KiuFunza II: Pre-Analysis Plan*

May 4, 2016

1 Introduction

This document presents the pre-analysis plan for the KiuFunza II project – a randomized evaluation of the impact of two different teacher incentive programs. Both of the programs are implemented in the same context and with the same budget, which allows us to calculate their relative efficiency. The project is being implemented and overseen by Twaweza – a non-profit initiative in Tanzania – and the evaluation is led by Innovations for Poverty Action (IPA), with the lead researchers being Prof. Isaac Mbiti, Prof. Karthik Muralidharan, and Mauricio Romero. This document is being prepared and registered prior to the analysis of any of the first year's data. The aim of this pre-analysis plan is to discipline the types of heterogeneity that we will analyze and to specify the specific questions in the survey instruments that will be used to define variables of interest. The KiuFunza II project features two treatment arms and a control group and is implemented across a representative sample of 180 schools across 10 districts in Tanzania. The treatment arms are:

1. A "levels" incentive that provides teachers and head teachers with bonus payments conditional on the skills that each student is able to demonstrate in a basic literacy and numeracy tests.

2. A "gains" incentive, that provides teachers and head teachers with bonus payments that are proportional to the relative gains of each student, when compared to other students with similar initial learning levels.

Each of the two treatments was assigned to 60 randomly-selected schools (6 in each of the districts) and an additional 60 schools served as a control group. The appeal of the "levels" design lies in its simplicity. The design is easy to communicate to teachers and easy to implement at a large scale. However its main drawbacks are twofold: giving a non-trivial amount of money to teachers who exert no effort, and the risk of some students not benefiting from (or even being negatively affected by) the program due to threshold effects(Neal & Schanzenbach, 2010). The

*Prepared by Erin Litzow, Jessica Mahoney, Isaac Mbiti, Karthik Muralidharan, and Mauricio Romero

appeal of the "gains" design lies in its theoretical foundation, which guarantees optimal teacher effort (Barlevy & Neal, 2012), but its complicated design makes it difficult to communicate and implement at a large scale. The project is strongly influenced by the design and findings of the Andhra Pradesh Randomized Evaluation Studies (see Muralidharan and Sundararaman (2011)), the Kiufunza I project (see Mbiti and Muralidharan (n.d.)), and by the theoretical work of Barlevy and Neal (2012).

The Kiufunza I project compromised on the 'ideal' design of the incentive program and instead chose a design that was more 'implementable' at scale. This 'ideal' design (based on teacher-level value addition, such as the one outlined in Muralidharan and Sundararaman (2011)) can be challenging to implement at scale in the settings of weak administrative capacity that is typical of developing countries. In particular, maintaining child-level databases of learning to calculate value-added and ensuring the integrity of testing are non-trivial administrative challenges. Thus, the bonuses to teachers were paid on the basis of the number of children who passed an absolute threshold of learning as opposed to on the basis of improvements in learning. The expectation of the project implementing partner (Twaweza) was that the simplicity of such a scheme would make it easy to understand, therefore causing it to be more effective at motivating teachers and head teachers than would a more complex (and thus more difficult to understand) formula. At the same time, such a design has some well known limitations – especially with respect to creating unequal rewards for improving students who are at different points of the achievement distribution and at different distances from the threshold (see Neal and Schanzenbach (2010)). Mbiti and Muralidharan (n.d.) 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 seek to answer what the trade-offs between the 'ideal' and the 'simple' designs are. This is an important question as the evidence on teacher performance pay is mixed, with some studies finding large positive effects (Glewwe, Ilias, & Kremer, 2010; Lavy, 2002, 2009; Muralidharan & Sundararaman, 2011; Balch & Springer, 2015), and others finding little or no effects at all (Goldhaber & Walch, 2012; Goodman & Turner, 2013; Springer et al., 2011). However, these studies are not directly comparable as they are performed in different contexts, with different incentive structures, and with different budgets. For example, Muralidharan and Sundararaman (2011) study the performance of a sophisticated design that rewards students for gains over baseline levels of learning in India, while (Goodman & Turner, 2013) study the performance of a threshold design in New York City. The two intervention arms of the Kiufunza II project implement two different teacher incentive programs in the same context and with the same budget.

The "levels" treatment is implemented as a threshold tournament, paying a bonus directly to teachers for each student who passes grade-specific skills outlined in the national curriculum. The amount paid to teachers per skill is set to ensure that payouts are equal across each grade-subject

combination. Because the amount paid out to teachers for each skill passed is dependent on how many students pass, the exact bonus rate is unknown before the end of the year tests. Payment amounts are calculated this way to ensure budget comparability across the "levels" and "gains" designs, and also in an effort to reward teachers proportionally to how many skills their students learn as well as how hard each skills is to learn.

The "gains" treatment arm rewards teachers for learning gains even when a student cannot meet the "levels" threshold, or is well beyond it. This intervention arm is also implemented as a tournament design, as described by (Barlevy & Neal, 2012). At the beginning of the school year students are grouped based on their initial levels of learning (based on the test scores from the previous year, and schools' historic test scores for students in grade 1 who have not been previously tested). At the end of the school year students are tested again and ranked within each ability group; teachers are paid proportionally to their students' ranks. This tournament in "gains" recognizes all learning improvements that occur, regardless of a student's initial learning achievements. This will be particularly important for teachers with students who have initial learning levels that are very low and far from the passing threshold in the "levels" design, as well as for students with initial learning levels that are well above the passing threshold.

This document is outlined as follows: Section 2 presents details on the experimental design (interventions, sampling procedures, and data collected). Section 3 presents the hypotheses that will be tested at the end of the first and the second year, while section 4 presents the specific methodologies that will be used to test the hypotheses - including a mapping from survey questions to variable definitions.

2 Experimental Design

2.1 Sample Selection

The evaluation is being implemented in 180 government primary schools in 10 districts in Tanzania between 2015 and 2016. All interventions were implemented directly by Twaweza and its district partners, with the money given to teachers also coming from Twaweza. Within each intervention arm, information describing the program was distributed to schools and the communities via public meetings in early 2015 and 2016. The district partners then followed up with additional school visits in July and August of each year to re-familiarize teachers with the program, gauge teacher understanding of the bonus payment mechanisms, and answer any remaining questions. All students in Grades 1, 2, and 3 in every school are tested in Kiswahili, English and, Math at the end of the school year (in 2015 and 2016) to determine teacher incentive payments.[1] Tanzanian education

[1] From 2015 to 2016 English was removed from grade's 1 and 2 curriculum, and as a consequence the curriculum for grade 3 was dramatically changed. We still test students in English.

professionals, following a similar structure to the Uwezo annual learning assessment, developed the subject tests used as the basis for measuring learning levels.

The sample of 180 schools was taken from a previous RCT((Mbiti & Muralidharan, n.d.)) for which all students in Grades 1, 2, and 3 had been tested at the end of 2014. This information was used to implement the "gains" treatment. We randomly assigned schools into treatment, but as a requirement from the implementing partner the probability of assignment into different treatments was a function of the previous RCT treatment status. The following table presents the number of schools that had to be randomly allocated to each treatment arm depending on the treatment status of the previous RCT.[2] The sample was also stratified by an index of the overall school quality. All of our specifications will control for the three levels of stratification: district, treatment in the previous RCT, and overall school quality. We are powered to detect an effect size of about 0.2 SD (with a size 5% and a power of 90%) in each treatment arm, as well as to detect any difference between the treatments larger than this magnitude. Appendix B presents the details of the power calculations used.

Table 1: Treatment allocation

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

2.2 Intervention

During the first year, a budget of $150,000 for teachers' incentives was split between the two treatment arms. Total enrollment in grades 1-3 was 22,296 and 24,928 across "gains" and "levels" schools. The budget was allocated proportionally to the number of students (that is, the budget for "gains" schools was $70,820 and the budget for "levels" schools was $79,180).

2.2.1 Levels

The levels treatment pays teachers proportionally to how many skills students in grades 1-3 are able to demonstrate in Mathematics, Swahili and English.[3] The more skills a student is able to

[2]The previous RCT had two treatment arms and a control group. One treatment arm had a simplified "levels" incentive structure, and the other had the sample incentive structure plus capitation grants given to schools. The implementing partner wanted to asses the long term impact of a "levels" design, and therefore asked that 4/7 schools in this treatment arm, remained as "level" schools and that 2/4 control schools remained as controls. In Table 1 C1 is the "levels" treatment arm, C2 is the "levels"+ capitation grant treatment arm, and C3 is the control group.

[3]From 2015 to 2016 English was removed from grade's 1 and 2 curriculum, and as a consequence the curriculum for grade 3 was dramatically changed. The incentive for English was removed in 2016 for grade 1/2. As of 2015 several schools had stopped teaching English in Grades 1 and 2.

demonstrate, the more a teacher earns. The harder a skill is to master for a student, the more the teacher earns for students who master that skill.

Table 2 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. Table 8 in the appendix shows the total amount of money available to teachers for each skill in each subject-grade combination. 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. Notice that for each skill teachers earn more money as more students in their class score higher than the passing threshold, but the payment is higher if overall across Tanzania fewer students are able to demonstrate learning in that skill. In other words, the reward is higher for teachers if students learn "harder" skills (we let the overall passing rate define the difficulty of each skill).

Table 2: Skills tested in the "levels" design

Swahili	**English**	**Math**
	Grade 1	
Letters	Letters	Count
Words	Words	Numbers
Sentences	Sentences	Inequalities
		Add
		Subtract
	Grade 2	
Words	Words	Inequalities
Sentences	Sentences	Add
Paragraph	Paragraph	Subtract
		Multiply
	Grade 3	
		Add
Story	Story	Subtract
Comprehension	Comprehension	Multiply
		Divide

Mbiti and Muralidharan (n.d.) study the effect of a teacher incentive program with a threshold design in Tanzania. The payment structure in this design is a single passing threshold for each subject (either a student is able to demonstrate a certain level of proficiency in a subject or not) and the payment for each student that passed the threshold was fixed beforehand. Mbiti and Muralidharan (n.d.) find that students who are well above or well below the passing threshold

do not see any improvements in test scores, but that students near the passing threshold see an increase in test scores of about 0.2 SD. Our design was built upon their findings, hoping that by having payments per skills (i.e., having more thresholds per subject) we would be able to incentivize teachers to improve learning outcomes for students with lower levels of learning. Additionally, by using a tournament design with a piece rate that is higher for skills with lower overall passing rates, we tried to improve the match between teacher effort and payments.

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 continous on teacher effort.

2.2.2 Gains

The "gains" design loosely compensates teachers in proportion to the ranks of their students within comparison sets. The design is based on the work of Barlevy and Neal (2012), who show that this incentive structure can, under certain conditions, induce teachers to exert socially optimal levels of effort. For each subject-grade combination we created student groups with similar initial learning levels based on test score data from the previous school year.[4] We then compensate 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. We divide students into k groups according to s_i^{t-1}. We divide 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 rank students (into 100 ranks) within each group according to their endline test score s_i^t, and within each group we give teachers points proportional to the rank of their students. For a student in the top 1% of group k a teacher gets 99 points, and for a student in the bottom 1% he gets no points. Within each group we have that:

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

where $b(i-1)$ is the amount of money paid for each student in rank i. Therefore we have that $b = \frac{A^g}{N_g}\frac{2}{99}$. 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, $A^g = \frac{X N_g}{3\sum_{g=1}^{3} N_g}$,

[4] 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.

where X is the total amount of money available for the "gains" design. This means that 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 $b = \frac{X}{3\sum_{g=1}^{3} N_g} \frac{2}{99}$. Given the budget ($70,820) and the number of students enrolled (22,296) in "gains" schools, the payment per "rank" is $0.0178. In other words, for a student in the top 1% a teacher receives $1.77 and for a student in the top 50% a teacher receives $0.89.

2.3 Data and Balance

Data collection is carried out by Economic Development Initiatives (EDI), a well-established, Kagera-based, survey firm. Data will be collected four times, two times during each school year (at the beginning and the end of the year). Detailed information is gathered for each school (e.g., facilities, management practices, and head teacher characteristics) and each teacher (e.g., education, age, experience and, self-reported time use). Additionally, student information (e.g. test scores, age, gender, and perception of school environment) is collected for a randomly selected sample of 40 students per school (10 students from Grades 1, 2, 3, and 4).

To be clear, there are two sets of tests performed to measure student learning levels. One test, the intervention test, is implemented by Twaweza and given to all students in grades 1, 2, and 3 in every school (including control schools). This test is used to calculate the bonus to be paid to each teacher. The second test, the research test, is implemented by the survey firm and is only taken by 40 randomly selected students from each school. The research test is a low-stakes test. The intervention test is used to calculate the incentive payments, but the impact evaluation is done using the research test.

Table 3 and 4 show 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 every analysis we do conditions on the variables on which we stratified 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 3: Balance between treatment arms: student and household characteristics

	Control	Gains	Levels	p-value
Panel A: Student characteristics				
Age	8.32	8.38	8.37	0.67
	(0.049)	(0.063)	(0.055)	
Gender	0.50	0.47	0.52	0.047**
	(0.013)	(0.012)	(0.012)	
Attend pre-school	0.80	0.77	0.78	0.73
	(0.024)	(0.027)	(0.028)	
Swahili test score	5.6e-10	-0.014	-0.0092	0.99
	(0.064)	(0.074)	(0.075)	
English test score	-2.6e-12	0.034	0.037	0.91
	(0.069)	(0.057)	(0.077)	
Math test score	5.1e-09	-0.023	-0.013	0.97
	(0.063)	(0.077)	(0.076)	
Other subjects test score	-2.8e-09	-0.067	-0.038	0.81
	(0.071)	(0.076)	(0.082)	
Panel B: Household characteristics				
Breadwinner employed	0.87	0.90	0.82	0.080*
	(0.032)	(0.020)	(0.033)	
Radio	0.55	0.52	0.52	0.78
	(0.035)	(0.035)	(0.033)	
TV	0.14	0.12	0.14	0.86
	(0.032)	(0.030)	(0.031)	
Bicycle	0.33	0.30	0.34	0.67
	(0.036)	(0.032)	(0.038)	
Car	0.037	0.027	0.017	0.39
	(0.014)	(0.010)	(0.0072)	
Motorbike	0.10	0.083	0.077	0.67
	(0.021)	(0.017)	(0.016)	
Refrigerator	0.040	0.043	0.053	0.89
	(0.018)	(0.017)	(0.021)	
Watch/Clock	0.17	0.14	0.16	0.75
	(0.028)	(0.029)	(0.028)	
Mobile Phone	0.71	0.70	0.73	0.78
	(0.035)	(0.035)	(0.034)	
Own Land	0.90	0.92	0.90	0.82
	(0.030)	(0.025)	(0.020)	
Exp. in child's education	24927.0	24324.7	27031.9	0.79
	(2450.1)	(2907.1)	(2991.6)	
Give to school (kind or cash)	0.50	0.50	0.53	0.83
	(0.045)	(0.040)	(0.042)	
Wall made out of mud	0.58	0.54	0.56	0.81
	(0.052)	(0.050)	(0.048)	
Floor made out of mud	0.67	0.70	0.68	0.88
	(0.047)	(0.039)	(0.044)	
Roof is durable	0.88	0.81	0.80	0.11
	(0.025)	(0.029)	(0.033)	
Improved water source	0.55	0.57	0.58	0.88
	(0.053)	(0.047)	(0.052)	
Improved toilet	0.11	0.080	0.10	0.74
	(0.029)	(0.025)	(0.027)	
Electricity	0.22	0.21	0.19	0.89
	(0.037)	(0.034)	(0.033)	

Standard errors, clustered at the school level, in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Balance between treatment arms: school and teacher

	Control	Gains	Levels	p-value
Panel A: School characteristics				
Grd 1 enrollment	118.1	122.5	132.3	0.71
	(9.01)	(10.5)	(14.9)	
Grd 2 enrollment	107.0	117.6	129.3	0.44
	(9.05)	(12.1)	(15.8)	
Grd 3 enrollment	91.7	102.7	116.0	0.16
	(6.27)	(9.83)	(11.5)	
Total enrollment	643.4	656.4	738.4	0.51
	(42.8)	(56.5)	(71.4)	
Kitchen	0.22	0.18	0.22	0.87
	(0.054)	(0.050)	(0.054)	
Library	0.22	0.18	0.23	0.79
	(0.054)	(0.050)	(0.055)	
Playground	0.80	0.83	0.92	0.14
	(0.052)	(0.049)	(0.036)	
Staff room	0.97	0.88	0.87	0.059*
	(0.023)	(0.042)	(0.044)	
Outer wall	0.17	0.050	0.050	0.088*
	(0.049)	(0.028)	(0.028)	
Newspaper	0	0.017	0.10	0.025**
	(0)	(0.017)	(0.039)	
Urban	0.15	0.13	0.17	0.88
	(0.046)	(0.044)	(0.049)	
Classes outside	0.067	0.13	0.13	0.34
	(0.032)	(0.044)	(0.044)	
Electricity	0.18	0.10	0.10	0.35
	(0.050)	(0.039)	(0.039)	
Computers	0.017	0.033	0.050	0.57
	(0.017)	(0.023)	(0.028)	
Preschool	0.90	0.85	0.87	0.69
	(0.039)	(0.046)	(0.044)	
Breakfast	0.13	0.17	0.12	0.73
	(0.044)	(0.049)	(0.042)	
Lunch	0.15	0.15	0.17	0.96
	(0.046)	(0.046)	(0.049)	
Piped Water	0.27	0.17	0.20	0.41
	(0.058)	(0.049)	(0.052)	
No Water	0.13	0.30	0.22	0.078*
	(0.044)	(0.060)	(0.054)	
Single shift	0.63	0.62	0.62	0.98
	(0.063)	(0.063)	(0.063)	
Track students	0.083	0.050	0.083	0.68
	(0.036)	(0.028)	(0.036)	
Panel B: Teacher characteristics				
Male	0.39	0.35	0.33	0.49
	(0.040)	(0.042)	(0.038)	
Yr born	1975.4	1976.5	1975.6	0.55
	(0.75)	(0.72)	(0.70)	
Yr started teaching	1999.3	2000.6	1999.7	0.54
	(0.79)	(0.83)	(0.73)	
Yr started teaching at this school	2007	2002.3	2007.7	0.37
	(0.49)	(5.41)	(0.46)	
Private school experience	0.023	0.020	0.037	0.34
	(0.0072)	(0.0079)	(0.0090)	
Travel time (mins)	19.1	22.6	19.6	0.63
	(2.10)	(3.12)	(1.86)	
Tertiary education	0.82	0.83	0.83	0.95
	(0.032)	(0.029)	(0.027)	

Standard errors, clustered at the school level, in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.4 A Note on English

Starting in 2015 English was removed from grade's 1 and 2 curriculum, and as a consequence the curriculum for grade 3 changed. Many schools had stopped teaching English in 2015. For 2016 only English teachers in grade 3 participated in our program in 'gains' and 'level' schools. Its unclear therefore how to interpret the results for English in Grade 1 and 2 in 2015, and for Grade 3 in both years.

3 Hypotheses

The hypotheses we present mainly test whether our treatment had any impact on learning outcomes, whether there is any heterogeneity created by the incentives structure in each treatment arm, and tries to get at the mechanisms behind the effects, if any. Specifically, our hypotheses are: ***Main Outcomes***

1. Impact of incentivizing teachers, using our 'levels' incentive structure, on students' learning outcomes
    - H_a (H_0): 'levels' treatment has (no) positive impact on test scores.
2. Impact of incentivizing teachers, using our 'gains' incentive structure, on students' learning outcomes
    - H_a (H_0): 'gains' treatment has (no) positive impact on test scores.
3. The relative impact of the 'levels' incentive structure, compared to the 'gains' incentive structure
    - H_a (H_0): 'gains' treatment has (no) greater impact on test scores than the 'levels' incentives structure.

Heterogeneity

4. Impact on learning outcomes by student ability
    - H_a (H_0): Treatment - 'gains' and 'levels' - impact on knowledge is different (the same) for all students.

Channels

5. Impact on teacher's behavior

- H_a (H_0): Treatment - 'gains' and 'levels' - has (no) impact on teacher's behavior (classroom presence, assignments, tutoring, etc.) in focal subjects in focal years.

6. Impact on school expenditure. This analysis will also be performed by grade.
   - H_a (H_0): Treatment - 'gains' and 'levels' - has (no) positive impact on text book and teaching input expenditures.
7. Impact on school's schedule
   - H_a (H_0): Treatment - 'gains' and 'levels' - has (no) impact on the number of hours taught in different subjects.
   - H_a (H_0): Treatment - 'gains' and 'levels' - has (no) impact on teacher assignments across subjects and grades.

The next section presents a detailed methodology of how we plan to test these hypotheses.

4 Methodology

In order to test the hypotheses outlined above we perform OLS regressions, clustering standard errors at the school level. We also perform non-parametric analysis using lowess (and bootstrapping to calculate clustered standard errors) in order to assess how treatment effects vary by baseline ability of the students. When appropriate, we control for student, teacher, and schools (baseline) characteristics. Table 5 presents the characteristics we control for and the corresponding question in the survey questionnaires from which they are taken. We do not directly control for all these characteristics. Instead, we create family-wise indices using principal component analysis (PCA) and control for these indices (e.g., school-infrastructure index).

Table 5: Control Variables

	Questionnaire	Question
Panel A: Student		
Age	Student (R7)	DETAILS Q.5
Gender	Student (R7)	DETAILS Q.6
Student attended pre-school (nursery) before attending elementary school	Student (R7)	DETAILS Q.7
Student has seen exercise books with Uwezo tests	Student (R7)	DETAILS Q.8
Student has footwear	Student (R8)	Observations of Student Q.1
What kind of footwear	Student (R8)	Observations of Student Q.2
Student has socks	Student (R8)	Observations of Student Q.3
Is the student dirty?	Student (R8)	Observations of Student Q.4
Is the student's uniform dirty?	Student (R8)	Observations of Student Q.5
Is the student's uniform torn?	Student (R8)	Observations of Student Q.6
Does the student have visible ringworm?	Student (R8)	Observations of Student Q.7
Baseline kiswahili score	Student test (R7)	Kiswahili set
Baseline english score	Student test (R7)	English set
Baseline math score	Student test (R7)	Hisabati set
Panel B: School		
School has a kitchen	School (R6)	Facilities Q.1
School has a library	School (R6)	Facilities Q.1
School has a playground	School (R6)	Facilities Q.1
School has a staff-room	School (R6)	Facilities Q.1
School has piped water	School (R6)	Facilities Q.1
School holds classes outside besides physical education	School (R6)	SCHOOL FACILITIES Q.6
Number of toilets/latrines per student	School (R6)	Facilities Q.1 (R6) / ((R7) Grade Breakdown Q.1+ Q.2)
Number of classroom per student	School (R6)	Facilities Q.1 (R6) / ((R7) Grade Breakdown Q.1+ Q.2)
Number of teachers per student	School (R7)	TEACHER ROSTER Q.2/(Grade Breakdown Q.1+ Q.2)
Volunteer/Student ratio	School (R7)	VOLUNTEERS Q.1/ (Grade Breakdown Q.1+ Q.2)
School is located in an urban area	School (R1)	B4.Q1
Schools has a public notice board with current expenditures	School (R6)	EXPENSES Q.9
Size of the school committee	School (R6)	SCHOOL COMMITTEE Q.1
Proportion of the members of the school committee that are female	School (R6)	SCHOOL COMMITTEE Q.2/Q.1
Proportion of the members of the school committee that are teachers	School (R6)	SCHOOL COMMITTEE Q.4/Q.1
Proportion of the members of the school committee that are parents	School (R6)	SCHOOL COMMITTEE Q.5/Q.1
Number of times school committee met in previous year	School (R6)	SCHOOL COMMITTEE Q.15
Managment PCA index	School (R7)	MANAGEMENT Q.1-Q.12
Head teacher digit span	School (R7)	DIGIT SPAN Q.1-Q.9
Personality PCA index	School (R7)	PERSONALITY QUESTIONS Q.1-Q.17
Gender of the head teacher	Teacher (R7)	Q.3
Year in which head teacher was born	Teacher (R7)	Q4
Year in which head teacher started teaching	Teacher (R7)	Q6
Year in which head teacher started teaching at that school	Teacher (R7)	Q7
Whether the head teacher has post-secondary education	Teacher (R7)	Q12
Salary (of head teacher)	Teacher (R7)	COMPENSATION Q.1
Panel C: Average Teacher Characteristics		
Proportion of teachers that are male	Teacher (R7)	Q3
Average year in which teachers at the school are born	Teacher (R7)	Q4
Average year in which teachers at the school started teaching	Teacher (R7)	Q6
Average year in which teachers at the school started teaching at that school	Teacher (R7)	Q7
Average salary	Teacher (R7)	COMPENSATION Q.1
Proportion of teachers with post-secondary school education	Teacher (R7)	Q12
Panel D: Teacher Characteristics by grade and subject		
Gender of the teacher	Teacher (R7)	Q3
Year in which teacher was born	Teacher (R7)	Q4
Year in which teacher started teaching	Teacher (R7)	Q6
Year in which teacher started teaching at that school	Teacher (R7)	Q7
Salary	Teacher (R7)	COMPENSATION Q.1
Whether the teacher has post-secondary education	Teacher (R7)	Q12
Whether the teacher is the Head (or deputy head) Teacher	Teacher (R7)	HEAD TEACHER Q.7

4.1 Effect on test scores: H1, H2, and H3

To estimate the effect on test scores (and test hypotheses 1, 2, and 3) we estimate the following equation

$$Z_{igsdt} = \alpha_0 + \alpha_1 G_s + \alpha_2 L_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_s\beta_2 + X_{gs}\beta_4 + \varepsilon_{igsdt},$$

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, G is a dummy variable that indicates whether the school was in the 'gains' group, L is an indicator variable of whether the school was in the 'levels' group, γ_d is a set of district fixed effects, γ_w is a set of week fixed effects[6], γ_g is a set of grade fixed effects, X_i is a series of student characteristics (see panel A in table 5), X_s is a set of school and average teacher characteristics (see panels B and C in table 5), and X_{gs} is a set of teacher characteristics (for a particular grade/subject, see panel D in table

[6]The time difference between the EDI and the Twaweza varies across schools, but the timing is balanced across treatment arms. The week fixed effects should increase the precision of our estimates.

5). The coefficients of interest are the αs, which test hypotheses 1-3 above. We will analyze each subject separately. Specifically, we have:

- H.1
  - H_0: 'levels' treatment has no positive impact on test scores (i.e. $\alpha_2 \leq 0$)
  - H_a: 'levels' treatment has a positive impact on test scores (i.e. $\alpha_2 > 0$)
- H.2
  - H_0: 'gains' treatment has no positive impact on test scores (i.e. $\alpha_1 \leq 0$)
  - H_a: 'gains' treatment has a positive impact on test scores (i.e. $\alpha_1 > 0$)
- H.3
  - H_0: 'gains' treatment has no greater impact on test scores than the 'levels' treatment (i.e. $\alpha_1 \leq \alpha_2$)
  - H_a: 'gains' treatment has a greater impact on test scores than the 'levels' treatment (i.e. $\alpha_1 > \alpha_2$)

4.2 Heterogeneity: H4

A key difference between the two designs is the following: Holding other teachers effort constant, the payoff to teacher effort is much more non-linear in the "levels" incentive treatment and will typically be higher for students near the thresholds of passing the test. Note also that returns to improving student learning are not continuous in improvement and rise sharply below a threshold and fall sharply after a threshold.

On the other hand, the payoff to teacher effort is much more continuous in the "gains" incentive treatment in two ways. First, all students matter equally because each student is compared against other students starting at the same point, and teachers are rewarded at the same rate for improvements in the performance of any student. Second, the returns increase continuously in improvements for each student because each percentile improvement in test scores is rewarded with a proportionately larger bonus. This aspect of pay-for-percentile mitigates against the standard concern in tournaments of non-linear payoffs that are highly sensitive to small amounts of noise.

An important implication of this difference is that we do not expect to find heterogeneity in impacts as a function of student's initial rank in the test score distribution in the "gains" treatment. However, in the "levels" treatment, we do expect heterogeneity. Specifically, we will test for this heterogeneity by using the gains in the control group as a benchmark and calculate the marginal return to a teacher of improving the test scores by the average treatment effect that we find (say

0.1 to 0.2 standard deviations) for each student. One testable prediction is that students whose gains provide greater marginal returns to teacher effort will see larger gains.[7]

If this prediction holds in the data, it would suggest that the model of education production is one where teacher attention in the classroom is differentially allocated across students based on marginal returns to teacher effort. If we find that the prediction is not supported in the data, that would imply either that teachers did not understand the formula well enough to optimize, or that the model of education production is one where classroom instruction is a public good and not differentiated much by student.

In order to test hypothesis 4 we run a locally weighted regression of the end line test scores on the baseline score of students. Specifically, we estimate the following equation

$$Z_{it} = f(\alpha_0 + \alpha_1 F(Z_{i,t=0}) + \varepsilon_{it}),$$

where F is the CDF of the baseline scores of students. Let $f(x;T)$ denote the estimated relation between baseline score and endline score for treatment T using the command *lowess* in STATA. The point-wise treatment effect is calculated as $g(x;T) = f(x;T) - f(x;Control)$ and the confidence intervals are estimated using bootstrapping. This enables us to estimate how the treatment effect varies for students with different initial abilities or knowledge. We also do a semi-parametric estimation, in which we regress both baseline and endline scores on student, school, and teacher controls, and then perform the above procedure on the residuals of those regressions. We also perform a semi-parametric test where we split the data by students' baseline test scores and test hypotheses 1-3 in the sub-samples.

Our hypothesis are:

- H.4.A
  - H_0: The treatment effect is the same for all students, regardless of the expected payoff per unit of effort
  - H_a: The treatment effect depends on the expected payoff per unit of effort
- H.4.B
  - H_0: The treatment effect is the same for all students, regardless of the students' initial levels of learning
  - H_a: The treatment effect depends on students' initial levels of learning
- H.4.C

[7] Note that the ability groups in the 'gains' are not perfect, and there is some left-over heterogeneity within group, which could lead teachers to focus on some students (e.g., the best ones within each group). We will search for this type of behavior the data.

- H_0: The treatment effect is the same for all students, regardless of the students' endline levels of learning
- H_a: The treatment effect depends on students' endline levels of learning

4.3 Effect on teachers: H5

To estimate the effect on teacher behavior we estimate the following equation

$$Y_{gsd} = \alpha_0 + \alpha_1 G_s + \alpha_2 L_s + \gamma_d + X_i\beta_1 + X_s\beta_2 + \varepsilon_{igsdt},$$

where Y_{igsd} is the outcome variable that measures the behavior of teacher i in school s in district d, G is a dummy variable that indicates whether the school was in the 'gains' group, L is an indicator variable of whether the school was in the 'levels' group, γ_d is a set of district fixed effects, X_s is a set of school characteristics (see panels B and C in table 5) and X_i is a set of teacher characteristics (see panel D in table 5). The coefficients of interest are the αs which test hypothesis 5 above. The outcome variables that we will focus on are presented in table 6 with the respective question in the surveys used to measure them. Most of the information used here is not self-reported, but instead reported by students or observed by the survey team. To avoid multiple-inference issues we create family-wise indices using principal component analisys, and only study the components within an index when we find an effect on the index. We create indices for: classroom environment, student reported teacher demeanor, and teacher behavior during class.

Table 6: Teacher outcomes

	Questionnaire	Question
Student missed school	Student	Classroom Environment Q.5
Singing in class	Student	Classroom Environment Q.7
Take books home	Student	Classroom Environment Q.9
Tutoring	Student	Teacher Perceptions Q.1
Teacher knows student name	Student	Teacher demeanor Q.1
Teacher leaves classroom	Student	Teacher demeanor Q.5
Teacher missed school	Student	Teacher demeanor Q.6
Positive feedback for doing good work	Student	Teacher demeanor Q.8
Homework on last normal schooling day	Student	Homework Q.1
How long is the assignment?	Student	Books Q.4
Assignment graded	Student	Books Q.5
Unobserved Classroom observation	Observation	Unobserved Classroom Observation Q.2
Children's work displayed on the walls	Observation	Classroom Environment Q.9
Other materials on the walls	Observation	Classroom Environment Q.10
Charts and posters	Observation	Classroom Environment Q.11
Trash in the classroom	Observation	Classroom Environment Q.12
Students misbehaving	Observation	Pupil Behavior Q.1
Teacher using textbook	Observation	Tools used by Teacher Q.1
Students using textbook	Observation	Tools used by Teacher Q.2
Blackboard written by teacher	Observation	Tools used by Teacher Q.3
Blackboard written by children	Observation	Tools used by Teacher Q.4
Teacher goes to students individually	Observation	Teacher demeanor Q.1
Teacher call students by name	Observation	Teacher demeanor Q.6
Teacher smiles and jokes	Observation	Teacher demeanor Q.8
Teachers uses aggressive language	Observation	Teacher demeanor Q.9
Teacher uses phone during class	Observation	Teacher demeanor Q.12
Teacher left classroom during class	Observation	Teacher demeanor Q.13
Language used during class	Observation	Language Q.1
Drill/Memorization	Observation	Teacher asking questions Q.1
Systematic explaining	Observation	Teacher asking questions Q.3
Demostrate knowledge	Observation	Teacher asking questions Q.4
Best students seating	Observation	Counting materials Q.4
Students with textbook	Observation	Counting materials Q.1
Students with exercise book and pencil	Observation	Counting materials Q.2
Hours taught	Teacher	DETAILS Q.1
Number of tests	Teacher	DETAILS Q.2
Tutoring	Teacher	DETAILS Q.5
Remedial teaching	Teacher	DETAILS Q.6
Number of tests	Teacher	DETAILS Q.5
Number of tests	Teacher	DETAILS Q.6
Time grading homework	Teacher	Daily activities roster Q.3
Time grading test	Teacher	Daily activities roster Q.3
Time extra classes	Teacher	Daily activities roster Q.3
Time in school	Teacher	3.5 Time Use Q.1-Q.2

Specifically, we have:

- H.5
  - H_0: Treatment has no impact on teacher behavior (i.e., $\alpha_i = 0$, $i = 1$ for 'gains' treatment and $i = 2$ for 'levels' treatment)
  - H_a: Treatment has an impact on teacher behavior (i.e., $\alpha_i \neq 0$, $i = 1$ for 'gains' treatment and $i = 2$ for 'levels' treatment)

4.4 Effect on school: H6 and H7

To estimate the effect on school behavior we estimate the following equation

$$Y_{sdt} = \alpha_0 + \alpha_1 G_s + \alpha_2 L_s + \gamma_d + X_s \beta_1 + \varepsilon_{sdt},$$

where Y_{sdt} is the outcome variable that measures the behavior of school s in district d at time t, G is a dummy variable that indicates whether the school was in the 'gains' group, L is an indicator variable of whether the school was in the 'levels' group, γ_d is a set of district fixed effects, X_s are a set of school characteristics (see panels in table C and D 5).The coefficients of interest are the αs which test hypotheses 6 and 7 above. The outcome variables that we will focus on are presented in table 7, along with the respective question in the surveys used to measure them.

Table 7: School outcomes

	Questionnaire	Question
Administrative expenses per student	School	School Expenses Q.1
Student expenses per student	School	School Expenses Q.1
Teaching aid expenses per student	School	School Expenses Q.1
Teacher expenses per student	School	School Expenses Q.1
Construction expenses per student	School	School Expenses Q.1
Textbook expenditure per student	School	TEXTBOOK AND PRACTICE EXAMS Q.1
Textbook expenditure per student per grade	School	TEXTBOOK AND PRACTICE EXAMS Q.1
Enrollment per grade	School	(Y2 Baseline) 4.1 GRADES Q.1 & 6.3 ENROLLMENT
Time spend per subject per week	School	TIME SPENT ON SUBJECTS Q.1-Q.10

Specifically, we have:

- H.6.a
  - H_0: Treatment has no impact on text book and teaching input expenditure (i.e., $\alpha_i = 0$, $i = 1$ for 'gains' treatment and $i = 2$ for 'levels' treatment)
  - H_a: Treatment has no impact on text book and teaching input expenditure (i.e., $\alpha_i \neq 0$, $i = 1$ for 'gains' treatment and $i = 2$ for 'levels' treatment)
- H.7
  - H_0: Treatment does not increase the amount of hours taught in incentivized subjects or the resources invested in incentived grades (i.e., $\alpha_i = 0$, $i = 1$ for 'gains' treatment and $i = 2$ for 'levels' treatment)
  - H_a: Treatment increases the amount of hours taught in incentivized subjects and/or the resources invested in incentived grades (i.e., $\alpha_i \neq 0$, $i = 1$ for 'gains' treatment and $i = 2$ for 'levels' treatment)

4.5 Other tests

4.5.1 Survey test vs intervention test

As mentioned before there are two sets of tests performed to measure student learning levels. Twaweza tests all students in grades 1, 2 and 3 in "levels" and "gains" schools to calculate the teacher payments. However, it also tests all Grades 1, 2, and 3 students in control schools. Additionally, EDI tests 40 students in all schools (10 each in grades 1, 2, 3, and 4) which allows us to compare treatment effects for all treatments compared to control schools, in a low stakes exam. Although the main analysis will be done using the EDI test, we test whether the treatment effects are different for the Twaweza test than for the EDI test. This will allow us to infer whether there is any cramming before the Twaweza exam and whether there is any teaching to the test (the EDI test has a wider range of questions).

4.5.2 Effect on non-incentivized subjects and grades scores

To estimate any spillover effect on non-incentivized grades (if resources at the school level are shifted by the treatment) we estimate the following equation

$$Y_{gsdt} = \alpha_0 + \alpha_1 G_s + \alpha_2 L_s + \gamma_z Y_{gsd,t-1} + \gamma_d + X_s\beta_1 + \varepsilon_{igsdt},$$

where Y_{igsdt} is a measure of learning for grade g at school s in district d at time t, G is a dummy variable that indicates whether the school was in the 'gains' group, L is an indicator variable of whether the school was in the 'levels' group, γ_d is a set of district fixed effects, X_s is a set of school characteristics (see panels B and C in table 5. The coefficients of interest are the αs. For Y_{igsdt} we use the average score and the pass rate in the national Grade 4 and 7 examinations, the score of the students tested in Grade 4, and the score for science in Grades 1-3. Specifically, we have:

- – H_0: Treatment has no impact on test scores (i.e., $\alpha_i = 0$, $i = 1$ for 'gains' treatment and $i = 2$ for 'levels' treatment)
  – H_a: Treatment has an impact on test scores (i.e., $\alpha_i \neq 0$, $i = 1$ for 'gains' treatment and $i = 2$ for 'levels' treatment)

4.5.3 Heterogeneous treatment effects by variance within a classroom

To estimate heterogeneous treatment effects by within classroom variance we perform the following regression

$$Z_{igsdt} = \alpha_0 + \alpha_1 G_s + \alpha_2 L_s + \lambda_0 C_i + \lambda_1 G_s \times V_{gs} + \lambda_2 L_s \times V_{gs} + \gamma_d + \gamma_w + \gamma_g + X_s\beta_1 + X_h\beta_2 + \varepsilon_{igsdt},$$

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, G is a dummy variable that indicates whether the school was in the 'gains' group, L is an indicator variable of whether the school was in the 'levels' group, γ_d is a set of district fixed effects, γ_w is a set of week fixed effects, γ_g is a set of grade fixed effects, and X_s is a set of school and teacher characteristics (see panels B and C in table 5. Finally V_{gs} is a variable that measures the variance of students ability within classroom (e.g., variance or interquantile range). The coefficients of interest are the λs, which test if there are any heterogeneous treatment effects by student characteristics. The idea is that it might be easier for teachers to improve outcomes in classrooms with low variance since students have similar initial learning levels or ability, and therefore teaching to the "median student" benefits more students.

4.5.4 Heterogeneous treatment effects by student characteristics

To estimate heterogeneous treatment effects by student characteristics we perform the following regression

$$Z_{igsdt} = \alpha_0 + \alpha_1 G_s + \alpha_2 L_s + \lambda_0 C_i + \lambda_1 G_s \times C_i + \lambda_2 L_s \times C_i + \gamma_d + \gamma_w + \gamma_g + X_s\beta_1 + X_h\beta_2 + \varepsilon_{igsdt},$$

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, G is a dummy variable that indicates whether the school was in the 'gains' group, L is an indicator variable of whether the school was in the 'levels' group, γ_d is a set of district fixed effects, γ_w is a set of week fixed effects, γ_g is a set of grade fixed effects, and X_s is a set of school and teacher characteristics (see panels B and C in table 5. Finally C_i is a student characteristic (grade, gender, age, proxys for socio-economics status, as in panel A of table 5). The coefficients of interest are the λs, which test if there are any heterogeneous treatment effects by student characteristics.

4.5.5 Heterogeneous treatment effect by school characteristics

To estimate heterogeneous treatment effect by school characteristics we perform the following regression

$$Z_{igsdt} = \alpha_0 + \alpha_1 G_s + \alpha_2 L_s + \lambda_0 C_s + \lambda_1 G_s \times C_i + \lambda_2 L_s \times C_i + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_p\beta_2 + X_h\beta_3 + \varepsilon_{igsdt},$$

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, G is a dummy variable that indicates whether the school was in the 'gains' group, L is an indicator variable of whether the school was in the 'levels' group, γ_d is a set of district fixed effects, γ_w is a set of week fixed effects, γ_g is a set of grade fixed effects, X_i is a set of student characteristics (see panel A in table 5), and X_p is a set of teacher characteristics (see panels B and C in table 5). Finally C_s is a school characteristic: An index between 0 and 6 of school facilities; whether the school has

piped water; whether the school has a single shift; the size of the school committee; the number of times the school committee met in 2014; the proportion of females, teachers and parents in the school committee; and whether the school keeps records of its expenses (and their quality) and publishes their expenditures on public noticeboards. We will also look for heterogeneity by head teacher characteristics (age, previous experience and education). See panel C in table 5. The coefficients of interest are the λs, which test if there are any heterogeneous treatment effects by school characteristics. Additionally, we will use the first component from a principal component analysis (PCA), using all the characteristics mentioned above, as a proxy for school quality. This index will explain variation across schools and allow for the use of a single index of school quality that is determined by the data itself, taking into account that several of the variables we used to measure school quality are correlated; however, the interpretation of this index and the associated coefficients is not as straightforward.

4.5.6 Heterogeneous treatment effects by teacher characteristics

To estimate heterogeneous treatment effects by teacher characteristics we perform the following regression

$$Z_{igsdt} = \alpha_0 + \alpha_1 G_s + \alpha_2 L_s + \lambda_0 C_s + \lambda_1 G_s \times C_i + \lambda_2 L_s \times C_i + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_p\beta_2 + X_h\beta_3 + \varepsilon_{igsdt},$$

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, G is a dummy variable that indicates whether the school was in the 'gains' group, L is an indicator variable of whether the school was in the 'levels' group, γ_d is a set of district fixed effects, γ_w is a set of week fixed effects, γ_g is a set of grade fixed effects, X_i is a set of student characteristics (see panel A in table 5), and X_s is a set of school characteristics (see panels B and C in table 5). Finally C_p is an average of teacher characteristics per school: proportion of male teachers, average year of birth, average year started teaching, average year started teaching at this school, proportion with experience in private schools, average time at school and average salary. See panel C in table 5. Aditionally, we will test heterogeneity by the amount of time teachers dedicate to each subject.[8] The coefficients of interest are the λs, which test if there are any heterogeneous treatment effects by teacher characteristics. As with school characteristics, we will use the first component from a principal component analysis (PCA), using all the characteristics mentioned above, as a proxy for teacher quality.

[8]The idea behind heterogeneity by teacher's schedule is to test any changes in effort across subjects. For example, take two teachers - one teachers English and Swahili and the other Math and Swahili. Since we believe English is more difficult then we may expect to see the teacher who has English invest more in Swahili than the teacher who has Math and Swahili (i.e., they internalize the effort costs and adjust accordingly).

4.5.7 Teacher Learning

Something we would like to explore is "teacher learning". By this, we refer to the possibility that after the first year of the program teachers may learn something about their students' learning abilities as well as their own teaching techniques. First, we would like to explore how internal ranking in schools compares to overall student ability distribution, and to see whether teachers with students that are "worst than they think" (for example, the highest rank student is below the average) perform in the second year compared to the first year, as well as those with students that are "better than they think". Additionally, we would like to see how teacher's performance in the first year correlates to performance in the second year, when there is variation in the quality of the students they get, to see if teachers that perform above their expected value added in the first year, also perform better in the second year.

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

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

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

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

Mbiti, I., & Muralidharan, K. (n.d.). *Inputs, incentives, and complementarities in primary education: Experimental evidence from tanzania.*

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., & 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

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*.

A Extra tables

Table 8: Total amount of money available to teachers for each skill in each subject-grade combination

	Money Paid
Grade 1	
Swahili	
Letters	$ 3,118
Words	$ 3,118
Sentences	$ 3,118
English	
Letters	$ 3,118
Words	$ 3,118
Sentences	$ 3,118
Math	
Count	$ 1,871
Numbers	$ 1,871
Inequalities	$ 1,871
Add	$ 1,871
Subtract	$ 1,870
Grade 2	
Swahili	
Words	$ 2,955
Sentences	$ 2,955
Paragraph	$ 2,955
English	
Words	$ 2,955
Sentences	$ 2,955
Paragraph	$ 2,955
Math	
Inequalities	$ 2,216
Add	$ 2,216
Subtract	$ 2,216
Multiply	$ 2,216
Grade 3	
Swahili	
Paragraph	$ 2,725
Story	$ 2,726
Comprehension	$ 2,725
English	
Paragraph	$ 2,725
Story	$ 2,725
Comprehension	$ 2,725
Math	
Add	$ 2,044
Subtract	$ 2,044
Multiply	$ 2,044
Divide	24 $ 2,044

B Power calculations that take into account the previous RCT treatment status

If we think that the first two years and the second two years have no "interaction" effects, then we calculate that the treatment for each group would be

Table 9: Treatment effects

		KiuFunza II		
		Levels	Gains	Control
KiuFunza I	C1	$\alpha_0 + \alpha_1 + \alpha_3$	$\alpha_0 + \alpha_3 + \alpha_2$	$\alpha_0 + \alpha_3$
	C2	$\alpha_0 + \alpha_1 + \alpha_4$	$\alpha_0 + \alpha_4 + \alpha_2$	$\alpha_0 + \alpha_4$
	C3	$\alpha_0 + \alpha_1$	$\alpha_0 + \alpha_2$	α_0

where

$$Y_i = \alpha_0 + \alpha_1 T^2_{Levels,i} + \alpha_2 T^2_{Gains,i} + \alpha_3 T^1_{C1,i} + \alpha_4 T^1_{C2,i} + \varepsilon_{i,g}$$

In matrix notation we would have

$$Y_i = X\beta + \varepsilon_i$$

where

$$X = \begin{pmatrix} 1 & T^2_{COD,1} & T^2_{Gains,1} & T^1_{C1,1} & T^1_{C2,i} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T^2_{COD,n} & T^2_{Gains,n} & T^1_{C1,n} & T^1_{C2,n} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}$$

And therefore the variance-covariance matrix of β would be

$$V(\hat{\beta}) = (X'\Omega^{-1}X)^{-1}$$

where $\Omega = Diag(\Omega_g)$ and

$$\Omega_g = \begin{pmatrix} \sigma^2_\alpha + \sigma^2_\varepsilon & \sigma^2_\alpha & \cdots & \sigma^2_\alpha \\ \sigma^2_\alpha & \sigma^2_\alpha + \sigma^2_\varepsilon & \cdots & \sigma^2_\alpha \\ \vdots & \ddots & \ddots & \vdots \\ \sigma^2_\alpha & \cdots & \sigma^2_\alpha & \sigma^2_\alpha + \sigma^2_\varepsilon \end{pmatrix}$$

$$
\begin{aligned}
V(\hat{\beta}) &= (X'(I_g \otimes \Omega_g)^{-1}X)^{-1} \\
&= \left((X_1', X_2', ..., X_J') \begin{pmatrix} \Omega_g^{-1} & 0 & ... & 0 \\ 0 & \Omega_g^{-1} & ... & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & ... & \Omega_g^{-1} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_J \end{pmatrix} \right)^{-1} \\
&= \left(\sum_{j=1}^{J} X_j' \Omega_g^{-1} X_j \right)^{-1}
\end{aligned}
$$

Now it is important to note that there are 9 types of schools (the subindex used to indicate each type is shown below). Let $A^k = X_j^{k'} \Omega_g^{-1} X_j^k$ for schools of type k.

Table 10: Add caption

		KiuFunza II		
		COD	Gains	Control
KiuFunza I	C1	1	2	3
	C2	4	5	6
	C3	7	8	9

Then we have that:

$$
A^1 = X_j^{1'} \Omega_g^{-1} X_j^1 = \begin{pmatrix} n_g\Sigma & n_g\Sigma & 0 & n_g\Sigma & 0 \\ n_g\Sigma & n_g\Sigma & 0 & n_g\Sigma & 0 \\ 0 & 0 & 0 & 0 & 0 \\ n_g\Sigma & n_g\Sigma & 0 & n_g\Sigma & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}
$$

$$
A^2 = X_j^{2'} \Omega_g^{-1} X_j^2 = \begin{pmatrix} n_g\Sigma & 0 & n_g\Sigma & n_g\Sigma & 0 \\ 0 & 0 & 0 & 0 & 0 \\ n_g\Sigma & 0 & n_g\Sigma & n_g\Sigma & 0 \\ n_g\Sigma & 0 & n_g\Sigma & n_g\Sigma & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}
$$

$$
A^3 = X_j^{3'} \Omega_g^{-1} X_j^3 = \begin{pmatrix} n_g\Sigma & 0 & 0 & n_g\Sigma & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ n_g\Sigma & 0 & 0 & n_g\Sigma & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}
$$

$$
A^4 = X_j^{4'} \Omega_g^{-1} X_j^4 = \begin{pmatrix} n_g\Sigma & n_g\Sigma & 0 & 0 & n_g\Sigma \\ n_g\Sigma & n_g\Sigma & 0 & 0 & n_g\Sigma \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ n_g\Sigma & n_g\Sigma & 0 & 0 & n_g\Sigma \end{pmatrix}
$$

$$
A^5 = X_j^{5'} \Omega_g^{-1} X_j^5 = \begin{pmatrix} n_g\Sigma & 0 & n_g\Sigma & 0 & n_g\Sigma \\ 0 & 0 & 0 & 0 & 0 \\ n_g\Sigma & 0 & n_g\Sigma & 0 & n_g\Sigma \\ 0 & 0 & 0 & 0 & 0 \\ n_g\Sigma & 0 & n_g\Sigma & 0 & n_g\Sigma \end{pmatrix}
$$

$$A^6 = X_j^{6'}\Omega_g^{-1}X_j^6 = \begin{pmatrix} n_g\Sigma & 0 & 0 & 0 & n_g\Sigma \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ n_g\Sigma & 0 & 0 & 0 & n_g\Sigma \end{pmatrix}$$

$$A^7 = X_j^{7'}\Omega_g^{-1}X_j^7 = \begin{pmatrix} n_g\Sigma & n_g\Sigma & 0 & 0 & 0 \\ n_g\Sigma & n_g\Sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A^8 = X_j^{8'}\Omega_g^{-1}X_j^8 = \begin{pmatrix} n_g\Sigma & 0 & n_g\Sigma & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ n_g\Sigma & 0 & n_g\Sigma & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A^9 = X_j^{9'}\Omega_g^{-1}X_j^9 = \begin{pmatrix} n_g\Sigma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where Σ is the sum of any given row of the inverse of Ω_g and n_g is the number of students per school we test. Therefore we have:

$$\begin{aligned} V(\hat{\beta}) &= \left(\sum_{j=1}^{J} X_j'\Omega_g^{-1}X_j\right)^{-1} \\ &= \begin{pmatrix} Jn_g\Sigma & 60n_g\Sigma & 60n_g\Sigma & 70n_g\Sigma & 70n_g\Sigma \\ 60n_g\Sigma & 60n_g\Sigma & 0 & 40n_g\Sigma & 10n_g\Sigma \\ 60n_g\Sigma & 0 & 60n_g\Sigma & 20n_g\Sigma & 30n_g\Sigma \\ 70n_g\Sigma & 40n_g\Sigma & 20n_g\Sigma & 70n_g\Sigma & 0 \\ 70 & 10n_g\Sigma & 30n_g\Sigma & 0 & 70n_g\Sigma \end{pmatrix}^{-1} \\ &= \frac{1}{10n_g\Sigma}\begin{pmatrix} 18 & 6 & 6 & 7 & 7 \\ 6 & 6 & 0 & 4 & 1 \\ 6 & 0 & 6 & 2 & 3 \\ 7 & 4 & 2 & 7 & 0 \\ 7 & 1 & 3 & 0 & 7 \end{pmatrix}^{-1} \\ &= \frac{1}{18840n_g\Sigma}\begin{pmatrix} 604 & -280 & -252 & -372 & -456 \\ -280 & 763 & 357 & -258 & 18 \\ -252 & 357 & 651 & -138 & -78 \\ -372 & -258 & -138 & 828 & 468 \\ -456 & 18 & -78 & 468 & 756 \end{pmatrix} \end{aligned}$$

It is easy to show that

$$\Omega_g^{-1} = \begin{pmatrix} \sigma_\alpha^2+\sigma_\varepsilon^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2+\sigma_\varepsilon^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 & \sigma_\alpha^2+\sigma_\varepsilon^2 \end{pmatrix}^{-1} \tag{2}$$

$$= \left(\sigma_\varepsilon^2(I_{n_g} - \frac{J_{n_g}}{n_g}) + (n_g\sigma_\alpha^2+\sigma_\varepsilon^2)\frac{J_{n_g}}{n_g}\right)^{-1} \tag{3}$$

$$\tag{4}$$

where I_{n_g} is the identity matrix of size n_g and J_{n_g} is an $n_g \times n_g$ matrix with the unit in all its entries. Therefore we have:

$$\Omega_g^{-1} = \sigma_\varepsilon^{-2}(I_{n_g} - \frac{J_{n_g}}{n_g}) + \frac{1}{n_g\sigma_\alpha^2+\sigma_\varepsilon^2}\frac{J_{n_g}}{n_g} \tag{5}$$

$$\tag{6}$$

and therefore

$$\Sigma = \frac{1}{n_g\sigma_\alpha^2+\sigma_\varepsilon^2} \tag{7}$$

Thus,

$$V(\hat{\beta}) = \frac{n_g\sigma_\alpha^2+\sigma_\varepsilon^2}{18840n_g}\begin{pmatrix} 604 & -280 & -252 & -372 & -456 \\ -280 & 763 & 357 & -258 & 18 \\ -252 & 357 & 651 & -138 & -78 \\ -372 & -258 & -138 & 828 & 468 \\ -456 & 18 & -78 & 468 & 756 \end{pmatrix}$$

Now without controls, we have that $\sigma_\varepsilon^2 = 1$ and an ICC of $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\varepsilon^2} = 0.3$. In that case we have:

Table 11: Power Calculations

Size	Power	MDE Levels	MDE Gains	MD Difference
0.05	0.8	0.34	0.31	0.33
0.1	0.8	0.43	0.40	0.42
0.05	0.9	0.40	0.37	0.38
0.1	0.9	0.49	0.46	0.47

However, we have control variables that explain about 30% of the variation in test scores and reduce the ICC. That is, $\sigma_\varepsilon^2 = 0.7$ and the ICC is $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\varepsilon^2} = 0.1$. Therefore we have:

Table 12: Power Calculations

Size	Power	MDE COD	MDE Gains	MD Difference
0.05	0.8	0.16	0.15	0.15
0.1	0.8	0.20	0.20	0.19
0.05	0.9	0.19	0.17	0.18
0.1	0.9	0.23	0.21	0.22