing teacher bonus programs.

29. Why is KiuFunza conducting this experiment?

We want to see if it is possible to improve education and learning by paying teachers. There are numerous studies in the world that have shown that such a system can help to improve the quality of education. So, we want to know if this can be successful in Tanzania.

30. Why did you choose this school?
We did not choose this school in any way. All schools in the district had an equal chance of being selected to participate in KiuFunza, we did the lottery and this school got lucky to be selected.

31. How long will this experiment be?
The initial phase of KiuFunza took place in 2013 and 2014. This trial will continue for two years as well: 2015 and 2016.

32. What will happen when the plan ends?
At the end of the program, KiuFunza will no longer offer bonuses to teachers and Head Teachers. scholarships to Senior Teachers and Teachers. But if the learning outcomes are positive, the Government will see the possibility to expand the program across the country.

33. What will happen between now and the time of the end of the year tests?
Researchers teams will come to the school to see what's going on. The researchers will be accompanied by letters of identification from Twaweza. The work of these researchers is crucial in carrying out this effort to empower teachers. Please give them your cooperation. We hope that between now and the end of the year, teachers and students will be diligent in their responsibility.

34. Who should I ask if I have more questions?
Ask the District Coordinator who will have to leave his/her phone number with the Head Teacher and you can ask him / her questions. The number of the District Coordinator will also be posted on the Posters that will be posted at the school and also at the end of this leaflet. However, if the District Coordinator will not be able to provide answers to policy questions, he or she will contact the KiuFunza Coordinator, at Twaweza so we can answer your questions.

35. What is Twaweza?
Twaweza is a non-governmental organization that deals with research and mobilizes the public to take action. Twaweza believes that citizens themselves have the power to take action, demand their rights and, ultimately, bring change. Twaweza also believes that cooperation between the Government and the people will bring prosperity to the country. Having scientific evidence is an important tool in making that change. Twaweza's research will help the Government develop better policies.
For more information see the website: www.twaweza.org or contact the District KiuFunza Coordinator through the information listed at the end of this leaflet.

F.2 Levels

KiuFunza 2015

Improving Learning in Tanzania

Information Paying Teachers for Skills learned.

Twaweza, a citizen centered initiative, in cooperation with the Commission for Science and Technology and the Government of Tanzania is working to find ways to improve learning in schools. In order to achieve this, scientific trials known as KiuFunza are being implemented to test various education policies.

KiuFunza was inaugurated by the Minister of State, The Prime Minister's Office, Regional Administration and Local Government (PMO-RALG) Hon. Hawa Ghasia (Mp) in January 2013.

In 2015 these scientific trials will be implemented in 21 Local Government Authorities (LGAs). As part of the trials a new program called Stadi is implemented in 66 schools in these LGAs.

Table 1 provides information on all participating LGAs. If these trials are successful the government will consider to scale up to cover more LGAs.

Table 1: Number of Stadi schools in KiuFunza LGAs

Names of Local Government Authorities	Number of Schools in the Program
Geita and Nyang'hwale	6
Kahama, Ushetu, na Msalala	6
Karagwe and Kyerwa	6
Kigoma and Uvinza	5
Kinondoni	7
Kondoa and Chemba	6
Korogwe	6
Lushoto and Bumbuli	6
Mbozi and Momba	6
Sumbawanga and Kalambo	6
Mbinga and Nyasa	6
Total	**66**

The selection of Local Government Authorities and schools was done through a lottery. Your school was chosen in this lottery to be part of the Stadi program. This flyer explains what this program is using questions and answers.

1. Last year my school was already part of a Twaweza program: is this the same program?

No, but some features of the program are the same. It has the same goal: to improve reading Kiswahili, English and doing Arithmetic in standards 1, 2 and 3. The program will use a bonus payment for teachers and head teachers in these subjects. The bonus is paid on top of the normal salary. You have been invited to participate because you teach one or more of these topics in these grades.

2. What do I need to do now to participate?

Today you just need to take time to understand the Stadi program and confirm your participation. Today is an important training day where Twaweza representatives explain how you as a beneficiary of the program can maximize your benefit.

3. How is the bonus program different this year?

Previously we offered bonuses on the basis of passing a complete test. As you know, these tests had different skill levels. For example we tested reading syllables, reading words, and reading sentences. But there were only two possible outcomes: the student passed or the student failed. In the new Stadi bonus program, teachers will receive a bonus for each skill level that a student passes. Teachers whose students master more skills will earn bigger bonuses.

4. What do you mean by skill level?

The national syllabus for each subject, Kiswahili, English and math, covers many skill levels. An example of these skills is below:

Grade 1 Kiswahili	reading syllables, reading words, reading sentences
Grade 1 English	reading letters, reading words, reading sentences
Grade 1 Maths	counting, identifying numbers, inequalities, addition, subtraction
Grade 2 Kiswahili	Reading words, reading sentences, reading paragraphs
Grade 2 English	Reading words, reading sentences, reading paragraphs
Grade 2 Maths	Inequalities, addition, subtraction, multiplication
Grade 3 Kiswahili	reading stories, answering comprehension questions
Grade 3 English	reading stories, answering comprehension questions

Grade 3 Maths	Addition, subtraction, multiplication, division

5. What do you mean by a bonus for each skill level?

KiuFunza Stadi will use a test to measure skills in each Grade for Kiswahili, English and Math. For example, in the Grade 1 Kiswahili test students are asked to do three skill levels: to read syllables, to read words, and to read sentences. In the Stadi program a teacher earns a bonus for each skill level a student successfully completes. The more levels a student can do, the more the teacher earns. For Grade 1 Kiswahili, if a student can read syllables, the teacher will earn. If the student can also read words, the teacher will earn more. If the student can also read sentences, the teacher will earn the most.

6. What about other topics and grades?

For all KiuFunza topics, Kiswahili, English and Math, the same principle is used. The more levels a student can do, the more the teacher will earn. For example in Grade 2 Math, skill levels are knowing which number is largest; adding numbers; subtracting numbers; and multiplying numbers. Again, if a student passes all these skill levels the teacher earns the most. But even if a student can only pass one level the teacher will earn.

7. But is this fair? Some topics and skills are really hard for students!

That is right. Certain subjects and skills are especially difficult for students, especially English. Despite their best efforts, teachers in English have a hard time getting good bonuses. But also some Math skills such as division are hard to learn. Therefore we will reward teachers more if a skill is harder to learn for a student.

8. Can you give me some explanation about the total bonus funds?

Twaweza has a total of TZS 142,524,140 to give out among teachers. We divide the money according to the number of students in each grade and evenly across subjects and skills.

9. How are funds divided over skills for Grade 1?

Grade 1		
Subject	Skill	Amount
Kiswahili	Sylabi	TZS 5,332,259
	Maneno	TZS 5,332,259
	Sentensi	TZS 5,332,259
English	Letters	TZS 5,332,259
	Words	TZS 5,332,259
	Sentences	TZS 5,332,259
Maths	Counting	TZS 3,199,355
	Identifying Numbers	TZS 3,199,355
	Inequalities	TZS 3,199,355
	Addition	TZS 3,199,355
	Subtraction	TZS 3,199,355

10. How are funds divided over skills for Grade 2?

Grade 2		
Subject	Skill	Amount
Kiswahili	Maneno	TZS 5,332,259
	Sentensi	TZS 5,332,259
	Aya	TZS 5,332,259
English	Words	TZS 5,332,259
	Sentences	TZS 5,332,259
	Paragraph	TZS 5,332,259
Maths	Inequalities	TZS 3,999,195
	Addition	TZS 3,999,195
	Subtraction	TZS 3,999,195
	Multiplication	TZS 3,999,195

11. How are funds divided over skills for Grade 3?

Grade 3		
Subject	Skill	Amount
Kiswahili	Hadithi	TZS 7,778,011
	Maswali ya Ufahamu	TZS 7,778,011
English	Story	TZS 7,778,011
	Comprehension Questions	TZS 7,778,011
Math	Addition	TZS 3,889,005
	Subtraction	TZS 3,889,005
	Multiplication	TZS 3,889,005

	Division	TZS 3,889,005

You see that math has more skills so we reward the passing each math skill differently than the languages. But the total reward for math is the same as the total reward for other subjects.

12. Can you tell me what I will earn exacly?

No, because how much you will earn this year depends on the total number of students that pass at the end of the year. The amount of money we pay per student that passes is equal to the total amount of bonus money available divided by the number of students that pass. But we have some idea about the expected amounts.

13. But what can a teacher expect to earn?

Suppose you are a Grade 1 Kiswahili teacher and you have 50 students in your class. If you help all of them do letters, you can expect to earn approximately TZS 75,850. If you help all of them do letters and words you can expect to earn approximately TZS 161,900. If you help all of them master all three skills, then you can expect to earn approximately TZS 261,800.

14. How is this different for English teachers?

English is often the hardest subject for students in Grades 1-3. Since fewer students master the skills in English, we divide the total bonus among fewer teachers. But since fewer students pass, English teachers will be able to get bonuses similar to their colleagues in other subjects as long as their students master some skills.

15. How can a teacher maximize his or her benefit?

The most successful in this bonus scheme will be teachers who find ways to help many students master many skills outlined in the syllabus. How? Perhaps by teaching longer hours; perhaps by grouping children within a class according to their starting skills; perhaps by using more books; perhaps by asking bright, more advanced students to help weakerstudents. KiuFunza does not tell teachers how to do this; we trust you know best.

16. What will Head Teachers earn?

As in previous years, Head Teachers will benefit whenever teachers benefit. For every TZS 5,000 that a teacher earns, the Head Teacher receives TZS 1,000. The idea is that Head Teachers play an important role in making sure that teachers can

teach effectively. We appreciate this role and invite Head Teachers and teachers to work as a team and therefore benefit as much as possible.

Questions and detailed information
Below we provide answers to some key questions about this programme of paying bonuses. Community meetings will be held at each school to provide additional information about the intervention. You can also contact your District KiuFunza Coordinator, his name and phone number are available at the school on the posters and at the end of this leaflet.

1. Why is learning to read Swahili, English and Arithmetic so important?
These skills are the foundation of education; and without these skills, students cannot succeed in the upper classes. Every student must have these basic skills to be successful in life.

2. Why do you focus on classes 1, 2 and 3?
Learning in these early classes creates a solid foundation for learning in life. By the end of Grade 3, the learner must know how to read, write and count.

3. Why did you choose only a few schools?
Since we are testing a new idea it is best to try it in a few schools in the first few districts. We have experimented for two years 2013 and 2014 and have decided to continue in 2015 by improving our experiment. If the experiment is successful this idea can be implemented by the Government throughout the country.

4. What does this bonus mean?
The teachers of classes 1, 2 and 3 will be paid a bonus after their pupils have done a test that will be given at the end of the year and prove that they can read Kiswahili, read English and do maths. This bonus money is extra after their salaries are paid. Our expectation is that teachers will increase accountability and make students learn these skills.

5. What kind of format will the exams have?
The KiuFunza tests are based on the curriculum standards set by government for Grades 1, 2 and 3. There will be three exams - reading Swahili, reading English and doing Arithmetic.

6. How do you come up with the tests?
The tests are based entirely on the national curriculum for each subject and each grade. We have a panel of experts, including people from the Curriculum Institute to come up with questions at the right level.

7. What are the exams like?
These exams are quite different than normal school exams. Twaweza researchers will come to your school at the end of the year. Each child will be tested individually face-to-face with the researcher. The test is verbal so answers are

spoken by the child but there will be paper available for them to write on if this helps them. We work hard to make sure that the child is not frightened and that the atmosphere is generally helpful to them.

8. How is my payment calculated?
The payment is calculated based on whether your student has achieved a specific skill in Kiswahili, English or Maths that is appropriate to their grade level. The student will be marked for each skill they are able to successfully complete. The actual amount you receive for this achievement is dependent on how many other students in KiuFunza across the country successfully complete each skill.

9. What types of skills will my students need to have?
The types of skills we are talking about are different for each subject but some examples are below. These are all based on the national curriculum for the different grades and subjects.

Grade 1 Kiswahili	reading syllables, reading words, reading sentences
Grade 1 English	reading letters, reading words, reading sentences
Grade 1 Maths	counting, identifying numbers, inequalities, addition, subtraction
Grade 2 Kiswahili	Reading words, reading sentences, reading paragraphs
Grade 2 English	Reading words, reading sentences, reading paragraphs
Grade 2 Maths	Inequalities, addition, subtraction, multiplication
Grade 3 Kiswahili	reading stories, answering comprehension questions
Grade 3 English	reading stories, answering comprehension questions
Grade 3 Maths	Addition, subtraction, multiplication, division

10. How does all of this translate into money?
Once the children have taken the tests administered by KiuFunza, we can calculate the payout for each teacher based on the total number of students who have achieved that skill. Here's a hypothetical example. Let's do a simple example. Say there are five students doing Grade 1 Kiswahili across the country. They all achieve different skills levels; there are three possible skill levels. And say the total bonus to be paid out for Grade 1 is TZS 90,000. After the tests we see that the five students are able to successfully read words, 3 can successfully read sentences, and only 2 can successfully read a paragraph. The payment for each skill a student successfully completes will be calculated as follows:

Skill	Number of students who achieved this skill	Total pot for skill (TZS 90,000/3)	Bonus payment to each teacher (per student)	Bonus pay to Head Teacher (per student)
Kusoma maneno	5	TZS 30,000	TZS 5,000	TZS 1,000

Kusoma sentensi	3	TZS 30,000	TZS 8,000	TZS 2,000
Kusoma aya	2	TZS 30,000	TZS 12,000	TZS 3,000

**These numbers are not representative of actual KiuFunza bonus amounts. They are simply meant to serve as an example to help teachers understand how the size of their bonus will be calculated.

11. Why can't you tell me how much money I will get?
As you can see in the example above, how much money you get will depend on how many of your students achieve a certain skill as well as how many other students in the same grade achieve that skill. We will only know this information after the tests at the end of the year.

12. You mentioned earlier that you will reward teachers more if a skill is harder to learn for a student. How do you know which skills are hard?

We calculcate the total number of students that are able to complete the skill. The fewer students that are able to complete the skill the harder it must be.

13. How do you determine how much I earn for harder skills vs easier skills?

To address this we have fixed the size of the total bonus funds available for teachers for each grade and subject. The total bonus funds are proportional to the number of enrolled students in each grade.

We have then divided the budget equally among the skills within each subject. We then divide this money by the total number of students in all Tanzania that are able to complete each skill and give the teachers of these students the money. Therefore, if fewer KiuFunza students nation-wide are able to complete the skills in a subject, it means that a teacher will earn more money for each student they are able to successfully teach skills in that subject.

14. Will the Headmaster of the school be given any bonus?
The head teacher will receive 20% or 1/5 of the total received by all the teachers in their school. Since the head teacher has an important role in overseeing the work of teachers, we want to reward them when teachers do well. At the same it is the teacher in the classroom who truly makes sure that children learn so it is fair that they get the most money. However, head teachers are also awarded for all teachers in their school so in the end may get a large bonus also.

15. Why should not the test be given at the beginning of the year and at the end to see real progress?

Few students will be given the exam at the beginning of the year, but it is difficult to do so for all. However, from the research, we have evidence that many students in classes 1, 2 and 3 have no reading, writing and numeracy skills.

16. Will teachers be trained to prepare for these exams?
No. Specialized training is not necessary since the test is about the subjects they teach in the curriculum. What teachers need is to use their skills and experience to teach students to read Kiswahili, and English and to do maths.

17. What should a teacher with problems in teaching do?
The teacher may ask for help from the Head Teacher, fellow teachers or retired teachers near the school. She can also ask for help at the nearest Teacher's Center. Twaweza believes that if a teacher decides to get good results he will find ways to achieve that goal. The key is for the teacher to be committed and to be accountable appropriately.

18. Will Twaweza help pay for in-service teacher training?
No. In this program there will be no payments for training or other expenses. The bonus payment is done after the students' results become known at the end of the year.

19. What should the teacher do if the students do not learn despite their best efforts?
This bonus is for those that have passed, not for the amount of effort. KiuFunza believes that if a teacher helps students, uses good teaching techniques, many students will learn the skills they need and pass the test.

20. What if the teacher is transferred to another school or classroom?
At the beginning of the year, KiuFunza collects the names of all teachers who teach classes 1, 2 and 3. We have contacted the District Education Officer (DEO) and requested her/him not to make the transfer if not necessary. However, when the transfer of a teacher occurs, KiuFunza will pay the teacher whose information was collected at the beginning of the year. It is the responsibility of the head teacher to report to the District Coordinator on the transfer as soon as possible. It is up to the new teacher to agree with the old teacher, ie the teacher whom KiuFunza recognized, about sharing the bonus between them. This agreement is for the benefit of both teachers, both old and new, and the Head teacher also ensures that many students succeed even if there is a teacher transfer.

21. Some schools employ volunteer teachers who are not paid by the government. Will teachers like these be paid by KiuFunza?
This bonus payments are meant for the teacher for each grade, subject whose names we collected at the beginning of the year. If a volunteer teacher assists the

main teacher to teach the grade/subject, this volunteer teacher should agree with the main teacher about the possibility of sharing the bonus payment, but should not be registered directly with Twaweza to receive payment. If this main teacher is a volunteer, this teacher may be registered with Twaweza to receive the bonus payment.

22. Wouldn't it be better to buy more books or build a better school than pay teachers?
Books and classrooms are important and are provided by the Government. But Twaweza believes that a dedicated teacher has a unique role to play in learning and that is why we try to pay this bonus to see if it will make him more diligent and make the students learn.

23. In what way will teachers be paid?
Teachers will be paid after the final exam results of 2015. Teachers have the freedom to choose to be paid via mobile money or bank accounts.

24. What information does the KiuFunza need to enable it to pay this bonus?
At the beginning of the year, KiuFunza needed to know:

- the names of teachers who teach Kiswahili, English and Arithmetic in each of the standards I, II and III;
- The payment method chosen by the teacher;
- the names of all students in classes 1,2 and 3; and also the photograph of students.

In some schools, this information will be partly collected by Economic Development Initiatives (EDI), a research organization working on behalf of Twaweza. They are assisting in collecting this data this year to decrease the amount of interview time needed with teachers.

We urge all teachers to make sure their information is complete and correct. All teachers should list their personal bank or personal phone account information and not that of other people such as their spouses, etc. Teachers who will list either non-personal information will delay their payments and those of their peers.

25. Will Twaweza be offering any other funding or program to address input shortages or other problems at the school?
No. Unlike parts of the KiuFunza program in past years, this year we are only implementing teacher bonus programs.

26. Why is KiuFunza conducting this experiment?
We want to see if it is possible to improve education and learning by paying teachers. There are numerous studies in the world that have shown that such a

system can help to improve the quality of education. So, we want to know if this can be successful in Tanzania.

27. Why did you choose this school?
We did not choose this school in any way. All schools in the district had an equal chance of being selected to participate in KiuFunza, we did the lottery and this school got lucky to be selected.

28. How long will this experiment be?
The initial phase of KiuFunza took place in 2013 and 2014. This trial will continue for two years as well: 2015 and 2016.

29. What will happen when the plan ends?
At the end of the program, KiuFunza will no longer offer bonuses to teachers and Head Teachers. scholarships to Senior Teachers and Teachers. But if the learning outcomes are positive, the Government will see the possibility to expand the program across the country.

30. What will happen between now and the time of the end of the year tests?
Researchers teams will come to the school to see what's going on. The researchers will be accompanied by letters of identification from Twaweza. The work of these researchers is crucial in carrying out this effort to empower teachers. Please give them your cooperation. We hope that between now and the end of the year, teachers and students will be diligent in their responsibility.

31. Who should I ask if I have more questions?
Ask the District Coordinator who will have to leave his/her phone number with the Head Teacher and you can ask him / her questions. The number of the District Coordinator will also be posted on the Posters that will be posted at the school and also at the end of this leaflet. However, if the District Coordinator will not be able to provide answers to policy questions, he or she will contact the KiuFunza Coordinator, at Twaweza so we can answer your questions.

32. What is Twaweza?
Twaweza is a non-governmental organization that deals with research and mobilizes the public to take action. Twaweza believes that citizens themselves have the power to take action, demand their rights and, ultimately, bring change. Twaweza also believes that cooperation between the Government and the people will bring prosperity to the country. Having scientific evidence is an important tool in making that change. Twaweza's research will help the Government develop better policies.
For more information see the website: www.twaweza.org or contact the District KiuFunza Coordinator through the information listed at the end of this leaflet.