KiuFunza: Pre-Analysis Plan for 2nd Year Results*

February 9, 2015

1 Introduction

This documents presents the pre-analysis plan for the second year's data of the KiuFunza project - which is a randomized evaluation of the impact of input and incentive based approaches to improving early grade learning outcomes, as well as the interaction of the two approaches. The project is being implemented and overseen by Twaweza - a non-profit initiative in Tanzania, and the evaluation is being led by Innovations for Poverty Action (IPA), with the lead researchers being Prof. Isaac Mbiti and Prof. Karthik Muralidharan.

This document is being prepared and registered after the first year of data has been collected and analyzed (to generate hypotheses) but before the second year of data have been analyzed. 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 project features 3 treatment arms and a control group and is implemented across a representative sample of 350 schools across 10 districts in Tanzania. The treatment arms are:

1. A capitation grant (CG) to schools that provides them with block grants (the "input" treatment)
2. A "cash on delivery" (CoD) treatment to schools that provides teachers and head teachers with bonus payments conditional on the number of students who pass basic literacy and numeracy tests (the "incentive" treatment)
3. A combination (Combo) treatment arm where schools were provided with both the CG and the CoD treatments.

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

Each of the three treatments was assigned to 70 randomly-selected schools (7 in each of 10 districts) and an additional 140 schools served as a control group. The project is strongly influenced by the design and findings of the Andhra Pradesh Randomized Evaluation Studies (see Muralidharan and Sundararaman (2011)), but has two critical differences.

First, this is to the best of our knowledge, the first experimental study that is explicitly designed (and adequately powered) to test for complementarities between inputs and incentives in improving education outcomes in developing countries. While cross-cutting designs have been widely used in education research before, to the best of our knowledge, the sample sizes in these studies have rarely been adequately powered to be able to detect interactions over and above the main effects. In practice, these cross-cutting designs have been employed mainly to reduce costs of evaluation (by treating schools with more than one treatment as having , and the evaluations have typically assumed away the role of interactions (see Duflo, Dupas, and Kremer (2011, 2012) for instance). Thus, the KiuFunza project is unique in allocating its sample size explicitly to test for the existence of complementarities between input and incentive based approaches to improving early grade learning outcomes.

Second, while the teacher incentives studied in Andhra Pradesh (see Muralidharan and Sundararaman (2011)) were designed in a sophisticated way (based on teacher-level value addition), such a design can be challenging to implement at scale in the settings of weak administrative capacity that are typical in 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. The KiuFunza project therefore compromised on the 'ideal' design of the incentive program and instead chose a design that was more 'implementable' at scale. 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 improvements. The expectation of the project implementing agency (Twaweza) was that the simplicity of such a scheme would make it easy to understand and be more effective at motivating teachers and principals than a more complex (but 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)). Thus, a key objective of the research is to test for such heterogeneity.

This document is outlined as follows: Section 2 presents details on the interventions, and the sampling procedures. Section 3 presents the hypotheses that will be tested at the end of 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 Overview

2.1 Treatment Selection

The evaluation is being implemented in 350 government primary schools in 10 districts in Tanzania between 2013 and 2014. All interventions were implemented directly by Twaweza and its District Partners, with money given to schools and teacher through the CG and COD interventions coming also from Twaweza. Within each intervention, information describing the intervention was distributed to schools and the communities via school and community meetings in early 2013. The District Partners then followed up with additional school visits in July and August to answer any questions regarding the program. All students in Grades 1, 2, and 3 in schools that received Cash on Delivery were tested in Kiswahili, English and Math at the end of the school year to determine teacher incentive payments. Tanzanian education professionals, following a similar structure as the Uwezo annual learning assessment, developed the subject tests for Grades 1, 2, and 3. The same schedule will be followed in 2014.

2.1.1 Intervention 1 - Making Capitation Grants Flow (CG)

This intervention is implemented in 70 schools (7 schools per district in 10 districts). The capitation grants are provided by Twaweza in two disbursements per year (at TZS 5,000 - ≈ $ USD 3 - each per pupil) on set dates towards the beginning of each school term (May and July). After obtaining cooperation and information from the district councils, funds will be transferred by electronic transfer directly into already established school bank accounts. Schools may only use the funds, consistent with present policy, for improving school quality via purchasing books, examinations, etc. but not for salaries or major construction. Twaweza did not establish any special systems for the planning, use, accounting and reporting of these funds but insist that schools are transparent and inform communities of how the money is being spent, consistent with government disclosure policy. Accounting for these funds is therefore conducted per existing government policy and mechanisms.

The evaluation will seek to measure the extent to which the funds reach schools, the level of citizen engagement on the use of funds, and ultimately the impact of funds and information on improving learning outcomes. In 2013, the average CG distributed to schools was 7,646,429 Tsh (≈ 4500 USD).

2.1.2 Intervention 2 - Local Cash on Delivery (COD)

This intervention is implemented in 70 schools (7 schools per district in 10 districts). For every child in Grades 1, 2 and 3 who passes the literacy (English and Kiswahili) and numeracy (Mathematics) assessment[1] at the end of the school year, the child's teacher will be paid TZS 5,000 (≈ $ USD 3

[1]Developed in the style of the proven Uwezo literacy and numeracy assessment.

) per subject the child passes (or up to TZS. 15,000 (≈ $ USD 9) per each child who is able to pass all three literacy and numeracy tests). Note this pays for absolute levels of learning, not gains in learning and that teachers are not penalized for students who do not pass. Additionally, the head teacher will be paid TZS 1,000 ((≈ $ USD 0.6) per subject each child passes. This incentive offer will be publicized at the beginning of the school year (around March), followed up through in person visits and/or phone calls, and children will be independently assessed towards the end of the school year. The intervention will be conducted with all pupils in chosen schools in Grade 1, 2, and 3.

We are aware that the incentive structure could lead to perverse outcomes as teachers might focus on students that are on or around the passing threshold (marginal students), and neglect students at the ends of the distribution of baseline knowledge (or ability), which could lead to negative effects on the latter group. An incentive structure that rewards gains at all points in the student achievement distribution would minimize the risk of this perverse outcome (Muralidharan & Sundararaman, 2011). However, having a incentive structure that rewards on levels and not on gains is much more simple administratively and so our results might be more informative to policy makers thinking about teacher performance payment programs.

2.1.3 Intervention 3 - Capitation Grants and Cash on Delivery

This intervention, which provides both capitation grants and local cash on delivery payments, is implemented in 70 schools (7 schools per district in 10 districts). The COD approach is designed to be "in addition to" or "on top of" existing programs and budgets, and in effect create an incentive to make better use of those resources. But if existing resources are significantly inadequate or not disbursed, particularly at the school level, schools and teachers may simply be unable to take the actions necessary to achieve results for which they will be rewarded later. In other words, a COD intervention can only reasonably be expected to work after one has ensured basic inputs have been provided for.

This idea was emphasized by several officials from the Ministry of Education and Vocational Training in a meeting between the Education Minister and CGD President, and in a workshop organized by the Ministry of Education and Embassy of Sweden in Dar es Salaam in April 2012. Thus the results from this intervention will inform on this issue.

2.2 Sample Selection

To be able to finance and manage the project, and so as to test the effectiveness of the idea before proposing that it go to scale, the interventions and study were implemented in 10 districts. The

sample was chosen randomly from the complete list of districts in Tanzania[2], with probability of district selection proportional to the number of primary school students. That is, districts with a higher number of primary school students had a higher chance of being in the sample. The selected sample includes the following 10 districts: Karagwe, Geita, Kahama, Kondoa, Korogwe Rural, Lushoto, Sumbawanga Rural, Mbozi, Mbinga and Kinondoni.

All government primary schools were eligible in each of the 10 districts, but 35 schools were randomly selected from each district to be part of the evaluation. The probability that a school was chosen was proportional to the number of students enrolled in the school. Furthermore, we stratified by school size and then randomized treatment within school size strata. In each sample district, 14 schools were randomly assigned to the control group and 7 schools were randomly assigned to each treatment group: Capitation Grant (CG), Cash on Delivery (COD) and Combination (Receiving both CG and COD). However, whenever schools were located next to each other (less than 1 km apart), we assigned the same treatment group to both schools.

3 Data

3.1 Survey Data

Data collection is carried out by Economic Development Initiatives (EDI), a well-established, Kagera-based, survey firm. Data is to be collected six times, three times during each school year (at the beginning, the middle and the end of the year). All information from the first year has already been collected. Detailed information is gather 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 and gender) is collected for a randomly selected sample of 30 students per school (10 students from Grades 1, 2, and 3).

Finally, household information (e.g. parents engagement in child's education, parents own education, household composition, and assets owned by the household) is collected for a random sample of students' households. In 2013 a total of 10 students were sampled per school for household interviews were, five from each second and third grade in the first year. In 2014 10 households were surveyed at the beginning of the school year (five from each second and third grade) and 15 at the end of the schools year (the same 10 households as at the beginning of the school year plus five more households from grade 1).

It is important to note there are two sets of tests performed to measure student learning levels. Twaweza tests all students in grades 1, 2 and 3 in COD and Combination schools to calculate the

[2]The list of districts and government primary schools within each district was provided by Tanzanian Prime Minister's Office - Regional Administration and Local Governments (PMO-RALG)

teacher payments and tests all Grades 1, 2, and 3 students in 40 randomly selected control schools (4 per district). Additionally, EDI tests 30 students in all schools which allows us to compare treatment effects for all treatments compared to control schools, in a low stakes exam. The Twaweza test is used to calculate the incentive payments, but the impact evaluation is done using the EDI test.

Table 1 presents a summary of the sample. Each year Twaweza tested all students in grades 1, 2 and 3 in COD and Combination schools and EDI tested 30 students in all schools. Additionally, household information was collected for 10 households in 2013 and will be collected for 15 households in 2014 for a total of 20 households (as 5 households are surveyed in 2013 and 2014). In total we have information for 7,000 households (20 per school) and 14,000 students (40 per school) in 350 schools. Notice that we observe 1,750 household and 7,000 students for two years, while the rest we only observe for one year[3].

Table 1: Sampled Design

Treatment	Control Schools				CG Schools				COD/Combination			
Grade in 2014	1	2	3	4	1	2	3	4	1	2	3	4
Panel A: 2013												
Twaweza Test[a]	No	Yes[f]	Yes[f]	Yes[f]	No	No	No	No	No	Yes	Yes	Yes
EDI Test[b]	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Household Survey[c]	0	0	5	5	0	0	5	5	0	0	5	5
Panel B: 2014												
Twaweza Test[a]	Yes[f]	Yes[f]	Yes[f]	No	No	No	No	No	Yes	Yes	Yes	No
EDI Test[b]	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Household Survey 1[d]	0	5	5	0	0	5	5	0	0	5	5	0
Household Survey 2[e]	5	5	5	0	5	5	5	0	5	5	5	0

[a] Done for all students
[b] Done for a random sample of 10 students per grade
[c] Beginning and end of school year
[d] Beginning of school year
[e] End of school year
[f] Done for a random sample of 40 schools

3.2 Administrative Data

We will use the results of the national Grade 4 and 7 examinations to assess the effect of our treatment on other grades.

4 Hypotheses

The hypotheses we present mainly test whether our treatment had any impact on learning outcomes and tries to get at the mechanisms behind the effects, if any. Specifically, our hypotheses are:

Main Outcomes

[3] We do not observe students enrolled in fourth grade in the second year; instead observe a new wave of students in first grade. In 2014, we do not re-survey households from students who were enrolled in grade 3 in 2013, and, in 2014, we only survey households from students currently enrolled in grade 1 at the end of the school year.

1. Impact of effectively delivering a Capitation Grant (CG) of 10,000 Tsh per student, as per goverment policy, to schools on students' learning
    - H_a (H_0): CG Treatment has (no) positive impact on test scores
2. Impact of incentivizing teachers, through our Cash on Delivery (COD) program, on student's learning outcomes
    - H_a (H_0): COD Treatment has (no) positive impact on test scores
3. The impact of the interaction of providing a Capitation Grant (CG) to the school and incentivizing teachers through COD on students' learning.
    - H_a (H_0): The interaction between COD and CG has (no) positive impact on test scores
4. Impact of CG, COD and their interaction on non-incentivized subjects learning outcomes. This magnitude and direction of this outcome depends on whether there is any substitution of teaching time and inputs away from non-incentive subjects and whether there are any complementarities on learning (e.g. higher language skills allow the students to read textbooks in other subjects).
    - H_a (H_0): CG, COD and their interaction have (no) impact on test scores of non-focal subjects

Channels

5. Impact of CG, COD and their interaction on teacher's behavior.
    (a) Impact on teacher's behavior in focal subjects in focal grades
        - H_a (H_0): Treatment - CG, COD or their interaction - has (no) impact on teacher's behavior (number of tests, remedial teaching, tutoring, homework assignments, time spend grading homework, time spend at school, time spend giving extra classes, attendance, teacher's likelihood to use new techniques, track students or use alternative resources to improve teaching) in focal subjects in focal years.
    (b) Impact on teacher's behavior in focal subjects in non-focal grades. It is possible that teachers substitute teaching time away from non-incentivized grades into incentivized grades. The same could be true at the school level for teaching inputs.
        - H_a (H_0): Treatment - CG, COD or their interaction - has (no) impact on teacher's behavior (number of tests, remedial teaching, tutoring, homework assignments, time spend grading homework, time spend at school, time spend giving extra classes, attendance, teacher's likelihood to use new techniques, track students or use alternative resources to improve teaching) in focal subjects in non-focal years.

(c) Impact on teacher's behavior in non-focal subjects.

- H_a (H_0): Treatment - CG, COD or their interaction - has (no) impact on teacher's behavior (number of tests, remedial teaching, tutoring, homework assignments, time spend grading homework, time spend at school, time spend giving extra classes, attendance, teacher's likelihood to use new techniques, track students or use alternative resources to improve teaching) in non-focal subjects.

6. Impact on school expenditure

- Impact on expenditure. This analysis will also be performed by grade, as it is possible that CG money is invest in 7th grade, when students take a high stakes exam.
  - H_a (H_0): Treatment - CG, COD or their interaction - has (no) positive impact on text book and teaching input expenditures
- Impact on how CG funds are spent. This analysis will also be performed by grade, to see if CG resources are spent in focal grades.
  - H_a (H_0): The interaction between CG and COD has (no) impact on how CG resources are spent.

7. Impact on school's schedule

- H_a (H_0): Treatment - CG, COD or their interaction - has (no) impact on the number of hours taught in different subjects.
- H_a (H_0): Treatment - CG, COD or their interaction - has (no) impact on teacher assignments across subjects and grades.

8. Impact on household's expenditure in education

- H_a (H_0): Treatment - CG, COD or their interaction - has (no) impact on household's expenditure in education (both through student supplies expenditure and through donations in money or kind to schools).

9. Impact on household's engagement in education

- H_a (H_0): Treatment - CG, COD or their interaction - has (no) impact on household's engagement in child's education.

Heterogeneity

10. Impact on learning outcomes according to baseline knowledge of the student.

- H_a (H_0): Treatment - CG, COD or their interaction - impact on knowledge is different (the same) for all students.

The next section presents a detail methodology of how we are going to test are hypotheses.

5 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, schools and household (baseline) characteristics. Table 2 presents the characteristics we control for and the corresponding question in the survey questionnaires from which they are taken.

Table 2: Control Variables

	Questionnaire	Question
Panel A: Student		
Student has seen exercise books with Uwezo tests	Student	DETAILS Q.6
Student attended pre-school (nursery) before attending elementery school	Student	DETAILS Q.5
Age	Student	DETAILS Q.3
Gender	Student	DETAILS Q.4
Ranking	Student	Collected during the second year, for the first and second year
Baseline kiswahili score	Uwezo Test	Kiswahili set
Baseline english score	Uwezo Test	English set
Baseline math score	Uwezo Test	Hisabati set
Panel B: School		
School has a kitchen	School	B4B.Q1
School has a library	School	B4B.Q2
School has a playground	School	B4B.Q3
School has a staff-room	School	B4B.Q4
School has a outer wall or fence	School	B4B.Q5
School receives newspaper	School	B4B.Q6
Travel time (in minutes) to closest bank	School	B4A.Q2
Travel time (in minutes) to closest post office	School	B4A.Q2
Travel time (in minutes) to closest high school	School	B4A.Q2
Travel time (in minutes) to closets government primary school	School	B4A.Q2
School has computers	School	B4.Q39
School has electricity	School	B4.Q13
School holds classes outside besides physical education	School	B4.Q11
School has piped water	School	B4.Q17
School has no water	School	B4.Q17
School has a single shift	School	B8.Q5
Number of toilets/latrines per student	School	B4.Q19/(B7.Q3+B7.Q4)
Number of classroom per student	School	B4.Q3/(B7.Q3+B7.Q4)
Number of teachers per student	School	B3.Q.2/(B7.Q3+B7.Q4)
School provides breakfast	School	B8.Q1
School has kinder-garden	School	B7.Q1
School tracks students	School	B8.Q7
School is located in an urban area	School	B4.Q1
Total number of students in school	School	B7.Q3+B7.Q4
Total number of students in the grade of the student	School	B7.Q3+B7.Q4
School keeps records of expenditures	School	B6B.Q1
Quality of school records	School	B6B.Q2
Schools has a public notice board with current expenditures	School	B6B.Q4
Size of the school committee	School	B2.Q1
Proportion of the members of the school committee that are female	School	B2.Q2/B2.Q1
Proportion of the members of the school committee that are teachers	School	B2.Q4/B2.Q1
Proportion of the members of the school committee that are parents	School	B2.Q5/B2.Q1
Number of times school committee met in previous year	School	B2.Q13
Gender of the head teacher	Teacher	B.Q4
Year in which head teacher was born	Teacher	B.Q5
Year in which head teacher started teaching	Teacher	B.Q7
Year in which head teacher started teaching at that school	Teacher	B.Q8
Previous experience in a private school (of head teacher)	Teacher	B.Q24
Travel time to school (of head teacher)	Teacher	D.Q9
Salary (of head teacher)	Teacher	B.Q11
Whether the head teacher has post-secondary education	Teacher	D.Q1
Panel C: Average Teacher Characteristics		
Proportion of teachers that are male	Teacher	B.Q4
Average year in which teachers at the school are born	Teacher	B.Q5
Average year in which teachers at the school started teaching	Teacher	B.Q7
Average year in which teachers at the school started teaching at that school	Teacher	B.Q8
Proportion of teachers with experience in private schools	Teacher	B.Q24
Average travel time to school	Teacher	D.Q9
Average salary	Teacher	B.Q11
Proportion of teachers with post-secondary school education	Teacher	D.Q1
Panel D: Teacher Characteristics by grade and subject		
Gender of the teacher	Teacher	B.Q4
Year in which teacher was born	Teacher	B.Q5
Year in which teacher started teaching	Teacher	B.Q7
Year in which teacher started teaching at that school	Teacher	B.Q8
Previous experience in a private school	Teacher	B.Q24
Travel time to school	Teacher	D.Q9
Salary	Teacher	B.Q11
Whether the teacher has post-secondary education	Teacher	D.Q1
Whether the teacher is the Head (or deputy head) Teacher	Teacher	D.Q9
Dummy variables for combinations of focal subjects with the same teacher	Teacher	Current grades and subjects Q.1
Panel E: Household Characteristics		
Household size	Household	HH ROSTER Label
Expenditure in education in 2012	Household	Expenditure Q.1
Expenditure in education in 2013 (baseline)	Household	Expenditure Q.2
Household owns a working radio	Household	HOUSE CHARACTERISTICS Q.9
Household owns a working tv	Household	HOUSE CHARACTERISTICS Q.10
Household owns a working bicycle	Household	HOUSE CHARACTERISTICS Q.11
Household owns a working car	Household	HOUSE CHARACTERISTICS Q.12
Household owns a working motorbike	Household	HOUSE CHARACTERISTICS Q.13
Household owns a working refrigerator	Household	HOUSE CHARACTERISTICS Q.14
Household owns a working watch	Household	HOUSE CHARACTERISTICS Q.15
Household owns a working mobile phone	Household	HOUSE CHARACTERISTICS Q.16
Main material of outer walls is mud or earth	Household	HOUSE CHARACTERISTICS Q.1
Main material of floors is mud or earth	Household	HOUSE CHARACTERISTICS Q.2
Roof made out of a durable material	Household	HOUSE CHARACTERISTICS Q.3
Parents Education	Household	EDUCATION Q.4
No. of rooms/No. Household members	Household	HH ROSTER Q.3
Household head had a job last week	Household	HOUSE CHARACTERISTICS Q.21
Improved water source	Household	HOUSE CHARACTERISTICS Q.4
Electric energy	Household	HOUSE CHARACTERISTICS Q.5
Improved sanitation facilities	Household	HOUSE CHARACTERISTICS Q.7
Attended any school meetings previous year	Household	PARENT INVOLVEMENT Q.2
Donated to school	Household	PARENT INVOLVEMENT Q.15
Child eats breakfast before class	Household	LEARNING ENVIRONMENT Q.19
Household head can read and understand kisawhili	Household	READING & MATH Q.1
Household head can read and understand english	Household	READING & MATH Q.2
Household head math ability	Household	READING & MATH Q.3 + Q.4

Note: All questionnaires are from the baseline survey performed in 2013 except for the incoming students in 2014 (entering first grade) for which the questionnaires correspond to baseline of 2014.

5.1 Effect on test scores: H1, H2, H3 and H4

To estimate the effect on test scores (and test hypotheses 1, 2, 3 and 4) we estimate the following equation

$$Z_{igsdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_s\beta_2 + X_h\beta_3 + X_{gs}\beta_4 + \varepsilon_{igsdt},$$

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, γ_d is a set of district fixed effects, γ_w is a set of week fixed effects[4], γ_g is a set of grade fixed effects, X_i is a series of student characteristics (see panel A in table 2), X_s is a set of school and average teacher characteristics (see panel B and C in table 2), X_{gs} is a set of teacher characteristics (for a particular grade/subject, see panel D in table 2), and X_h is a set of household characteristics[5] (see panel E in table 2). The coefficients of interest are the α's, which test hypotheses 1-4 above. We will analyze each subject separately, as well as the combination of all of them.

Specifically, we have:

- H.1
  - H_0: CG Treatment has no positive impact on test scores (i.e. $\alpha_1 \leq 0$ when Z_{igsdt} is the score in a focal subject)
  - H_a: CG Treatment has a positive impact on test scores (i.e. $\alpha_1 > 0$ when Z_{igsdt} is the score in a focal subject)
- H.2
  - H_0: COD Treatment has no positive impact on test scores of focal subjects (i.e. $\alpha_2 \leq 0$ when Z_{igsdt} is the score in a focal subject)
  - H_a: COD Treatment has a positive impact on test scores of focal subjects (i.e. $\alpha_2 > 0$ when Z_{igsdt} is the score in a focal subject)
- H.3
  - H_0: The interaction between COD and CG has no positive impact on test scores of focal subjects (i.e. $\alpha_3 \leq 0$ when Z_{igsdt} is the score in a focal subject)
  - H_a: The interaction between COD and CG has a positive impact on test scores of focal subjects (i.e. $\alpha_3 > 0$ when Z_{igsdt} is the score in a focal subject)

[4]The EDI test was performed before the Twaweza test, but the timing is balanced across treatment arms. The week fixed effects should the increase the precision of our estimates.

[5]These are not included in the main specification as we only have data for a subsample of households.

- H.4
  - H_0: CG and COD (and their interaction) have no impact on test scores of non-focal subjects (i.e. $\alpha_1 = \alpha_2 = \alpha_3 = 0$ when Z_{igsdt} is the score in a non focal subject)
  - H_a: CG or COD (or their interaction) have an impact on test scores of non-focal subjects (i.e. $\alpha_1 \neq 0$ or $\alpha_2 \neq 0$ or $\alpha_3 \neq 0$ when Z_{igsdt} is the score in a non focal subject)

To construct the standardized test scores, Z_{igsdt}, in each grade-subject we normalize using the mean and dividing by the standard deviation of the test scores in the control group. Once we have subject-grade standardized test scores, we will add these up across grades and the re-normalize (dividing by the standard deviation of the test scores in the control group); this will yield subject standardized test scores for a subject. We specifically collect some information on non-incentivized subjects to test hypothesis H4. The "Test Booklet-EDI version 09.06" is the test done at the end of 2013 school year and "EDI EL2014 Darasa" is the test that will be done at the end of 2014 school year.

5.2 Effect on teachers: H5

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

$$Y_{gsd} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \gamma_d + X_i\beta_1 + X_s\beta_2 + \varepsilon_{igsdt},$$

where Y_{igsd} is the outcome variable that measures behavior of teacher i in school s in district d, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, γ_d is a set of district fixed effects, X_s is a set of school characteristics (see panel B and C in table 2) and X_i is a set of teacher characteristics (see panel D in table 2). The coefficients of interest are the α's which test hypothesis 5 above. The outcome variables that we will focus on are presented in table 3 with the respective question in the surveys used to measure them.

Table 3: Teacher outcomes

	Hypothesis	Questionnaire	Question
Test in focal subjects in focal grades	H5	Teacher	3A.Q.2
Test in non-focal subjects in focal grades	H5	Teacher	3A.Q.2
Test in focal subjects in non-focal grades	H5	Teacher	3A.Q.2
Test in non-focal subjects in non-focal grades	H5	Teacher	3A.Q.2
Homework in focal subjects in focal grades	H5	Teacher	3A.Q.3
Homework in non-focal subjects in focal grades	H5	Teacher	3A.Q.3
Homework in focal subjects in non-focal grades	H5	Teacher	3A.Q.3
Homework in non-focal subjects in non-focal grades	H5	Teacher	3A.Q.3
Tutoring in focal subjects in focal grades	H5	Teacher	3A.Q.6 & 3A2.Q.1-3A2.Q.7
Tutoring in non-focal subjects in focal grades	H5	Teacher	3A.Q.6 & 3A2.Q.1-3A2.Q.7
Tutoring in focal subjects in non-focal grades	H5	Teacher	3A.Q.6 & 3A2.Q.1-3A2.Q.7
Tutoring in non-focal subjects in non-focal grades	H5	Teacher	3A.Q.6 & 3A2.Q.1-3A2.Q.7
Offers remedial in focal subjects in focal grades	H5	Teacher	3A.Q.7 & 3A3.Q.1-3A2.Q.8
Offers remedial in non-focal subjects in focal grades	H5	Teacher	3A.Q.7 & 3A3.Q.1-3A2.Q.8
Offers remedial in focal subjects in non-focal grades	H5	Teacher	3A.Q.7 & 3A3.Q.1-3A2.Q.8
Offers remedial in non-focal subjects in non-focal grades	H5	Teacher	3A.Q.7 & 3A3.Q.1-3A2.Q.8
Time grading homework	H5	Teacher	Daily activities roster Q.3
Time grading test	H5	Teacher	Daily activities roster Q.3
Time extra classes	H5	Teacher	Daily activities roster Q.3
Time in school	H5	Teacher	3.5 Time Use Q.1-Q.2
Start using new teaching technique this year	H5	Teacher	D.Q.21
Split students according to ability	H5	Teacher	D.Q.24
Use resources to improve teaching	H5	Teacher	P.Q.9
Above average teaching inputs	H5	Teacher	O.Q.1
Above average help from head teacher	H5	Teacher	O.Q.4 and J
Above average help from other teachers	H5	Teacher	O.Q.5 and A2
Attendance	H5	Teacher	O.Q.5 and A2
Attendance	H5	Teacher	3.7 Q.1 -Q.9

Note: All questionnaires are from the endline survey performed in 2014.

Specifically, we have:

- H.5
  - H_0: Treatment has no impact on teacher behavior (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)
  - H_a: Treatment has an impact on teacher behavior (i.e. $\alpha_i \neq 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)

5.3 Effect on school: H6 and H7

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

$$Y_{sdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \gamma_d + X_s\beta_1 + \varepsilon_{sdt},$$

where Y_{sdt} is the outcome variable that measures behavior of school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, γ_d is a set of district fixed effects, X_s are a set of school characteristics (see panel in table C and D 2).The coefficients of

interest are the α's which test hypotheses 7 and 8 above. The outcome variables that we will focus on are presented in table 4 with the respective question in the surveys used to measure them.

Additionally, to test hypothesis 7b, we will test whether the teacher characteristics (age, gender, education, and experience) change in focal grade/subjects.

Table 4: School outcomes

	Hypothesis	Questionnaire	Question
Administrative expenses per student	H6	School	School Expenses Q.1
Student expenses per student	H6	School	School Expenses Q.1
Teaching aid expenses per student	H6	School	School Expenses Q.1
Teacher expenses per student	H6	School	School Expenses Q.1
Construction expenses per student	H6	School	School Expenses Q.1
Student/teacher ratio	H6	Teacher	TEACHER ROSTER Q.2
Volunteer/teacher ratio	H6	School	VOLUNTEERS Q.1
Student/teacher ratio per grade	H6	Teacher	TEACHING GRADES Q.1
Textbook expenditure per student	H6	School	TEXTBOOK AND PRACTICE EXAMS Q.1
Textbook expenditure per student per grade	H6	School	TEXTBOOK AND PRACTICE EXAMS Q.1
Enrollment per grade	H6	School	(Y2 Baseline) 4.1 GRADES Q.1 & 6.3 ENROLLMENT
Time spend per subject per week	H7a	School	TIME SPENT ON SUBJECTS Q.1-Q.10
Grades taught	H7b	Teacher	GRADES Q.1
Reason to change teaching assignment	H7b	Teacher	2.1A EXIT Q.1-Q5
Subjects taught	H7b	SUBJECTS Q.1	
Reason to change teaching assignment	H7b	Teacher	2.1B1 EXIT/ENTRY Q.1- Q.7

Note: All questionnaires are from the endline survey performed in 2014.

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 COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)
  - H_a: Treatment has no impact on text book and teaching input expenditure (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)
- H.6.b
  - H_0: CG resources are not used more in focal subjects when COD is given as well (i.e. $\alpha_3 = 0$) when we restrict ourselves to focal grades data.
  - H_a: CG resources are used more in focal subjects when COD is given as well (i.e. $\alpha_3 \neq 0$) when we restrict ourselves to focal grades data.
- H.7
  - H_0: Treatment does not increase the amount of hours taught in incentivized subjects. (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)

- H_a: Treatment increase the amount of hours taught in incentivized subjects. (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)

5.4 Effect on households: H8 and H9

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

$$Y_{igsdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \gamma_d + \gamma_g + X_i\beta_1 + X_s\beta_2 + \varepsilon_{igsdt},$$

where Y_{igsdt} is the outcome variable that measures behavior of household i, which has a child in grade g at school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, γ_d is a set of district fixed effects, γ_g is a set of grade fixed effects, X_i are a series of household characteristics (see panel E in table 2), X_s are a set of school characteristics (see panel B and C in table 2). The coefficients of interest are the α's which test hypothesis 9 above. The outcome variables that we will focus on are presented in table 5 with the respective question in the surveys used to measure them.

Table 5: Household outcomes

	Hypothesis	Questionnaire	Question
Expenditure in education	H8	Household	Expenditure Q1
Donations to school	H8	Household	FOCUS CHILD TEACHERS AND PARENT INVOLVEMENT Q.10
Attend school meetings	H9	Household	FOCUS CHILD TEACHERS AND PARENT INVOLVEMENT Q.2-Q3
Meet teachers	H9	Household	FOCUS CHILD TEACHERS AND PARENT INVOLVEMENT Q.6-Q7
Adult at home when child gets home	H9	Household	LEARNING ENVIRONMENT AND FOCUS CHILD PERFORMANCE Q.1
Tutoring for child	H8	Household	LEARNING ENVIRONMENT AND FOCUS CHILD PERFORMANCE Q.10
Child eat breakfast before school	H9	Household	LEARNING ENVIRONMENT AND FOCUS CHILD PERFORMANCE Q.18

Note: All questionnaires are from the endline survey performed in 2014.

- H.8
  - H_0: Treatment does not change household expenditure in education (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)
  - H_a: Treatment changes household expenditure in education (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)
- H.9
  - H_0: Treatment does not change household engagement in child's education (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)

– H_a: Treatment changes household engagement in child's education (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)

5.5 Heterogeneous treatment effects: H10

In order to test hypothesis 12 we do a locally weighted regression of the end line test scores on the baseline score of students. Specifically, we estimate the following equation

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

where F is the CDF of the baseline scores of students. Let $f(x;T) = \alpha_0^T + \alpha_1^T z$ denote the estimated relation between baseline score and endline score for treatment T using the command *lowess* in STATA. The pointwise 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 perform a semi-parametric test where we split the data by students baseline and test hypothesis 1-3 in the sub-samples. We split students by baseline ability as follows: Students above the passing threshold (those who passed the threshold for teachers COD payments at baseline), students below the threshold (those who did not score a single question right and therefore require the most investment in order to the get the teacher a COD payment), and students around the passing threshold (those who got some questions right at baseline but did not enough to get the teacher a COD payment). One consequence of the incentive design in which teachers are rewarded based on absolute levels of learning, is that the incentives to focus on students near the threshold are larger and therefore we would expect the effect to be larger near the threshold (and more significant). Specifically, for the different sub samples we estimate the following equation

$$Z_{igsdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \gamma_z Z_{isd,t=0} + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_s\beta_2 + X_h\beta_3 + \varepsilon_{igsdt},$$

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, 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 series of student characteristics (see panel A in table 2), X_s is a set of school and teacher characteristics (see panel B and C in table 2), and X_h is a set of household characteristics[6] (see panel E in table 2). Then our hypothesis are

- H.12.A

[6]These are not included in the main specification as we only have data for a subsample of households.

- H_0: Treatment has no impact on test scores for student below the threshold (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)
- H_a: Treatment has an impact on test scores for student below the threshold (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)

- H.12.B
  - H_0: Treatment has no impact on test scores for student above the threshold (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)
  - H_a: Treatment has an impact on test scores for student above the threshold (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)

- H.12.C
  - H_0: Treatment has no impact on test scores for student around the threshold (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)
  - H_a: Treatment has an impact on test scores for student around the threshold (i.e. $\alpha_i = 0$, $i = 1$ for COD treatment, $i = 2$ for CG treatment and $i = 3$ for the interaction between COD and CG)

5.6 Other tests

5.6.1 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 COD and Combination schools to calculate the teacher payments and tests all Grades 1, 2, and 3 students in 40 randomly selected control schools (4 per district). Additionally, EDI tests 30 students in all schools 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 (as EDI test have a wider range of questions).

5.6.2 Effect on non-incentivized grades scores

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

$$Y_{gsdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \gamma_z Y_{gsd,t-1} + \gamma_d + X_s\beta_1 + \varepsilon_{igsdt},$$

where Y_{igsdt} is a measure of learning for grade g at school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, γ_d is a set of district fixed effects, X_s is a set of school characteristics (see panel B and C in table 2. The coefficients of interest are the α's, which test hypotheses 5 above. For Y_{igsdt} we use the average score and the pass rate in the national Grade 4 and 7 examinations.

Specifically, we have:

- – H_0: CG and COD (and their interaction) have no impact on test scores (i.e. $\alpha_2 = \alpha_3 = \alpha_4 = 0$)
  – H_a: CG or COD (or their interaction) have an impact on test scores (i.e. $\alpha_2 \neq 0$ or $\alpha_3 \neq 0$ or $\alpha_4 \neq 0$)

5.6.3 Capitation grant and funding substitution

It could be possible that the capitation grant causes a substitution from other sources of funding, leaving the total amount of funds available to the school unchaged. To asses this we estimate the following equation

$$Y_{sdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \gamma_d + X_s\beta_1 + \varepsilon_{sdt},$$

where Y_{sdt} is the outcome variable that measures funding from other sources for school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, γ_d is a set of district fixed effects, X_s are a set of school characteristics (see panel in table B and C 2).The coefficients of interest are the α's. The outcome variables that we will focus on are measures of funding from other sources and are based on questions: Funding Q.1-Q.2, Funding details Q.1-Q.16 and IN-KIND Q.1-Q.2 from the school questionnaire.

5.6.4 Heterogeneous treatment effects by student characteristics

To estimate heterogeneous treatment effects by student characteristics we perform the following regression $Z_{igsdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \lambda_0 C_i + \lambda_1 COD_s \times C_i + \lambda_2 CG_s \times$

$C_i + \lambda_3 CG_s \times COD_s \times C_i + \gamma_d + \gamma_w + \gamma_g + X_s\beta_1 + X_h\beta_2 + \varepsilon_{igsdt}$, where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, 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_s is a set of school and teacher characteristics (see panel B and C in table 2), and X_h is a set of household characteristics[7] (see panel E in table 2). Finally C_i is a student characteristic (grade, gender and age as in panel A of table 2). The coefficients of interest are the λ's, which test if there are any heterogeneous treatment effects by student characteristics. We also test for heterogeneity by students ranking (within the school).

5.6.5 Heterogeneous treatment effect by school characteristics

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

$Z_{igsdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \lambda_0 C_s + \lambda_1 COD_s \times C_s + \lambda_2 CG_s \times C_s + \lambda_3 CG_s \times COD_s \times C_s + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_p\beta_2 + X_h\beta_3 + \varepsilon_{igsdt}$,

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, 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 2), X_p is a set of teacher characteristics (see panel B and C in table 2), and X_h is a set of household characteristics[8] (see panel E in table 2). Finally C_s is a school characteristic: An index between 0 and 6 of school facilities, the average proximity to other facilities, whether the school has piped water, whether the school has a single shift or not, the size of the school committee, the number of times the school committee met in 2012, the proportion of females, teachers and parents in the school committee, and whether the school keeps records of their 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 2. The coefficients of interest are the λ's, 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 (and school committee). 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 and 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 as before.

[7] These are not included in the main specification as we only have data for a subsample of households.
[8] These are not included in the main specification as we only have data for a subsample of households.

5.6.6 Heterogeneous treatment effects by teacher characteristics

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

$Z_{igsdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \lambda_0 C_p + \lambda_1 COD_s \times C_p + \lambda_2 CG_s \times C_p + \lambda_3 CG_s \times COD_s \times C_p + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_p\beta_2 + X_h\beta_3 + \varepsilon_{igsdt}$,

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, 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 2), X_s is a set of school characteristics (see panel B and C in table 2), and X_h is a set of household characteristics[9] (see panel E in table 2). 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 2. Aditionally, we will test heterogeneity by teacher's schedule.[10] The coefficients of interest are the λ's, 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.

5.6.7 Heterogeneous treatment effects by household characteristics

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

$Z_{igsdt} = \alpha_0 + \alpha_1 COD_s + \alpha_2 CG_s + \alpha_3 CG_s \times COD_s + \lambda_0 C_h + \lambda_1 COD_s \times C_h + \lambda_2 CG_s \times C_h + \lambda_3 CG_s \times COD_s \times C_h + \gamma_d + \gamma_w + \gamma_g + X_i\beta_1 + X_s\beta_2 + \varepsilon_{igsdt}$,

where Z_{igsdt} is the test score of student i in grade g at school s in district d at time t, COD is a dummy variable that indicates whether the school received cash on delivery or not, CG is an indicator variable of whether the school received a capitation grant, 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 2), and X_s is a set of school and teacher characteristics (see panel B and C in table 2). Finally C_h is a household characteristic for student i: an asset level index between 0 and 8, dwelling characteristics and parents education. See panel E in table 2. The coefficients of interest are the λ's, which test if there are any heterogeneous treatment effects by household characteristics.

[9] These are not included in the main specification as we only have data for a subsample of households.

[10] 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 hard then we may 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).

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 household socioeconomic status.

5.6.8 Teachers Learning

Something we would like to explore is "teacher learning". Specifically, after the first year they learn something about their students as well as their own teaching techniques. First, we would like to explore how internal ranking in schools compare to overall students' ability distribution, and see whether teachers with students that are "worst than they think" (for example, the best rank student is below the average) perform in the second year compare 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 kids 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

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, *101*(5), 1739-74. Retrieved from `http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.5.1739` doi: 10.1257/aer.101.5.1739

Duflo, E., Dupas, P., & Kremer, M. (2012, March). *School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from kenyan primary schools* (Working Paper No. 17939). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w17939` doi: 10.3386/w17939

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`

A Baseline Z-score

At the baseline not all students had to answer all questions, thus we have to input some values in order to create a standardized test score. Grade 1 students started in the easier level and stopped

when they couldn't pass a given level. We assume that they would have score zero in the levels they did not answer in this cases. For grades 2 and 3, in math, students started in the easier level and stopped when they couldn't pass a given level. We assume that they would have score zero in the levels they did not answer in this cases.

For grades 2 and 3, in Swahili and English, things are a little bit more complicated as students started in question 3. If the student passes the starting level, then he moved on to the next level (question 4) and continued to move levels until couldn't pass a given level. In this case, we assume they would have gotten zero in the questions they did not answer and a full score in the first two questions. If the student did not pass question 3, then he moved down to the previous level (question 2) and we assume he would have score zero in questions 4, 5 and 6. If he scored zero, then he was asked question 1. However, if he scored above zero, we do not observe his answer to question 1. In order to cope with the last case, we estimate a Poisson regression, using the sample of students in grade 1 that score more than zero in question 2, with the number of correct answers in question 1 in the left hand side and the number of correct answers in question 2, age, gender and district fixed effects in the right hand side. Using this model we estimate the number of questions a student in grade 2 and 3 would have score in question 1 when he scored more than zero in question 2.