Designing Teacher Performance Pay Programs: Experimental Evidence from Tanzania[*]

Isaac Mbiti[†] Karthik Muralidharan[‡] Mauricio Romero[§] Youdi Schipper[¶]

September 27, 2017

Abstract

A growing body of evidence suggests that well-designed teacher performance pay systems can improve student learning outcomes. However, education systems in developing countries face a trade-off between more complex theoretically effective systems which are harder to implement, and simpler systems which are easier to administer but theoretically less effective. Given the limited research that directly compares different teacher performance pay systems, we use a field experiment to compare the effectiveness of the pay for percentile system, which can induce optimal effort among teachers, with a simple system that rewards teachers based on student proficiency levels. Despite the theoretical advantages of the pay for percentile system, we find the proficiency system is at least as effective in raising student learning.

1 Introduction

Developing countries invest considerable resources in the education sector. Tanzania spends about 3.5 percent of GDP on education, which is slightly below the sub-saharan African average of 4.5 percent (World Bank, 2013). Teacher compensation, including wages and other pecuniary benefits, typically accounts for the majority of the education budget. For instance, teacher compensation accounts for almost two-thirds of the Tanzanian budget, compared to 80 percent in Uganda (UNICEF, 2016, and World Bank EDStats, 2016). Additionally, the average teacher

[*] We are grateful to Joseph Mmbando and seminar participants at UC San Diego for comments. Erin Litzow and Jessica Mahoney provided excellent research assistance through Innovations for Poverty Action. We are also grateful to EDI Tanzania for their excellent data collection and implementation efforts. The EDI team included Respichius Mitti, Phil Itanisia, Timo Kyessey, Julius Josephat, Nate Sivewright and Celine Guimas. The evaluation was partially funded by a grant from the REACH Trust fund at the World Bank.

[†] University of Virginia, J-PAL, IZA; imbiti@virginia.edu

[‡] University of California - San Diego, NBER, J-PAL; kamurali@ucsd.edu

[§] University of California - San Diego; mtromero@ucsd.edu

[¶] Twaweza; yschipper@twaweza.org

in a sub-Saharan African country earns almost four times GDP per capita, compared to OECD teachers who earn 1.3 times GDP per capita (OECD, 2017; World Bank, 2017). Despite the relatively lucrative wages of teachers, the government has faced sustained pressure to increase teacher pay, including a 2012 strike where the teachers union demanded a 100 percent increase in pay (Reuters, 2012; PRI, 2013).

Given the low levels of student learning and high rates of teacher absenteeism in Tanzania (Uwezo, 2014), which serves as proxy measure of teacher motivation, proponents argue that increasing teacher pay will increase teacher motivation, which will in turn increase student learning outcomes. However, a large body of evidence in both developing and developed countries has consistently shown that the correlation between teacher compensation and student learning is extremely low (Kane, Rockoff, & Staiger, 2008; Bettinger & Long, 2010; Woessmann, 2011). In contrast, teacher performance pay programs could potentially be more effective as they link teacher pay to student learning. However, there is mixed evidence on the efficacy of these programs. A number of studies find that teacher performance pay leads to increases in student learning (Glewwe, Ilias, & Kremer, 2010; Lavy, 2002, 2009; Muralidharan & Sundararaman, 2011; Balch & Springer, 2015), while another set of studies find that these programs have limited impact on student learning (Goldhaber & Walch, 2012; Goodman & Turner, 2013; Springer et al., 2011). Although direct comparisons between these studies is difficult due to the differences in context, design, and budgets, there is growing consensus that heterogeneity in program effectiveness is driven in large part by the different incentive designs used (Neal, 2011; Loyalka, Sylvia, Liu, Chu, & Shi, 2016; Ganimian & Murnane, 2016). In particular, incentives designs that reward teachers on the basis of student learning gains are typically more effective than systems that reward teachers if their students attain a specific proficiency level. In addition, incentives programs seem to work best in conjunction with other complementary programs such as resources or student incentives (Behrman, Parker, Todd, & Wolpin, 2015; Mbiti et al., 2017). Moreover, there may be trade-offs between (theoretically) more effective incentive schemes and simpler schemes such as proficiency targets. More effective incentive designs are more complex and harder to implement, which may hinder countries with limited administrative capacity, such as Tanzania, from adopting them. In contrast, simpler schemes such as incentives that use a single proficiency target are easily understood and implemented. However, they typically do not benefit students who are far from the passing threshold, and rarely induce optimal effort from teachers (Neal & Schanzenbach, 2010; Neal, 2011). In addition, using a single uniform proficiency target across all schools may penalize teachers who serve students from disadvantaged backgrounds.

In this paper, we conduct a randomized experiment in a set of 180 Tanzanian schools, where we compare the effectiveness of two different teacher incentive programs, implemented in the same context and with the same budget. The program focused on English, Swahili, and Math in Grades 1, 2, and 3. We compare the effectiveness of the pay for percentile scheme proposed by Barlevy and Neal (2012) to a simple proficiency threshold design. The pay for percentile scheme can induce optimal effort among teachers (Barlevy & Neal, 2012) but is difficult to implement and communicate to teachers. As proficiency designs typically lead teachers to focus on students close to the passing threshold, we include several passing thresholds to allow teachers to earn bonuses for helping a broader set of students.

In both incentive designs, we fix the size of the total bonus pool for each subject-grade combination and determine teacher reward payments based on actual student performance. While we ensure budget comparability across our treatments, the ex-post determination of the actual payments introduces some uncertainty which may reduce teacher responsiveness.

In schools assigned to receive incentives based on proficiency targets, teachers earn bonuses based on their students mastery of several grade-specific skills that are outlined in the national curriculum. These skill thresholds range from very basic skills to more complex skills, which allow teachers to earn rewards across the entire distribution of students. As reward payments for each skill are inversely proportional to the number of students that attain the skill, harder-to-pass skills are rewarded more.

In pay for percentile schools, students are first tested and grouped based on their initial levels of learning at the beginning of the school year. At the end of the school year, students are tested and ranked within their assigned group, and teachers are paid proportionally to their students' ranks within each group. By effectively handicapping the differences in initial student performance across teachers, the pay for percentile system does not penalize teachers who serve disadvantaged students. In addition, since teachers can earn similar rewards for exceptional performance from students in each group, it does not encourage teachers to focus on marginal students. As the system essentially employs a modified tournament design, Barlevy and Neal (2012) show that pay for percentile can induce socially optimal levels of effort among teachers.

Despite the theoretical advantages of the pay for percentile system, the proficiency incentive system is at least as effective as the pay for percentile system. This is in contrast with the findings of Loyalka et al. (2016), who find that student Math test scores increased the most under a pay for percentile system compared to other systems. In the second year of our evaluation, test scores in math increase by about 0.07SD under both systems, while Swahili scores increase by 0.11SD under the proficiency system compared to only a 0.06SD increase (but insignificant) under the pay for percentile. As teacher comprehension of the incentive systems was similar, these differences were not merely driven by the relative lack of understanding among pay for percentile teachers. Teacher behavior was different across the two different incentives systems, where pay for percentile classrooms were more likely to be off task, and assigned less homework. Teachers in pay for percentile classrooms were more likely to report that students were disengaged, and their classrooms were in worse condition. Overall, teachers exerted more effort under the proficiency incentive, relative to the pay for percentile system. In addition, teachers under the pay for percentile tend to focus on their best students.

Given the administrative capacity of developing countries such as Tanzania, simpler proficiency incentive systems may be more suitable for large scale implementation. Results from a previous study in Tanzania found that a system using a single uniform proficiency threshold was not effective in raising test-scores as it induced teachers to focus on marginal students.

By introducing multiple proficiency thresholds, our design mitigates this issue by providing teachers with bonus opportunities across the entire distribution of students. Given the recent push to introduce stronger accountability in the education system through initiatives such as the "Big Results Now," which featured a school based incentive system, evidence that provides policy makers with guidance on how to best structure incentive designs is especially important.

2 Experimental Design

2.1 Context

Overall student learning levels remain extremely low across East Africa despite a decade plus of major reforms and significant new investments in public education. In Kenya, Tanzania and Uganda, recent nationwide surveys show that large majorities of children are unable to read or do arithmetic at the required level (Uwezo, 2012). A foundation in basic literacy and numeracy is commonly considered an important foundation upon which to build new skills. Without this, children may be denied the opportunity to develop fully in the future. Findings from Tanzania indicate that students entering secondary institutions are generally woefully ill-prepared, unable to read in the English language, the medium of instruction for secondary education (Uwezo, 2012). While these challenges are well known, existing reforms and aid instruments have largely failed to improve the situation (Uwezo, 2012).

Under current arrangements no one is held accountable or incentivized to achieve learning. Administrators and teachers are paid regardless of their attendance or performance; quality assurance systems such as the inspectorate function poorly; and appointments of education administrators do not take learning outcomes into account. Still, teachers are generally seen as key potential drivers of learning in schools. Their salaries are typically paid on time and make up about 60% of all spending in primary education. However, teachers in primary education, while at school, are often not in class teaching their students. Surveys estimate instructional time losses of some 60 percent (see Wolrd Bank (2011)). The lack of adequate attention to accountability and incentives may in part explain why increased budgets for education have not resulted in improved learning outcomes. So while government programs have largely focused on providing educational inputs, recent evidence suggests that it may be more effective to incentivize the delivery of learning outcomes, particularly at the local level (Glewwe & Kremer, 2006; Kremer & Holla, 2009).

2.2 Interventions and Implementation

The interventions examined in this paper were formulated and managed by Twaweza, an East-African civil society organization that focuses on citizen agency and public service delivery. The intervention was part of a series of projects launched under a broader program umbrella named KiuFunza ("Thirst for learning" in Swahili). The first set of interventions were launched in 2013 and evaluated by Mbiti et al. (2017). This first project compromised on the 'ideal' design of the incentive program and instead chose a design that was more 'implementable' at scale. More optimal designs, such as those based on teacher level value-added measures, can be challenging to implement at scale, particularly in settings with weak administrative capacity such as Tanzania. For instance, maintaining the child-level databases of learning that are required to calculate teacher value-added, and ensuring the integrity of the testing system are non-trivial administrative challenges. Consequently, we decided to use a proficiency system where bonuses to teachers were paid on the basis of the number of children who passed an absolute threshold of learning. In addition, as the proficiency system is easier for teachers to understand, it is more likely to increase motivation among teachers and head teachers than a more complex and more difficult to understand system. However, proficiency bonus systems have some well known limitations such as their inability to adequately account for differences in the initial distribution of student preperation across schools and classrooms. In additoin, this type of system may lead teachers to focus on students who are close to the proficiency threshold, at the expense of students who are sufficiently above or below the threshold (Neal & Schanzenbach, 2010). Mbiti et al. (2017) found suggestive evidence of this heterogeneity: students who are well above or well below the passing threshold do not see any improvements in test scores, but students near the passing threshold see an increase in test scores of about 0.2 SD.

In this project, we compare the effectiveness two different incentive schemes, implemented in a set of Tanzanian primary schools. A budget of $150,000 for teachers' incentives was split between the two treatment arms each year proportional to the number of students enrolled. As a result, the total prize in each treatment arm was approximately $3 per student. All interventions were implemented by Twaweza in partnership with EDI, a Tanzanian research firm, and set of local district partners. Within each intervention arm, information describing the program was distributed to schools and the communities via public meetings in early 2015 and 2016. The implementation teams also conducted additional mid-year school visits to re-familiarize teachers with the program, gauge teacher understanding of the bonus payment mechanisms, and answer any remaining questions. At the end of the school year, all students in Grades 1, 2, and 3 in every school,including control schools, are tested in Swahili, English and, Math. As this test was used to determine teacher incentive payments, it is "high-stakes" (from the teacher's perspective).[1] The tests were developed by Tanzanian education professionals, following a similar test development framework as the Uwezo annual learning assessment that is widely used in East Africa.

2.2.1 Proficiency Thresholds (Levels) Design

The levels treatment pays teachers proportionally to how many skills students in grades 1-3 are able to demonstrate in Mathematics, Swahili and English.[2] Teacher's earn larger bonuses if

[1] Starting in 2015, English was removed from the curriculum of Grades 1 and 2, and as a consequence the curriculum for grade 3 was dramatically changed. We still test students in English.

[2] Due to the curriculum changes, the incentive for English was removed in 2016 for Grades 1 and 2.

their students' are proficient in more skills. They can also earn larger bonuses if their students' can master harder skills

Table 1 shows the skills to be tested in each grade-subject combination in the "levels" design. The total amount of money is then split across grades proportionally to the number of students enrolled in each grade and then divided equally among subjects and skills within each subject. At the end of the year teachers are paid according to the following formula:

$$P_j^s = \frac{X_s}{\sum_{i \in T} 1_{T_i > b_s}} \sum_{k \in J} 1_{T_k > b_s} \tag{1}$$

where P_j^s is the payment of teacher j for skill s, J is the set of students of teacher j, T_k is the test score of student k, b_s is the passing threshold for skill s, X_s is the total amount of money available for skill s, and T is the set of all students in schools across Tanzania in the "levels" treatment. For each skill teachers earn more money as more students in their class score higher than the passing threshold, and the payment is higher if fewer students are able to demonstrate learning in that skill. In other words, the reward is higher for teachers if students master a "harder" skill, which is defined by the overall passing rate of each skill.

Table 1: Skills tested in the "levels" design

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

An important feature of the "levels" design (even with multiple thresholds) is that it does not offer rewards for increasing test scores for all students (e.g., for students far above the highest threshold, increases in test scores do not increase teacher payouts), and these rewards are not continuous on teacher effort.

2.2.2 Pay for Percentile Design (Gains)

The pay for percentile design (gains) is based on the work of Barlevy and Neal (2012), who show that this incentive structure can, under certain conditions, induce teachers to exert socially optimal levels of effort. For each subject-grade combination we created student groups with similar initial learning levels based on test score data from the previous school year.[3] We then compensated teachers proportionally to the rank of their student at the end of the school year relative to other students with a similar baseline level of knowledge.

More formally, let s_i^{t-1} be the score of student i at the end of the previous school year. Students are divided into k groups according to s_i^{t-1}. We divided the total pot of money allocated to a subject-grade combination A^g into k groups, proportional to the number of students in the group. That is, $A_k = \frac{A^g n_k}{N_g}$, where N_g is the total number of students in grade g, n_k is the number of students in group k, and A_k is the amount of money allocated to group k. At the end of the year, we ranked students (into 100 ranks) within each group according to their endline test score s_i^t, and within each group we gave teachers points proportional to the rank of their students. For a student in the top 1% of group k a teacher received 99 points, and for a student in the bottom 1% the teacher received no points. Within each group we have:

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

where $b(i-1)$ is the amount of money paid for each student in rank i. Therefore, $b = \frac{A^g}{N_g}\frac{2}{99}$. The total money A^g allocated to a subject-grade is proportional to the number of students in each grade and is divided equally among the three subjects. In other words, $A^g = \frac{XN_g}{3\sum_{g=1}^{3} N_g}$, where X is the total amount of money available for the "gains" design. The total amount of money paid per rank is the same across all groups, in all subjects, and in all grades, and is equal to $b = \frac{X}{3\sum_{g=1}^{3} N_g}\frac{2}{99}$. For example, in the first year, the total prize money was $70,820 and total enrollment was 22,296 in "gains" schools. Therefore, the payment per "rank" was $0.0178. In other words, for a student in the top 1%, a teacher would earn $1.77 and for a student in the top 50% a teacher would earn $0.89.

2.3 Sample Selection

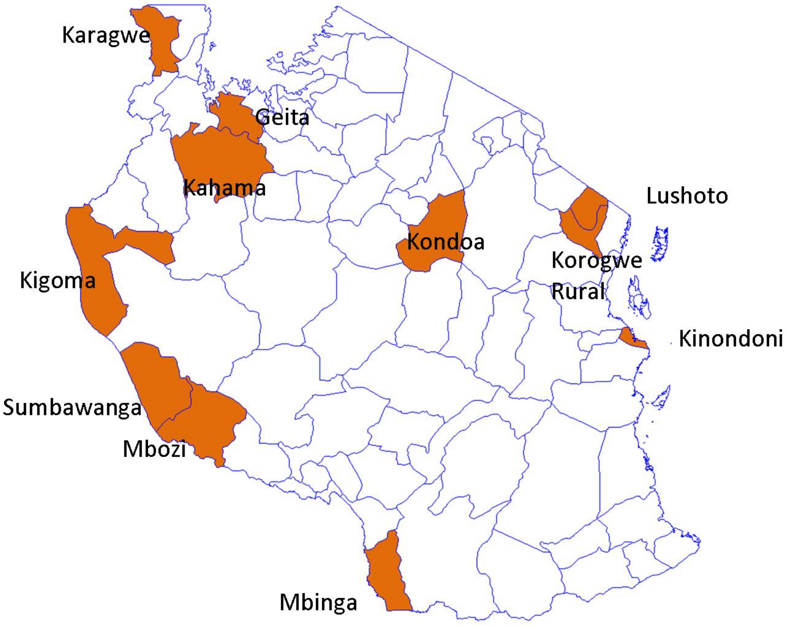
The teacher incentive programs were evaluated using a randomized design. First, ten districts were randomly selected (see Figure 1)[4]. The sample of 180 schools was taken from a previous field experiment (Mbiti et al., 2017) where all students in Grades 1, 2, and 3 had been tested at the end of 2014. These school level tests provided the necessary baseline student-level test score information that we needed in order to implement the "gains" treatment. Within each district, we randomly allocated schools to one of our three experimental groups. Thus, in each

[3] As noted previously, Grade 1 students were grouped according to historic test scores at the school level. Students without test scores in any other grade were grouped together in a "unknown" ability group.

[4] However, 11 district participate in the treatment arms, as one district was included in the program non-randomly by the implementing partner. We do not survey schools in this district.

district 6 schools were assigned to the "levels" treatment, 6 schools to the "gains" treatment, and 6 schools were served as our controls. In total, we have 60 schools in each group. The sample was also stratified by treatment of the previous RCT and by an index of the overall learning level of students in each school. All of our specifications control for the three levels of stratification: district, treatment in the previous RCT, and overall school quality.

Figure 1: Districts in Tanzania from which schools are selected



2.4 Data and Balance

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

Within each school we surveyed and tested a randomly selected sample of 40 students (10 students from Grades 1, 2, 3, and 4). 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. While students in Grade 4 in the first year were not tracked into Grade 5 due budget constraints. In the second year of the program we sampled an additional 10 incoming Grade 1 students. We collected a variety of data from our student sample including test scores, individual characteristics such as age and gender, and perceptions of the school environment. Crucially, the test scores collected on the sample of students are “low-stakes” for teachers and students and can serve as a more reliable measure of student learning. We can supplement this set of “low-stakes” student tests scores with the test scores from the “high-stakes” tests which are used to determine teacher bonus payments, and are conducted in all schools including control schools.

Table 2 shows the balance between students, school, teachers, and household characteristics in each treatment arm. Columns 1-3 shows the conditional mean of the variable for different treatment arms and Column 4 shows the p-value of a test of equality of these means. We show the conditional mean since all of our our analysis includes controls for the set of stratification variables used during randomization.[5]

[5] Randomization was stratified by district, previous treatment arm, and “quality strata”. The quality strata variable for schools was created using principal component analysis on students’ test scores. Schools were categorized into one of two strata depending on whether they were above or below the median for the first principal component. This was done to ensure balance in test scores at baseline.

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

	(1) Control	(2) Gains	(3) Levels	(4) p-value (all equal)
Panel A: Students				
Age	8.88	8.94	8.89	0.35
	(1.60)	(1.67)	(1.60)	
Male	0.50	0.48	0.51	0.05*
	(0.50)	(0.50)	(0.50)	
Swahili test score	0.00	0.01	0.01	0.15
	(1.00)	(1.00)	(0.98)	
English test score	-0.00	0.04	-0.02	0.71
	(1.00)	(1.03)	(1.04)	
Math test score	0.00	-0.01	-0.01	0.56
	(1.00)	(1.04)	(1.00)	
Tested in yr0	0.91	0.89	0.90	0.41
	(0.29)	(0.31)	(0.30)	
Tested in yr1	0.87	0.87	0.88	0.20
	(0.33)	(0.34)	(0.32)	
Tested in yr2	0.88	0.88	0.89	0.56
	(0.33)	(0.32)	(0.32)	
Poverty index (PCA)	0.01	-0.07	0.01	0.44
	(1.99)	(1.94)	(1.98)	
Panel B: Schools				
Total enrollment	643.42	656.35	738.37	0.67
	(331.22)	(437.74)	(553.33)	
Facilities index (PCA)	0.18	-0.11	-0.24	0.07*
	(1.23)	(0.97)	(1.01)	
Urban	0.15	0.13	0.17	0.92
	(0.36)	(0.34)	(0.38)	
Preschool	0.90	0.85	0.87	0.53
	(0.30)	(0.36)	(0.34)	
Piped Water	0.27	0.17	0.20	0.15
	(0.45)	(0.38)	(0.40)	
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	0.37	0.34	0.12
	(0.49)	(0.48)	(0.47)	
Age (Yrs)	38.43	37.37	37.88	0.04**
	(11.51)	(11.14)	(11.04)	
Experience (Yrs)	14.50	13.29	13.80	0.04**
	(12.14)	(11.36)	(11.21)	
Private school experience	0.02	0.01	0.03	0.11
	(0.15)	(0.11)	(0.16)	
Tertiary education	0.86	0.87	0.86	0.88
	(0.34)	(0.34)	(0.35)	

This tables presents the mean and standard error of the mean (in parenthesis) for several characteristics of students (Panel A), schools (Panel B), and teachers (Panel C) across treatment groups. Column 4 shows the p-value from testing whether the mean is equal across all treatment groups (H_0 := mean is equal across groups). The poverty index is the first component from a Principal Component Analysis of the following assets: Mobile phone, watch/clock, refrigerator, motorbike, car, bicycle, television, and radio. The school facilities index is the first component from a Principal Component Analysis of indicator variables for: Outer wall, staff room, playground, library, and kitchen. Standard errors are clustered at the school level for test of equality.

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

2.5 A Note on English

Starting in 2015 English was removed from the curriculum in Grade 1 and 2. As a consequence the curriculum for grade 3 changed. However, as there was little guidance from the Ministry of Education, there was a lot of variance in how the curriculum changes were actually implemented by schools. Some schools stopped teaching English in 2015, while others stopped in 2016. There was also no official guidance on whether to use Grade 1 English materials in Grade 3 as there were no new books issued to reflect the curriculum changes. As a result we dropped English from the incentives in Grade 1 and 2 in 2016, but included Grade 3 English teachers in the incentives. To avoid confusion, we also communicated that the end of year English test in 2016 would still use existing Grade 3 curriculum. Given the chaotic implementation of the curriculum reform, it's unclear how to interpret the results for English in Grade 1 and 2 in 2015, and for Grade 3 in both years.

2.6 Empirical Specification

As our interventions are assigned randomly to schools, we can estimate the effect of each treatment using OLS regressions. To estimate the effect that each intervention had on students test scores, we estimate the following equation:

$$Z_{isdt} = \delta_0 + \delta_1 Levels_s + \delta_2 Gains_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_w + \gamma_g + X_i \delta_3 + X_s \delta_4 + \varepsilon_{isd}, \qquad (2)$$

where Z_{isdt} is the test score of student i in school s in district d at time t. $Levels$ and $Gains$ are binary variables which capture the treatment assignment of each school. We normalize the test scores using the mean and variance of the control schools to facilitate a clear interpretation of our results. We include baseline test scores in our specifications to increase precision, as well as district and survey week fixed effects, γ_w, to account for any learning trends over time. By ensuring that the timing of our survey activities, including the low-stakes tests, was balanced on a weekly basis across treatment arms, we mitigate concerns that our results could be driven by imbalanced survey timing where, for example, students in a treatment group are systematically surveyed later, potentially giving them an advantage on the test. γ_g is a set of grade fixed effects, X_i is a series of student characteristics (age, gender and grade), X_s is a set of school and teacher characteristics (facilities, students per teacher, school committee characteristics, teacher's age, experience, qualifications, and gender). We can use a similar specification to examine teacher behavioral responses. As we have two sets of student test scores we can also examine the impacts of the incentives using the high stakes testing data. However, this would have limited controls given the lack of student characteristics in the data.

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

3 Results

3.1 Test scores

Table 3 shows the impact of the incentives on student learning in Math and Swahili using both the low-stakes data (Panel A) and the high-stakes data (Panel B). In the first year of the program, there were limited learning gains found in the low-stakes data, although the learning improvements found in proficiency levels system were consistently larger than those in the pay for percentile (gains) system (Panel A, Columns 1 and 2). The differences were statistically significant for Swahili at the 10 percent level (Panel A, Column 2). In the second year of the program both systems raised math test scores by a modest (0.07SD) but statistically significant amount (Panel A, Column 3). However, we only find statistically significant increases in Swahili for students in the proficiency levels system (0.11SD), although the difference between the two systems was no longer significant (Panel A, Columns 3 and 4).

As most of the existing literature on pay for performance uses a single high-stakes test to determine teacher rewards as well as evaluate the program, we present the treatment effects of our interventions using our high-stakes data in Panel B. Generally, the estimated treatment effects are larger compared to those estimated using the low-stakes data in Panel A. In addition, we see smaller differences in the estimated treatment effects between the two incentive designs using this data. Overall, across both years test scores increased between 0.10 and 0.14SD in Math, and 0.04 and 0.11SD in Swahili. The larger treatment effects found in the high-stakes data are likely driven by test-taking effort, where teachers have incentives to motivate their students to take the tests seriously. The importance of student test-taking effort has been documented in other settings such as an evaluation of teacher and student incentives in Mexico city (Behrman et al., 2015). As formal hypothesis tests in Panel C show that only the statistically significant differences between the two types of tests were in Math in year one, we argue that this provides some validation that the implementation protocols used were effective in minimizing manipulation and gaming.

Table 3: Effect on Test Scores

	(1)	(2)	(3)	(4)
	Year 1		Year 2	
	Math	Swahili	Math	Swahili
Panel A: Low-stakes				
Levels (α_1)	0.04	0.04	0.07*	0.11***
	(0.05)	(0.05)	(0.04)	(0.04)
Gains (α_2)	-0.01	-0.03	0.07**	0.06*
	(0.04)	(0.04)	(0.04)	(0.04)
N. of obs.	4,781	4,781	4,869	4,869
Gains-Levels (α_3) = $\alpha_2 - \alpha_1$	-0.05	-0.08*	0.00	-0.06
p-value ($H_0 : \alpha_3 = 0$)	0.23	0.08	0.92	0.17
Panel B: High-stakes				
Levels (β_1)	0.12**	0.13***	0.12***	0.17***
	(0.05)	(0.05)	(0.04)	(0.04)
Gains (β_2)	0.06	0.02	0.10**	0.08*
	(0.04)	(0.04)	(0.04)	(0.04)
N. of obs.	48,118	48,118	59,755	59,755
Gains-Levels (β_3) = $\beta_2 - \beta_1$	-0.06	-0.11**	-0.02	-0.09*
p-value ($H_0 : \beta_3 = 0$)	0.23	0.022	0.54	0.051
Panel C: High-stakes − Low-stakes				
$\beta_1 - \alpha_1$	0.07	0.08	0.05	0.05
p-value($\beta_1 - \alpha_1 = 0$)	0.11	0.09	0.19	0.29
$\beta_2 - \alpha_2$	0.07	0.05	0.02	0.02
p-value($\beta_2 - \alpha_2 = 0$)	0.10	0.27	0.53	0.63
$\beta_3 - \alpha_3$	0.00	-0.03	-0.03	-0.03
p-value($\beta_3 - \alpha_3 = 0$)	0.96	0.51	0.50	0.56

Clustered standard errors, by school, in parenthesis.

3.2 Spillovers to Other Grades and Subjects

As our incentives focused on Math, English, and Swahili in Grade 1, 2, and 3, schools could focus on these grades and subjects to the detriment of other grades and subjects. For example, schools may shift resources such as textbook purchases from higher grades to Grades 1, 2, and 3. Additionally, teachers may cut back on teaching non-incentivized subjects such as Science. On the other hand, if our incentive programs improve literacy and numeracy skills, they may promote student learning in other subjects. Moreover, these gains may persist over time, and continue to be reflected in test-scores. In order to asses these concerns, we examine learning outcomes in Grade 4 math and Grade 4 Swahili, and Science for Grades 1, 2, and 3.

The results of potential spillover effects are shown in Table 4. Panel A uses the data from a random sample of fourth grade students to ascertain any potential detrimental shifts in resources or focus away from higher grades. Overall, we do not see decreases in test scores of fourth graders, which suggests that schools were not disproportionately shifting resources away from higher grades. As third graders in the first year of our program transitioned to the fourth grade in the second year of our evaluation, we can use the fourth grade results in the second year (Panel A, Columns 3 and 4) to ascertain the persistence of any learning gains produced by the incentive programs. Although the point estimates are mostly positive, they are not significant, perhaps due to the smaller sample size.

Panel B shows the effects of our incentives on Science test scores. Contrary to the concerns of performance pay critics, the point estimates are generally positive suggesting that the incentives are more complementary to learning other subjects.

Table 4: Spillovers to Other Grades and Subjects

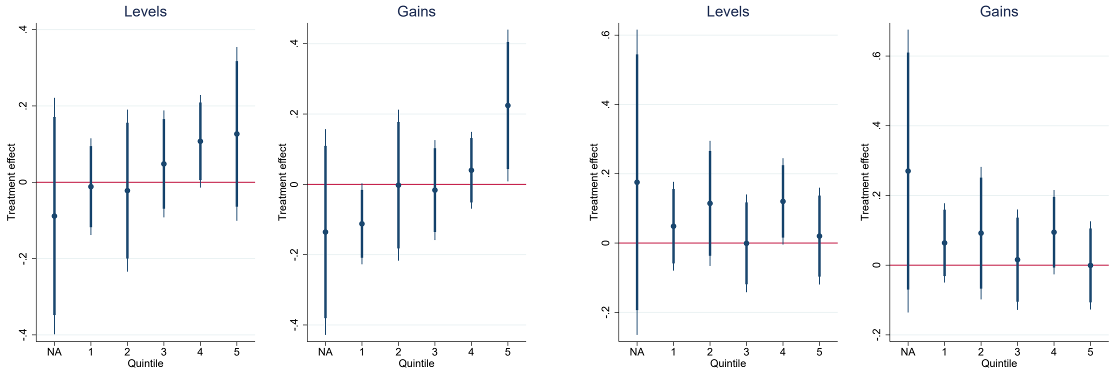
	(1)	(2)	(3)	(4)
Panel A: Grade 4				
	Year 1		Year 2	
	Math	Swahili	Math	Swahili
Levels (α_1)	0.13**	0.04	0.06	0.04
	(0.06)	(0.05)	(0.06)	(0.07)
Gains (α_2)	-0.03	-0.03	-0.00	0.03
	(0.05)	(0.05)	(0.06)	(0.06)
N. of obs.	1,513	1,513	1,482	1,482
Gains-Levels (α_3) $= \alpha_2 - \alpha_1$	-0.16**	-0.08	-0.06	-0.02
p-value ($H_0 : \alpha_3 = 0$)	0.01	0.13	0.27	0.79
Panel B: Science (Grades 1-3)				
	Year 1	Year 2		
Levels (α_1)	0.07	0.08		
	(0.06)	(0.06)		
Gains (α_2)	-0.00	0.08		
	(0.05)	(0.06)		
N. of obs.	4,781	4,869		
Gains-Levels (α_3) $= \alpha_2 - \alpha_1$	-0.07	-0.01		
p-value ($H_0 : \alpha_3 = 0$)	0.26	0.92		

Clustered standard errors, by school, in parenthesis.

3.3 Heterogeneity

We explore the heterogeneity in treatment effect across the distribution of student baseline test-scores in Figures 2 (Math) and 3 (Swahili). In the first year of the program, math teachers in the pay for percentile system (labeled "gains") focused a lot of attention on their very best students, whereas teachers in the proficiency system (labeled "levels") focused on the top half of their class (Figure 2a). In the second year of the program, we do not see such overt focus on top students in either incentive system (Figure 2b). In the first year of the program, Swahili teachers under the proficiency system again focus on the top half of the class, while teachers in the pay for percentile system focus on the the very best students (Figure 3a). In the second year of the program, Swahili teachers under the proficiency system seem to help all their students, while teachers in the pay for percentile system again focused on the very best students (Figure 3b).

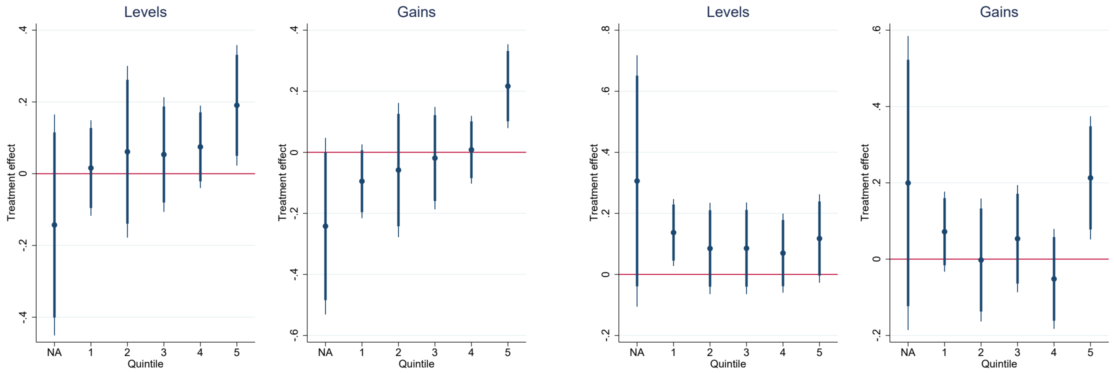
Figure 2: Math



(a) Year 1

(b) Year 2

Figure 3: Swahili



(a) Year 1

(b) Year 2

We further explore heterogeneity by student, school and teacher characteristics in Table 5. There were generally no differences in treatment effects by gender or pre-school attendance (see Panel A). However, we do find some limited differences by age for Swahili, where older students were less likely to benefit under the proficiency incentive system. Panel B explores heterogeneity by school characteristics. Overall, we do not observe differences in treatment effects by school facilities or distance to urban areas. Schools with higher pupil-teacher ratio benefited less in Math under the pay for percentile system. Panel C explores heterogeneity by teacher characteristics including an index of teacher content knowledge (created by an IRT model), teacher gender, age. Overall, we do not see any statistically significant patterns.

Table 5: Heterogeneity by School, Teacher, and Student Characteristics

Panel A: Student characteristics

	Math			Swahili		
	Male	Age	Preschool	Male	Age	Preschool
Levels*Covariate	-0.025	0.011	-0.10	0.011	-0.024	-0.024
	(0.039)	(0.015)	(0.064)	(0.039)	(0.016)	(0.058)
Gains*Covariate	0.0095	0.0089	0.018	0.0023	-0.0051	0.059
	(0.042)	(0.016)	(0.062)	(0.039)	(0.016)	(0.056)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650

Panel B: School characteristics

	Math			Swahili		
	Facilities	PTR	Distance	Facilities	PTR	Distance
Levels*Covariate	0.031	-0.00015	0.0040	0.033	-0.0019	-0.0022
	(0.023)	(0.0015)	(0.017)	(0.031)	(0.0013)	(0.016)
Gains*Covariate	-0.027	-0.0025**	0.0060	0.0024	-0.0021	-0.0047
	(0.026)	(0.0012)	(0.014)	(0.032)	(0.0013)	(0.013)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650

Panel C: Teacher characteristics

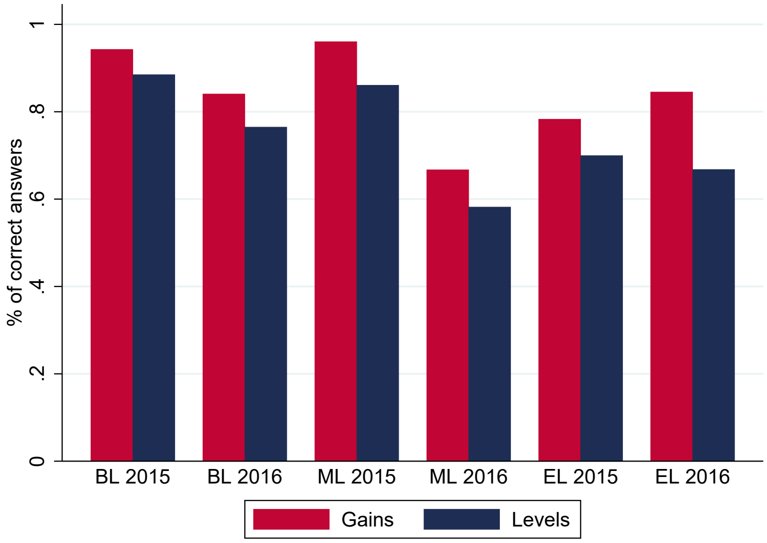
	Math			Swahili		
	IRT	Male	Age	IRT	Male	Age
Levels*Covariate	0.020	0.030	0.00088	0.00051	-0.083	0.0000096
	(0.037)	(0.070)	(0.0016)	(0.033)	(0.069)	(0.0011)
Gains*Covariate	-0.014	-0.016	0.00055	0.011	0.013	0.000033
	(0.038)	(0.060)	(0.0016)	(0.030)	(0.066)	(0.0011)
N. of obs.	9,650	9,650	9,650	9,650	9,650	9,650

Clustered standard errors, by school, in parenthesis.

3.4 Mechanisms

As the pay for percentile system is much more complex, our results may reflect differences in teacher understanding of the incentive systems. As part of the intervention we developed culturally appropriate materials, examples and analogies, which we used to communicate the details of the incentive programs with teachers. We also had intervention teams visit schools multiple times to reinforce the features of the program. During our visits and surveys we gave teachers a comprehension test to ensure they understood the details of the incentive program they were assigned to. We then debriefed with teachers to go over the answers to the questions to further ensure that teachers understood the design details. The results of the teacher comprehension tests are shown in Figure 4. As we asked different questions during each survey round (Baseline, Midline and Endline) we cannot compare the trends in understanding over time. Dsepite the lack of temporal comparability, the data suggest that teacher comprehension was generally high and roughly equal across both types of incentive programs. This provides some reassurance that our results are not driven by differences in program comprehension.

Figure 4: Do Teachers Understand the Interventions?



We examine teacher responsiveness to the incentives in Table 6. We use teacher presence in school and in the classroom as broad measures of teacher effort (Panel A). Overall, we do not find any differences in this dimension of teacher effort. We also examine student reports about teacher effort such as assigning homework and providing extra help in Panel B. We find some suggestive evidence of more help provided by teachers under proficiency systems in the first year but not in the second year. We also observe generally higher propensities to assign homework by teachers under our proficiency system compared to pay for percentile teachers. These differences are statistically significant (α_3), although the individual point estimates (α_1 and α_2) are not significant.

Table 6: Teacher Behavioral Responses

	(1)	(2)	(3)	(4)
Panel A: Spot-checks				
	Year 1		Year 2	
	In school	In classroom	In school	In classroom
Levels (α_1)	0.012	0.0061	-0.025	0.025
	(0.053)	(0.057)	(0.050)	(0.053)
Gains (α_2)	-0.012	-0.023	-0.0050	0.023
	(0.044)	(0.050)	(0.044)	(0.044)
N. of obs.	180	180	180	180
Gains-Levels (α_3) $= \alpha_2 - \alpha_1$	0.71	0.32	0.67	0.37
p-value ($H_0 : \alpha_3 = 0$)	0.65	0.60	0.71	0.97
suma	-0.02	-0.03	0.02	-0.00
Panel B: Student reports				
	Year 1		Year 2	
	Extra help	Homework	Extra help	Homework
Levels (α_1)	0.011	0.033	0.0052	0.0029
	(0.018)	(0.024)	(0.0096)	(0.018)
Gains (α_2)	-0.022	-0.0055	0.016*	-0.023
	(0.017)	(0.024)	(0.0097)	(0.019)
N. of obs.	9,006	9,006	9,557	9,557
Mean control	0.12	0.10	0.02	0.09
Gains-Levels (α_3) $= \alpha_2 - \alpha_1$	0.07	0.16	0.29	0.24
p-value ($H_0 : \alpha_3 = 0$)	-0.03*	-0.04	0.01	-0.03

Clustered standard errors, by school, in parenthesis.

4 Conclusion

In this paper we conduct a randomized experiment in a set of 180 Tanzanian schools, where we compare the effectiveness of two different teacher incentive programs, implemented in the same context and with the same budget. Specifically, we compare a simple threshold proficiency incentive design with a more complex pay for percentile system that theoretically induces optimal effort. Despite the theoretical advantage of the pay for percentile system, the simple proficiency system is at least as effective in improving student test-scores. Contrary to expectations, the pay for percentile lead teachers to almost exclusively focus on the very best students, while the simpler system benefited a wider range of students. Overall, a simpler system with multiple thresholds can actually outperform a more complex incentive system, especially in countries with low administrative capacity such as Tanzania.

References

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

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

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

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

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

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

Glewwe, P., & Kremer, M. (2006, June). Schools, Teachers, and Education Outcomes in Developing Countries. , *2*, 945-1017. Retrieved from http://ideas.repec.org/h/eee/educhp/2-16.html

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

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

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

Kremer, M., & Holla, A. (2009). Improving education in the developing world: What have we learned from randomized evaluations? *Annual Review of Economics*, *1*(1), 513-542. Retrieved from http://dx.doi.org/10.1146/annurev.economics.050708.143323 doi: 10.1146/annurev.economics.050708.143323

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

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

Loyalka, P. K., Sylvia, S., Liu, C., Chu, J., & Shi, Y. (2016). *Pay by design: Teacher performance pay design and the distribution of student achievement.* (mimeo)

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

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

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

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

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

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

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

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

Uwezo. (2012). *Are our children learning? annual learning assessment report 2011* (Tech. Rep.). Author. Retrieved from http://www.twaweza.org/uploads/

files/UwezoTZ2013forlaunch.pdf (Accessed on 05-12-2014)

Uwezo. (2014). *Are our children learning? literacy and numeracy across east africa 2013* (Tech. Rep.). Author. Retrieved from http://www.twaweza.org/uploads/files/UwezoTZ2013forlaunch.pdf (Accessed on 05-12-2014)

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

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

World Bank. (2013). *Government expenditure on education, total (% of gdp).* (data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS)

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