When Student Incentives Don't Work: Evidence from a Field Experiment in Malawi

James Berry, Hyuncheol Bryant Kim, and Hyuk Son[*]

April 6, 2019

PRELIMINARY AND INCOMPLETE: PLEASE DO NOT CITE OR CIRCULATE

Abstract

Financial Incentives have been often proposed to enhance students' performance in school, but their impacts are theoretically ambiguous, and empirical evidence is mixed. "Tournament" incentives that reward only top performers could crowd out intrinsic motivation, and may affect a subgroup of students at the near the performance threshold. Through a field experiment among 5th to 8th graders in Malawi, we study impacts of two scholarship programs: a *Standard* scholarship program that rewarded top overall performers on an exam and a *Relative* scholarship program that rewarded the top performers within smaller groups of students with similar baseline scores. We find that the *Standard* scholarship program significantly decreased test scores and motivation to study, especially for those least likely to win the scholarship. By contrast, we find no evidence for test score impacts among those in the *Relative* scholarship program.

JEL Classifications: I21, O15

Keywords: student incentives, education policy, merit-based scholarship, feedback, field experiments

[*] Berry: University of Delaware, jimberry@udel.edu; Kim: Cornell University, hk788@cornell.edu; Son: Cornell University, hs924@cornell.edu. We are grateful to Won Bae and Jiwon Kim and staff members of Africa Future Foundation for their excellent field assistance in Malawi. Seminar participants at the American Economic Association Annual Meeting, the International Food Policy Research Institute, Yonsei University, New York University, Syracuse University, and NEUDC provided helpful comments and suggestions.

1 Introduction

In recent years, performance-based incentives for students have received increasing research attention as a means to improve learning outcomes in both developed and developing countries (Gneezy, Meier, and Rey-Biel (2011)). *Standard* economic theory predicts that financial incentives can induce student effort and thereby increase academic outcomes. On the other hand, a common argument against such incentives is that they may crowd out intrinsic motivation that may reverse positive impacts (Bénabou and Tirole (2006); Gneezy, Meier, and Rey-Biel (2011)). Empirical evidence on the effectiveness of performance-based incentives is largely mixed (Kremer, Miguel, and Thornton (2009); Angrist and Lavy (2009); Sharma (2010); Bettinger (2011); Fryer (2011); Levitt et al. (2012); Jackson (2010); Li et al. (2014)), with mixed impacts on intrinsic motivation as well (Visaria et al. (2016); Bettinger (2011)).[1] Understanding why incentive programs do and don't work remains an important open research area.

One particular incentive scheme that has received substantial research attention is an individual tournament scheme in which the top performing students on an exam are provided with a reward. Such schemes allow for the policy maker to set a fixed budget for the incentives, and have been generally shown to be incentive compatible to induce effort (Lazear and Rosen (1981)). However, tournament schemes, in which relatively few students receive the reward may induce effort only from top students. Indeed, several studies in developed countries find that effects of the programs were concentrated among those who were most likely to receive the reward (Angrist and Lavy (2009); Leuven, Oosterbeek, and van der Klaauw (2010); Bettinger (2011)). In the same vein, the bottom students who are unlikely to receive the reward may not be motivated to exert effort. These effects could could result in increased inequality in academic performance.

In addition, to the extent that student effort depends on the likelihood of obtaining the incentive, enhancing a student's information on his or her initial learning level or progress may enhance the distributional impacts of incentives. If students' responses to the financial incentives vary by initial level of academic achievement, these responses would depend on the perception of their initial position, rather than the actual position. Students may respond to information for reasons unrelated to external incentives: for example, such feedback may allow a student to better focus effort or may induce a sense of competition among students (Bandiera, Larcinese, and Rasul (2015); Tran and Zeckhauser (2012)).

In this paper, we study the impacts of two types of incentive programs, as well as performance feedback, on 5th to 8th graders at 119 classrooms of 31 primary schools in Malawi. The two

[1] Within the psychology literature, there is no clear consensus on effects of performance-based incentives on intrinsic motivation (Cameron and Pierce (1994); Deci, Koestner, and Ryan (1999)).

incentive programs, framed as scholarship schemes, provided rewards of MWK 4500 (USD 9.70) if the corresponding test score goal was met.[2] The first, which we call the *Standard* merit-based scholarship scheme, provided a scholarship to students in the sample who scored in the top 15 percent on the final end-of-year exam. This scholarship scheme is similar to that of Kremer, Miguel, and Thornton (2009), in which scholarships were given to the top 15 percent of 6th grade female students in a sample of schools in Kenya.

In the second scholarship scheme, the *Relative* merit-based scholarship, students were grouped into bins by baseline test score, and the top 15 percent of students within each bin received the incentive. Because students compete only with others that have similar baseline test scores, initially low-performing students are more likely to receive the rewards compared with a standard tournament. We hypothesized that this scheme would increase effort and reduce discouragement that may accompany the *Standard* scholarship. In addition, like a standard tournament incentive, the *Relative* scheme allows for a fixed incentive budget, as the number of students who obtain the incentive is known ex ante. The design was based on Barlevy and Neal (2012) who propose a similar scheme for teachers, which they call "pay for percentile."[3]

We interviewed all students in 5th to 8th grade at baseline for a short-term follow-up right before the final exam was administered. In addition, for students in 5th and 6th grade at baseline, we implemented a long-term follow-up survey and exam six months after the experiment was completed. This long-term follow-up survey and exam allow us to understand the impacts of and behavioral responses to the incentive for students after the incentives disappeared.

Our main finding is that the *Standard* scholarship scheme reduced final exam scores by 0.27 standard deviations across the full sample, with the largest negative impacts on students with the lowest initial test scores. The *Standard* scholarship scheme also reduced survey-measured motivation of the students, again with the results concentrated among the initially lowest-performing students. These results are consistent with arguments that financial incentives may crowd out intrinsic motivation (Gneezy, Meier, and Rey-Biel (2011)). By contrast, the relative merit-based scholarship scheme did not have significant impacts on test score performance or motivation.

This paper contributes to the existing literature along several dimensions. First, it contributes to the growing literature on financial incentives in education. Evidence on these programs is generally mixed, both in developing countries (Kremer, Miguel, and Thornton (2009); Sharma (2010); Behrman et al. (2015); Hirshleifer (2017)) and in developed countries (see Gneezy, Meier, and

[2] The exchange rate at the time of the study was 464 MWK: 1 USD.

[3] Our paper is, to our knowledge, the first test of the Barlevy and Neal (2012) "pay for percentile" scheme on students. Several papers evaluate this incentive structure for teachers.(Loyalka et al. (2016); Mbiti, Romero, Mauricio, and Schipper, Youdi (2018); Gilligan et al. (2018)). The structure is closely related to schemes that provide incentives based on improvement relative to baseline (Behrman et al. (2015); Berry (2015)).

Rey-Biel (2011) for a review).[4] The work closest to ours is that of Kremer, Miguel, and Thornton (2009), who study a merit scholarship program girls in Kenyan primary schools. In this program, scholarships were awarded to girls scoring in the top 15 percent of the endline exam. They find that the program increased test scores both for the targeted girls and for boys who were not eligible for the program. Our *Standard* incentive scheme is structured similarly, although it applied to all students. A key difference is that in our setting, that students are aware of their initial test score and percentile rank within the district. This has important implications on sustainability of the merit-based scholarship programs because, even though students may be unaware of their relative score initially, they would know if the scheme were repeated in a future period.

Although the types of incentive schemes vary across studies, most study a single incentive scheme. A smaller but growing literature evaluates the structure of incentive schemes by comparing multiple schemes within the same experiment. Studies have compared group and individual incentives (Li et al. (2014); Blimpo (2014)), incentives for effort and incentives for achievement (Hirshleifer (2017)), incentives targeted to parents and incentives targeted to children (Berry (2015)), and incentives for students and incentives for teachers (Behrman et al. (2015)). To our knowledge, our study is the first to compare incentives to top performers with incentives for relative performance.

Next, we contribute to the literature that studies how educational incentives influence motivation and other non-cognitive skills and behaviors. Although numerous studies within the psychology literature examine impacts of incentives on intrinsic motivation in controlled laboratory settings, there is no consensus on whether incentives do decrease motivation (Cameron and Pierce (1994); Deci, Koestner, and Ryan (1999)). Within the economics literature, evidence is also mixed. For example, in a study of U.S. middle students, Bettinger (2011) finds that incentives for exam performance did not decrease survey-based intrinsic motivation, while Visaria et al. (2016) find that incentives for attendance among primary students in India decreased intrinsic motivation.

Finally, our study is related to assessing the impact of feedback regarding students' relative performance on academic performance. Tran and Zeckhauser (2012) and Azmat and Iriberri (2010) find that providing rank information improves academic performance. By contrast, Ashraf, Bandiera, and Lee (2014) study the effects of providing relative rank information in a job training setting and show that rank information may lower exam performance by discouraging those at the bottom of the distribution.

The remainder of this paper is organized as follows. Section 2 provides a description of the context and scholarship schemes. Section 3 presents the estimating equations, and Section 4 presents the

[4] Within the developed-country literature, of particular note is Leuven, Oosterbeek, and van der Klaauw (2010) who study financial rewards given to Dutch University students for passing first-year requirements. Similar to our results, they find positive impacts for high-ability students and negative impacts on low-ability students.

results. We discuss the results and conclude in Section 5.

2 Context, Programs, and Experimental Design

2.1 Primary education in Malawi

The education system in Malawi is composed of eight years of primary education followed by four years of secondary education. Similar to other countries in Sub-Saharan Africa, the government of Malawi abolished primary school fees in the early 1990s, leading to near-universal primary enrollment. However, like many countries in the developing world, learning outcomes among Malawian primary students are low. Even among developing countries, Malawi lags behind. Among the 15 countries in Sub-Saharan Africa taking the Southern and Eastern Africa Consortium for Monitoring Education Quality standardized assessments, 6th graders in Malawi scored near the bottom in both reading and mathematics (SACMEQ, 2011). Schools are characterized by high pupil-teacher ratios and low levels of infrastructure.[5]

The academic calendar, starting on September, consists of three semesters. At the end of each semester, students in primary school take exams in six subjects: Chichewa (the vernacular language), English, mathematics, primary science, social studies, and art and life skills. Students typically must pay a fee about USD 0.5 to 1 to take the exam, to cover printing of exam copies. Passing the exams at the end of the third semester of each year is required for a student to proceed to the next grade. At the end of eighth grade, students take the Primary School Leaving Certificate Exam (PSLCE), a national-level exam for 8th graders, to obtain secondary school admission.

2.2 Program Descriptions and Experimental Design

The scholarship program was implemented in grades 5 to 8 at 119 classrooms of 31 public primary schools in TA Chimutu.[6] TA Chimutu is a rural area located about 15 km from the capital city of Lilongwe, and has three school districts. Each grade typically consists of one or two classes. The scholarship programs were implemented by Africa Future Foundation (AFF), an international NGO focused on health and education programs in Malawi and several other countries in Africa.

[5] For example, no school in our sample had electricity in the classrooms, and only 67% of students had their own desk and chair. The average pupil-teacher ratio was 85:1.

[6] TA stands for Traditional Authority and is the administrative division below the level of district.

2.2.1 Project chronology

The project chronology is summarized in Figure 1. The baseline survey and baseline exams were implemented during the first semester of the 2014-2015 academic year (December 2014 to January 2015). The results of the scholarship randomization were announced in the middle of the second semester (February 2015). The feedback intervention was based on the midterm exam, administered at the end of the second semester (March 2015). The follow-up survey was implemented shortly before the the final exam. Eligibility for the scholarships was based on the final exam, administered at the end of the third semester (June 2015). Eighth graders took the PSLCE, the national exam, in the third semester instead of the final exam. Awards were distributed in an area-wide awards ceremony that took place after the experiment was completed (October 2015). Finally, longer-term follow-up exams and surveys for baseline 5th and 6th graders were administered six months after the experiment was completed (March 2016).

2.2.2 Scholarship Programs

In February 2015, we stratified the 119 classrooms in the sample grades and randomly assigned classrooms into three groups: the *Standard* merit-based scholarship, the *Relative* merit-based scholarship, or the control group. Under the *Standard* merit-based scholarship scheme, within each grade, students scoring in the top 15 percent in the district on the final exam were eligible to receive the award.[7] Under the *Relative* merit-based scholarship scheme, students were grouped into bins of 100 by baseline test score, and the top 15 percent of each bin in the final exam were eligible to receive the award.[8]

The awards for *Standard* and *Relative* merit-based scholarships were identical. The award was a choice among a cash award of USD 9.70 (MWK 4,500) or an in-kind award including a pair of shoes, a school bag, or a school uniform of similar value.[9] This represents a significant amount considering that Malawi GDP per capita was only around USD 362.7 in 2014 (World Bank, 2019).

To ensure that students fully understood the scholarship programs and the conditions of winning the scholarships, AFF conducted a one-hour session to describe the program to students. Because the randomization was conducted within schools, all three treatment and control groups were explained to all students. During the announcement session, ach student was provided an individualized note describing his or her treatment assignment. Figure 2 provides examples of notes for each

[7] For 8th graders, eligibility was determined by PSLCE results.

[8] Students were eligible for the scholarship program if they took one of two baseline exams, administered in December of 2014 and January of 2015. A total of 8597, 89.7% of enrolled students, met this criterion.

[9] About 95 percent of eligible students chose the cash award.

treatment group, as well as the control group. For the *Standard* scholarship group, information on the student's overall rank as well as the scholarship eligibility condition (top 15 percent) were provided. For the *Relative* scholarship group, information on overall rank and rank within bin as well as the scholarship eligibility condition (top 15 percent within bin) were provided. For the control group, only information on the student's overall rank was provided.

At the end of the session, students were informed of their treatment and control assignments, and took a short a quiz to measure their understanding of the programs. The quiz, shown in Figure A1, contained 5 questions about hypothetical students who were assigned to one of the scholarship groups and whether they would receive the scholarship given their absolute or relative rank in the final exam. To measure understanding of the treatment assignments we asked students their perceived likelihood of receiving the scholarship after providing them with the individualized announcements.

With the exception of the eighth-grade PSLCE, exams used in this study were developed by a district-level exam committee to ensure uniformity across schools. The exam committee consisted of eight teachers, one vice-principal, and one principal (head teacher) of the schools within the district.[10] The exams were jointly administered by AFF and local primary education authorities. Additionally, AFF provided exam copies for the students during the study period, exempting them from exam fees.

2.2.3 Feedback intervention

The second intervention of the study was provision of feedback on the student's ranking as of the midterm exam. Across all three scholarship treatment groups, students in grades 5 to 7 were individually randomized into a "feedback" or "no-feedback" group.[11] In March of 2015, each student received a note providing their ranking as of the midterm exam privately in a separated place and encouraged not to share with their peers. Figure 4 presents examples of these notes. The feedback treatment group received information on their rank at the baseline and midterm exams (Panels 3a, 3c, and 3e), while the control group received an information only on the baseline exam (Panels 3b, 3d, and 3f). Feedback differed depending on the scholarship treatment group. In the *Standard* scholarship group, students in the feedback treatment received their overall rankings in the midterm exam relative to all students in the program. Students in the *Relative* scholarship group received information on their rankings in the midterm relative to students in their respective

[10] Prior to this study, each school created its own end-of-semester exams. AFF organized an exam committee under the supervision of the district education authority to form common questions for the whole district.

[11] Eighth graders were excluded from the feedback experiment because there was insufficient time between the feedback announcement and the the final PSLCE exam early in the third semester.

bins.

What is unique in our setting compared to the previous literature is that we are in an environment where feedback could potentially be more effective because it is directly linked to the scholarship eligibility. There is potential complementarity between feedback on relative performance and test score in a performance-based incentive setting if students are encouraged or discouraged when their test score is high or low. On the other hand, students in this study already had information on their previous academic performance through the scholarship announcement (Figure 2), which could make feedback effect less effective.

2.3 Data

We use several sources of data: AFF's administration data, district-level test score data (the baseline, midterm, final exam, and long-term follow-up exams), students' school attendance data, and student and parent surveys.

Our main source of data is student performance on district-level exams. Baseline exams were conducted twice, at the end of the first semester (December 2014) and beginning of the second semester (January 2015).[12] The midterm exam data used for the feedback intervention was implemented at the end of second semester in the 2014-2015 academic year. The final exam used to measure school achievement and select scholarship recipients was conducted at the end of third semester in the 2014-2015 academic year.[13] Lastly, the long-term follow-up exam was administered nine months after the experiment was completed (July 2016) by 5th and 6th graders. The main outcome variables are test scores and district level rank in these tests.

In addition to the exams, we measured students' school attendance through unannounced checks. These checks were conducted every month between April and June, four times before the scholarship announcement, and four times after.

To measure intermediate outcomes we also conducted surveys of students at the time of the baseline exams and right before the follow-up exams. A primary objective of the surveys was to measure non-cognitive skills and motivation. Self esteem is based on the Rosenberg self-esteem scale which measures both positive and negative feelings about oneself (Rosenberg (1965)). We also measured conscientiousness by questions based on Big Five Inventory scale (John and Srivastava

[12] Only 6728 (70.2 percent) students were able to take the first baseline exam due to the exam fee. AFF covered the exam fee in the second baseline exam, and thus 7945 (82.9 percent) students took the second baseline exam. The mean (and standard deviation) of the first and second exam is very similar: 11.5 (3.2) and 11.5 (3.4), respectively.

[13] For 8th graders who took PSLCE, instead of regular final exam, we were able to obtain letter grades for each subject, not raw test score. Score and rank for the reward were calculated based on following calculation. We treat A, B, C, D, and F as 6, 5, 4, 3, and 1, and standardize total scores.

(1999)).[14] Motivation was measured by asking how strongly the students agree to the statement "I am motivated to study hard" in a scale of five where one being strongly disagree and five being strongly agree.

In addition to these measures, the surveys collected students' reports on their own effort, as well as that of teachers and parents. Student effort was measured through reports of weekly study hours and attendance. To measure teachers' effort, students answered 21 questions on how the teachers encouraged students, and challenged them, and were responsive to participation. To measure parental effort, we elicited student reports of how much parents encourage, help, and ask students to study.

We collected a list of 9,419 enrolled students in the participating schools during the first semester. Among them, 7,638 students (81 percent) completed the baseline survey and 8,491 (90.1 percent) participated in the baseline exam. The final study sample consists of 7,386 students (78.4 percent) who participated in both the baseline survey and baseline exam.

Table 2 presents baseline characteristics and the checks of balance of the scholarship and feedback randomizations. Columns 1 and 2 display summary statistics of key variables for the whole sample and the control group, respectively. The average age is 14.2, and 47.3 percent of the sample are males. At the time of the baseline survey, the attendance rate of the students was 85 percent, and the average study hours per week was 16.1.

Columns 3 and 4 of Table 2 show the test of differences in means across scholarship treatment groups, and Column 6 presents the differences between the feedback and no-feedback groups. Overall, we observe few significant differences. Of the 16 variables examined, only one variable between the *Standard* scholarship and control group is significantly different at the 10% level. In the feedback randomization, four out of 16 are significantly different at the 10% level, but the differences are relatively small in magnitude. For example, the average grit score (out of 5) is 0.02 higher in the feedback group, a difference of 0.64 percent compared.

Table A6 displays sample attrition across treatment groups. On average 88, 83, and 90 percent of the study sample participated in the midterm exam, follow-up survey, and final exam, respectively. For the long-term follow-up survey and exam, 63 and 57 percent of the long-term study sample participated on average. We observe one statistically significant difference between the scholarship groups and the control group: students in the relative scholarship group are 3.2 percentage points more likely to take the final exam (significant at the 10 percent level). There are no significant differences in attrition between the feedback and no-feedback groups.

[14] Survey questions used to measure self-esteem, grit, and conscientiousness are shown in Appendix Figure A2

3 Estimating Equations

The randomized assignment of treatment groups allows for straightforward estimation of treatment effects. To estimate the average impacts of the *Standard* and *Relative* scholarship programs, we use the following equation:

$$Y_{ijgk1} = \beta_0 + \beta_1 Standard_{ij} + \beta_2 Relative_{ij} + Y_{ijgk0} + \eta_g + \gamma_k + \varepsilon_{ijgk} \tag{1}$$

where Y_{ijgk1} is the outcome of interest for student i in classroom j of grade g, and district k. *Standard* and *Relative* are indicators for being *Standard* and *Relative* scholarship groups, respectively. Y_{ijgk0} is the outcome of interest at baseline. η_g is a grade fixed effect and γ_k is district fixed effect. In some specifications, we include X_{ijgk}, a set of student-level controls, including age, race, household size, and a household asset index. Standard errors are clustered at the the classroom level, the level of randomization.

Because the distributional impact of the programs is a key research question, we present several methods of estimating heterogeneity by initial test score. First, we graph impacts across the baseline test score distribution. To examine heterogeneity using regression, we interact the treatment groups with an indicator for whether the student's baseline rank was in the top 15 percent. We select top 15% because students' responses to the *Standard* scholarship might differ based on whether they are above or below the cutoff for the scholarship eligibility. This implies the following regression:

$$\begin{aligned} Y_{ijgk} =& \beta_0 + \beta_1 Standard_{ij} + \beta_2 Relative_{ij} + \beta_3 Top15_{ijgk} \\ &+ \beta_4 Standard_{ij} * Top15_{ijgk} + \beta_5 Relative_{ij} * Top15_{ijgk} + \eta_g + \gamma_k + X_{ijgk} + \varepsilon_{ijgk} \end{aligned} \tag{2}$$

Where $Top15_{ijgk}$ is an indicator of being within top 15 percent in the baseline test. In these specifications, β_1 and β_2 represent the impacts of the *Standard* and *Relative* scholarships on the bottom 85 percent of students, and β_4 and β_5 capture the differences in the impacts of the *Standard* and *Relative* merit-based scholarship group between the top 15 and bottom 85 percent students. In addition to defining the top 15 percent based on the full baseline test score distribution, we run a similar regression interacting the treatment groups with an indicator for whether the student was in the top 15 percent within the narrower bins used in the *Relative* scholarship scheme.

Lastly, to analyze the impacts of feedback, we regress the outcome on inclusion in the feedback treatment group:

$$Y_{ijgk} = \beta_0 + \beta_1 Feedback_{ijg} + \eta_g + \gamma_k + X_{ijgk} + \varepsilon_{ijgk} \tag{3}$$

where *Feedback* indicates student *i* 's assignment to receive feedback. We also examine the impacts of feedback in each scholarship group by interacting Feedback with inclusion in each scholarship group:

$$\begin{aligned} Y_{ijgk} = &\beta_0 + \beta_1 Standard_{ij} + \beta_2 Relative_{ij} + \beta_3 Feedback_{ijg} + \beta_4 Standard_{ij} * Feedback_{ijg} \\ &+ \beta_5 Relative_{ij} * Feedback_{ijg} + \eta_g + \gamma_k + X_{ijgk} + \varepsilon_{ijgk} \end{aligned} \tag{4}$$

In these specifications, β_3 shows how feedback affect those in the control group. β_4 and β_5 capture whether feedback affects students assigned to the *Standard* and *Relative* scholarship group differently. We assess heterogeneity in treatment effects of feedback by running equation (4) separately for the top 15 percent and bottom 85 percent at baseline.

4 Results

4.1 Understanding of Program and Expectation of Scholarship

Before turning to the main impact results, we first discuss students' understanding of the program and expectation that they would receive the scholarship. As described in Section 2.2 above, students' understanding and expectations were elicited at the time of the program announcement, and again during the follow-up survey before the final exam.

Panel (a) of Figure A3 presents graphs of average scores on the test of understanding of the scholarship schemes (y-axis) by baseline rank (x-axis), by scholarship treatment group. Columns (1) and (2) of Table 3 presents the corresponding regressions. The results confirm that students understood the scholarship program quite well. For example, students answered 92 percent of questions correctly at the time of the program announcement, falling to about 64 percent as of the follow-up survey. Panel A of Table 3 shows that there are no significant differences in students' understanding between the treatment and control groups either right after the program announcement or right before the exam, with confidence intervals able to rule out differences above about five percentage points. There is some evidence of heterogeneity in understanding by baseline test score: while there is no significant difference between top 15 percent students and lower 85 percent students right after the announcement (Column 1), in the follow-up survey, the difference is about 8 percentage points (Column 2).

Panel (b) of Figure A3 displays students' expectations of winning the scholarship by baseline rank. Columns (3) and (4) of Table 3 display the corresponding regression results. For students in the *Standard* scholarship group, expectation of receiving the scholarship should increase with overall baseline rank; for students in the *Relative* scholarship group, expectations should not be related with baseline test score; and for students in the control group, expectations should be close zero. Panel (b) of Figure 4 generally confirms this pattern, particularly at the time of program announcement. Formal regression results in Columns (3) and (4) of Table 3 show that students in the scholarship groups were 29-35 percentage points more likely to expect the scholarship. Examining differences across baseline rank, those in the top 15 percent in the *Standard* merit-based scholarship group were significantly more likely to expect the scholarship 45 and 15 percentage points more than the control group after the announcement and 1st follow-up survey, respectively. It is worth noting that general understanding of the scholarship scheme decreased over time while expectation of winning the scholarship increased over time for all three groups.

Figure **??** shows expectation of winning the scholarship by distribution within bin where baseline subgroup rank (within bin) is on the *x*-axis. We expect no change in expectation with subgroup rank in the *Standard* scholarship group and no or minimal increase with subgroup rank in the *Relative* scholarship group. We also find expected patterns in which overall expectation of the *Standard* and *Relative* scholarships are similar or higher than that of the control group.

In sum, results in Figure A3 and Table 3 confirm students generally understood the scholarship scheme and had expectations consistent with their assigned groups.

4.2 Test Scores

We now turn to the impacts of the scholarship programs on test scores. Panel A of Table 4 presents the results of estimating Equation (1) on normalized test scores.[15] The *Standard* scholarship had substantial negative impacts on student performance: students performed 0.27 to 0.28 standard deviations worse than those in the control group (significant at the 10 percent level). The effects of the *Relative* scholarship were closer to zero and not statistically significant.

[15] For each outcome, we present two specifications with and without control variables, but the results are robust in various specifications.

Figure A4presents nonparametric plots of endline test scores in each treatment group by baseline rank. As shown in the figure, the negative impacts of the *Standard* scholarship are concentrated among those with the low baseline test scores, and the impacts turn positive for students above the 90th percentile of the baseline distribution. In contrast with the *Standard* scholarship, the impacts of the *Relative* scholarship were decreasing in test scores, with positive impacts at the bottom of the baseline test score distribution and negative impacts at the top of the distribution.

Panel B of Table 4 presents an additional analysis of heterogeneity by baseline rank by interacting the treatment with an indicator for being in the top 15 percent of baseline test scores, as per Equation (2). These results confirm that the decrease in academic achievement in the *Standard* treatment is driven by students with initial test scores in the bottom 85 percent: the coefficient on *Standard* merit scholarship is negative and significant, and that on the interaction between *Standard* merit scholarship and being in the top 15 percent at baseline is of opposite sign and more than half the magnitude, although it is not statistically significant. By contrast, the coefficient on the interaction of the *Relative* treatment and the top-15 dummy is negative, reflecting the negative impacts at the top of the test score distribution, although the coefficient is again not statistically significant.

We explore the heterogeneous impacts further by looking at the impact in each 10% bin at the baseline where those around the cutoff (between top 10% and 20%) are the reference group. The following linear regressions are estimated:

$$\begin{aligned} Y_{ijgk} =& \beta_0 + \beta_1 Standard_{ij} + \beta_2 Relative_{ij} + \sum_{l=1}^{10} \gamma 1_l Topl + \sum_{l=1}^{10} \gamma 2_l Standard_{ij} Topl \\ &+ \sum_{l=1}^{10} \gamma 3_l Relative_{ij} Topl + \eta_g + \gamma_k + X_{ijgk} + \varepsilon_{ijgk} \end{aligned} \tag{5}$$

Figure 6 presents ɣ2 and ɣ3 which present relative impacts of the *Standard* and *Relative* treatment for those in each bin compared to those at the cutoff. It also confirms that the negative impacts of the *Standard* scholarship are largest among those with the lowest baseline test scores (although the estimates are not statistically significant).

Finally, we examine whether the impacts vary by subgroup rank--that is, the ranking within the 100-student bins used in the *Relative* merit-based scholarship. The results are shown in Panel C of Table 4. The results show no statistically significant heterogeneity by subgroup rank, even in the *Relative* scholarship group.

4.3 Intermediate Outcomes

In this subsection we analyze intermediate outcomes in order to shed light on mechanisms for the test score results presented in the previous section. We start by analyzing responses of students including school attendance, time spent studying, motivation to study, self-esteem, and conscientiousness. These results are presented in Columns (1) to (5) of Table 5, with average impacts in Panel A, and heterogeneity by baseline rank in Panels B.

We find few impacts on observed and self-reported student effort. As shown in Column 1 of Table 5, there is a small increase in the attendance rate among the *Standard* scholarship group (Panel A), but we find no evidence for heterogeneity by baseline test score (Panel B). We find no statistically significant impacts on self-reported weekly study hours measured in the follow-up surveys (Column (2)), but point estimates suggest slightly less study effort in both scholarship treatment groups (Panel A), and slightly lower effort among students with the highest baseline scores in the treatment groups (Panel B).

Turning to impacts on non-cognitive measures including motivation to study, self esteem, and conscientiousness, we do find changes that generally correspond to the test score results (Columns (3) to (5) of Table 5). As shown in Panel A, the *Standard* scholarship program had negative impacts on all three measures on average, with statistically significant impacts on motivation and self esteem. The *Relative* scholarship program also had negative effects on average, although these impacts were smaller and not statistically significant. Turning to heterogeneity by baseline score, Panel B of Table 5 shows that the negative impacts of the *Standard* scholarship were concentrated among the bottom 85 percent of students. By contrast, there is no consistent pattern of heterogeneity in the *Relative* scholarship program.

Outcomes of the *Standard* scholarship treatment on non-cognitive measures correspond to the argument that financial incentives may crowd out intrinsic motivation. In our case, a merit-based scholarship program may discourage those who are unlikely to win the scholarship to study and negatively affect self-esteem and other non-cognitive measures. Results of the *Relative* scholarship treatment support this argument in that we do not find such negative impacts when the chance of winning the scholarship is similar across baseline test scores.

Columns (6) to (8) of Table 5 present impacts on students' perceptions of teacher and parental effort. We do not find evidence for changes in teacher effort as a result of either scholarship program. We do find that parents mentioned the scholarship program more often in the standard scholarship group, with effects concentrated among children with the highest baseline test scores. However, even though parents of the *Standard* scholarship group mentioned the opportunity more, it did not appear to translate into actual parental effort.

It is worth noting that a large portion of parents in our sample had little or no education and therefore may not have had the skills to effectively help their children at home.[16] A lack of capacity and resources may explain the null impacts parental effort. However, the results in Column (8) suggest that parents were aware of the program and discussed it with their children. The attendance results in Column (1) may therefore have been partially a result of parental encouragement to attend school.

4.4 Longer-term impacts

As discussed previously, the *Standard* scholarship program resulted in large negative impacts on the scores of the test that was incentivized as well as non-cognitive skills. In this section, we analyze impacts on test scores in the next semester, 9 months after the incentivised final exam, and show that these impacts did not persist after the incentives programs ended. As described in Section 2.3, second follow-up tests were conducted in the school year after the incentive programs took place, with students who were originally in grades 5 and 6. When presenting our longer-term follow-up results we also display short-term results for the grade 5 and 6 subsample to confirm that the results presented in the previous subsections hold for the sample that was followed into the next school year.

Table 6 displays the long-term results of the scholarship programs on test scores. As shown in Panel A, the negative effects of the *Standard* scholarship program have largely disappeared: the average long-term impacts (Columns (3)-(4)) much smaller in absolute value than the short-term impacts (Columns (1) and (2)) and are no longer statistically significant. As displayed in Panel B, point estimates reveal a similar pattern of heterogeneity by pretest score in the longer-term results compared with the short-term results, although the estimates are again smaller and not statistically significant. We do not find evidence on long-term impacts of the *Relative* scholarship program.

Table 7 presents corresponding short- and long-term results on attendance, self-reported student effort, and non-cognitive skills for 5th and 6th graders at the baseline. Even though there were negative effects on non-cognitive skills in the short-term (Columns 1-5), we do not find persistent changes in the long-term (Columns 6-9), which corresponds to the absence of long-term effects in the test scores.

[16] Only 54% of parents in our study sample graduated primary school.

4.5 Discussion

In the previous sections, we have shown that financial incentives may decrease students' test scores and negatively affect non-cognitive skills, particularly for those who are unlikely to win the reward. A natural question that arises is how much of the impacts on academic performance are driven by changes in non-cognitive skills. With the caveat that non-cognitive skills are endogenous, we provide suggestive evidence by controlling for measures of non-cognitive skills at the follow-up survey such as motivation to study, self-esteem, grit, and conscientious in our scholarship impact regressions. We find that test scores are explained at least partially by these control variables (Table A5): controlling for these variables reduces the impacts on test scores by about 11%.

The scholarship effects could have been driven by several other factors. First, introduction of a financial incentive based on relative performance might affect the classroom environment (even though it is district level competition, not within class). For example, students in the scholarship classrooms may have become more competitive as a result of the program and students may have been less likely to help each other to study. Our follow-up survey collected student reports of the classroom environment, allowing us to test for this possibility. As shown in Table A6, we do not find evidence that either scholarship group changed the classroom environment.

Second, the scholarship programs may have influenced cheating on the final exam. In particular, students in scholarship classrooms may have been less likely to cheat. If students in the scholarship classrooms prevented cheating with each other it may have explain the decrease in test scores. However, these arguments should apply to both the *Standard* and *Relative* scholarship programs, and thus they do not explain the fact that only the *Standard* program, not *Relative* program, decreased students achievement.

Another important point why there are no long-term effects, despite the relatively large short-term effects. This result is consistent with temporary decreases in effort and motivation while the incentives were in place, and these detrimental effects did not persist after the incentives.

4.6 Impacts of Feedback

Lastly, we study how feedback on students' midterm rankings influences student test performance and self-evaluated performance, and how these effects vary across scholarship groups. Panel (a) of Figure 7 plots final exam rank by midterm exam rank, for the randomly assigned feedback and no-feedback groups. As shown in Panel (a), those ranked in the top 15 percent in the feedback treatment group performed slightly better than those in the no-feedback group for the whole sample. Panel (b) repeats these plots for scholarship treatment and control groups. As shown in this panel, all three groups had similar patterns, with small positive impacts of feedback among those in the top 15 percent and limited impacts elsewhere. The impacts appear most pronounced for those in the *Standard* scholarship, although the differences between scholarship groups do not appear large.

These results are tested formally in Table 8, which presents estimates of Equation (3) and (4) for the full sample (Column (1) and (4)), as well as for the top 15 percent (Column (2) and (5)) and bottom 85 percent of students at baseline (Columns (3) and (6)). Across the full sample, there is no evidence of an impact of feedback in any of the three scholarship treatments and the control group. However, Column (2) of Panel A confirms the pattern in Panel (a) of Figure 10 by showing a modest positive impact of feedback among the top 15 percent of students, although the estimate is not statistically significant. We interpret the lack of impacts as suggesting that the feedback provided did not provide much new information to students, who were already told there baseline scores rank.

5 Conclusion

Financial incentives may not be successful in promoting educational achievement if such incentives have negative psychological effects. Through a randomized-controlled trial in Malawi, we study the impacts a *Standard* merit-based scholarship program that provided scholarships for students whose test scores were within the top 15 percent with a novel *Relative* scholarship scheme. The design of the *Relative* scholarship follows Barlevy and Neal (2012), in which students are grouped by baseline score, and incentives are awarded to the top 15 percent performers in each group. Using an additional randomized intervention, we study the impacts of feedback on student rank under these scholarship schemes, in which the results of a midterm exam were randomly provided to students in the middle of the study period.

We find that the *Standard* merit-based scholarship significantly decreased test scores compared to the control group, with the largest decreases concentrated among those least likely to win the scholarship. These decreases in test scores correspond to decreases in motivation to study among those least likely to win. However, we do not find such negative impacts among the *Relative* scholarship group. We find limited evidence that feedback on ranking may influences test scores, although point estimates suggest that it may increase test scores for initially high-performing students.

Our results suggest that tournament incentive schemes such as the standard scholarship that we study may exacerbate inequality in education outcomes. This may be especially pronounced in environments such as ours, in which students knew their baseline ranking from which they could gauge their chances of winning the scholarship. This may partially explain the differences between our results and those of Kremer, Miguel, and Thornton (2009) in which such information was not provided. This finding also corresponds to the literature that incentives may not work due to the psychological effects (Bénabou and Tirole (2006); Gneezy, Meier, and Rey-Biel (2011)). We further speculate that in contexts such as ours, with relatively few education inputs at home or in schools, students and their parents may have few resources to draw upon in order to improve achievement. This may induce discouragement and decrease effort.

References

Angrist, Joshua and Victor Lavy (2009). “The effects of high stakes high school achievement awards: Evidence from a randomized trial”. *The American Economic Review* 99.4, pp. 1384–1414.

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee (2014). “Awards unbundled: Evidence from a natural field experiment”. *Journal of Economic Behavior & Organization* 100, pp. 44–63.

Azmat, Ghazala and Nagore Iriberri (2010). “The importance of relative performance feedback information: Evidence from a natural experiment using high school students”. *Journal of Public Economics* 94.7, pp. 435–452.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul (2015). “Blissful ignorance? A natural experiment on the effect of feedback on students’ performance”. *Labour Economics*. European Association of Labour Economists 26th Annual Conference 34, pp. 13–25.

Barlevy, Gadi and Derek Neal (2012). “Pay for Percentile”. *American Economic Review* 102.5, pp. 1805–1831.

Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin (2015). “Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools”. *Journal of Political Economy* 123.2, pp. 325–364.

Bénabou, Roland and Jean Tirole (2006). “Incentives and Prosocial Behavior”. *American Economic Review* 96.5, pp. 1652–1678.

Berry, James (2015). “Child Control in Education Decisions An Evaluation of Targeted Incentives to Learn in India”. *Journal of Human Resources* 50.4, pp. 1051–1080.

Bettinger, Eric P. (2011). “Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores”. *Review of Economics and Statistics* 94.3, pp. 686–698.

Blimpo, Moussa P. (2014). “Team incentives for education in developing countries: A randomized field experiment in Benin”. *American Economic Journal: Applied Economics* 6.4, pp. 90–109.

Cameron, Judy and W. David Pierce (1994). “Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis”. *Review of Educational Research* 64.3, pp. 363–423.

Deci, Edward L., Richard Koestner, and Richard M. Ryan (1999). “A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation”. *Psychological Bulletin* 125.6, pp. 627–668.

Fryer, Roland G. (2011). “Financial Incentives and Student Achievement: Evidence from Randomized Trials”. *The Quarterly Journal of Economics* 126.4, pp. 1755–1798.

Gilligan, Daniel O, Naureen Karachiwalla, Ibrahim Kasirye, Adrienne M Lucas, and Derek Neal (2018). “Educator Incentives and Educational Triage in Rural Primary Schools”. Working Paper 24911. National Bureau of Economic Research.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel (2011). “When and Why Incentives (Don’t) Work to Modify Behavior”. *Journal of Economic Perspectives* 25.4, pp. 191–210.

Hirshleifer, Sarojini (2017). “Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance”. 201701. University of California at Riverside, Department of Economics.

Jackson, C. Kirabo (2010). “A little now for a lot later a look at a texas advanced placement incentive program”. *Journal of Human Resources* 45.3, pp. 591–639.

John, Oliver P. and Sanjay Srivastava (1999). “The Big Five trait taxonomy: History, measurement, and theoretical perspectives”. *Handbook of personality: Theory and research* 2.1999, pp. 102–138.

Kremer, Michael, Edward Miguel, and Rebecca Thornton (2009). “Incentives to Learn”. *Review of Economics and Statistics* 91.3, pp. 437–456.

Lazear, Edward P. and Sherwin Rosen (1981). “Rank-Order Tournaments as Optimum Labor Contracts”. *Journal of Political Economy* 89.5, pp. 841–864.

Leuven, Edwin, Hessel Oosterbeek, and Bas van der Klaauw (2010). “The Effect of Financial Rewards on Student Achievement: Evidence from a Randomized Experiment”. *Journal of the European Economic Association* 8.6, pp. 1243–1265.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff (2012). “The behavioralist goes to school: Leveraging behavioral economics to improve educational performance”. National Bureau of Economic Research.

Li, Tao, Li Han, Linxiu Zhang, and Scott Rozelle (2014). “Encouraging classroom peer interactions: Evidence from Chinese migrant schools”. *Journal of Public Economics* 111, pp. 29–45.

Loyalka, Prashant Kumar, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi (2016). “Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement”.

Mbiti, Isaac, Romero, Mauricio, and Schipper, Youdi (2018). “Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania”. Working Paper.

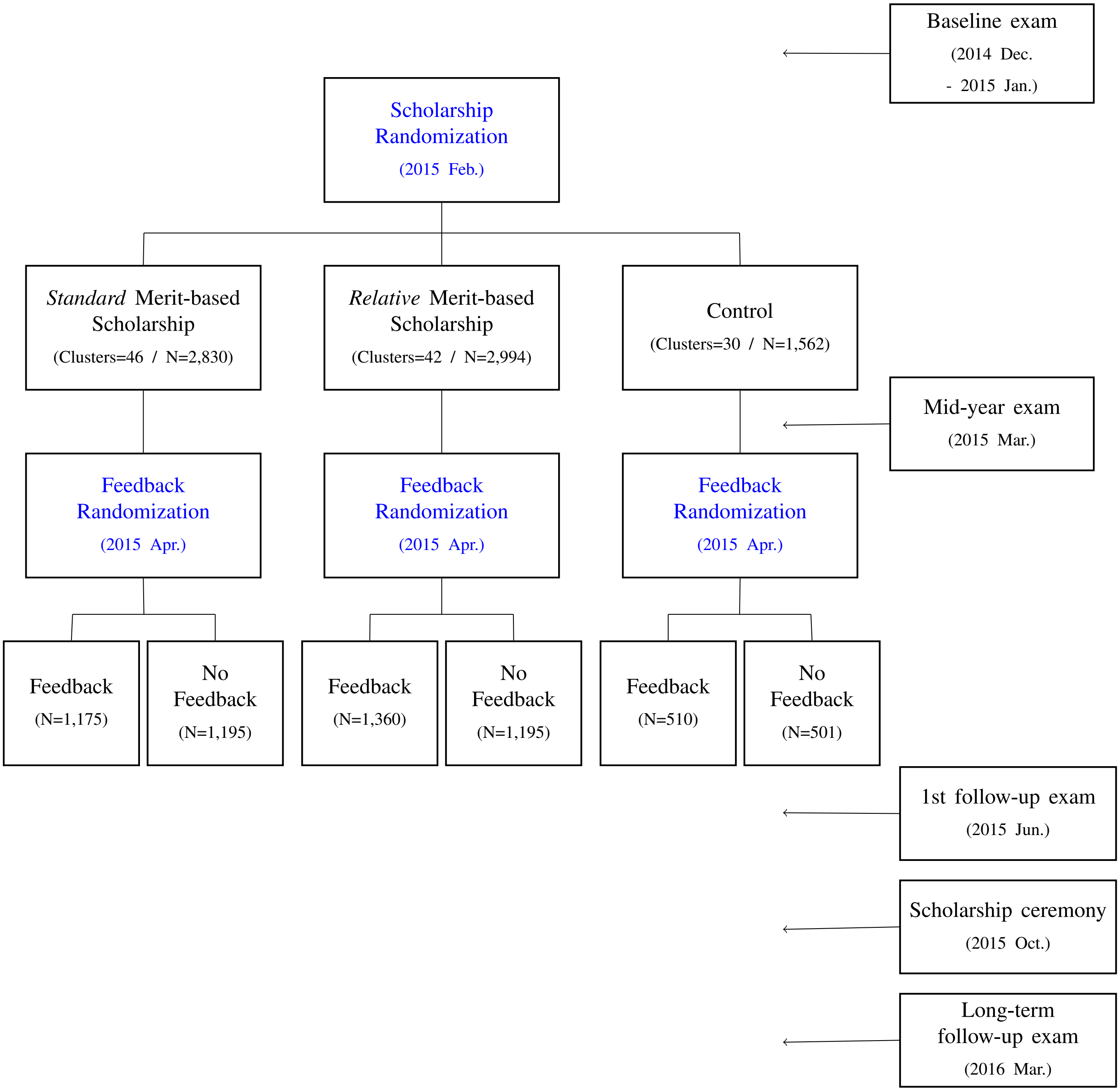
Rosenberg, Morris (1965). “Society and the Adolescent Self-Image.” *Science* 148.3671, pp. 804–804.

Sharma, Dhiraj (2010). “The impact of financial incentives on academic achievement and household behavior: Evidence from a randomized trial in Nepal”.

Tran, Anh and Richard Zeckhauser (2012). “Rank as an inherent incentive: Evidence from a field experiment”. *Journal of Public Economics* 96.9, pp. 645–650.

Visaria, Sujata, Rajeev Dehejia, Melody M. Chao, and Anirban Mukhopadhyay (2016). “Unintended consequences of rewards for student attendance: Results from a field experiment in Indian classrooms”. *Economics of Education Review* 54, pp. 173–184.

Figure 1: Experimental Design



Note: The experiment was implemented for 2014-2015 school year. School calendar year consists of three semester. Baseline, mid-term, and final exams are were administrated at the end of each semester. 8th graders took PSLCE, a national-level exam to obtain secondary school admission, instead of the final exam. Randomization was stratified at school-grade level, which we marked as clusters in the graph.

Figure 2: Scholarship Randomization result announcement note

(a) *Standard* merit-based scholarship group

ID	XXXXXXX	**School**	XXX
STD	7	**Name**	XXX
Group	A		

Current Position

25% [759 out of 1928]

You can receive a present when you are reanked at:

15%(455th) or above

(b) *Relative* merit-based scholarship group

ID	XXXXXXX	**School**	XXX
STD	5	**Name**	XXX
Group	B		

Current Position

75% [2286 out of 3037]

86% [86 out of 100 learners with similar score]

You can receive a present when you are reanked at:

15th or above among 100 learners of similar score

(c) Control group

ID	XXXXXXX	**School**	XXX
STD	6	**Name**	XXX
Group	C		

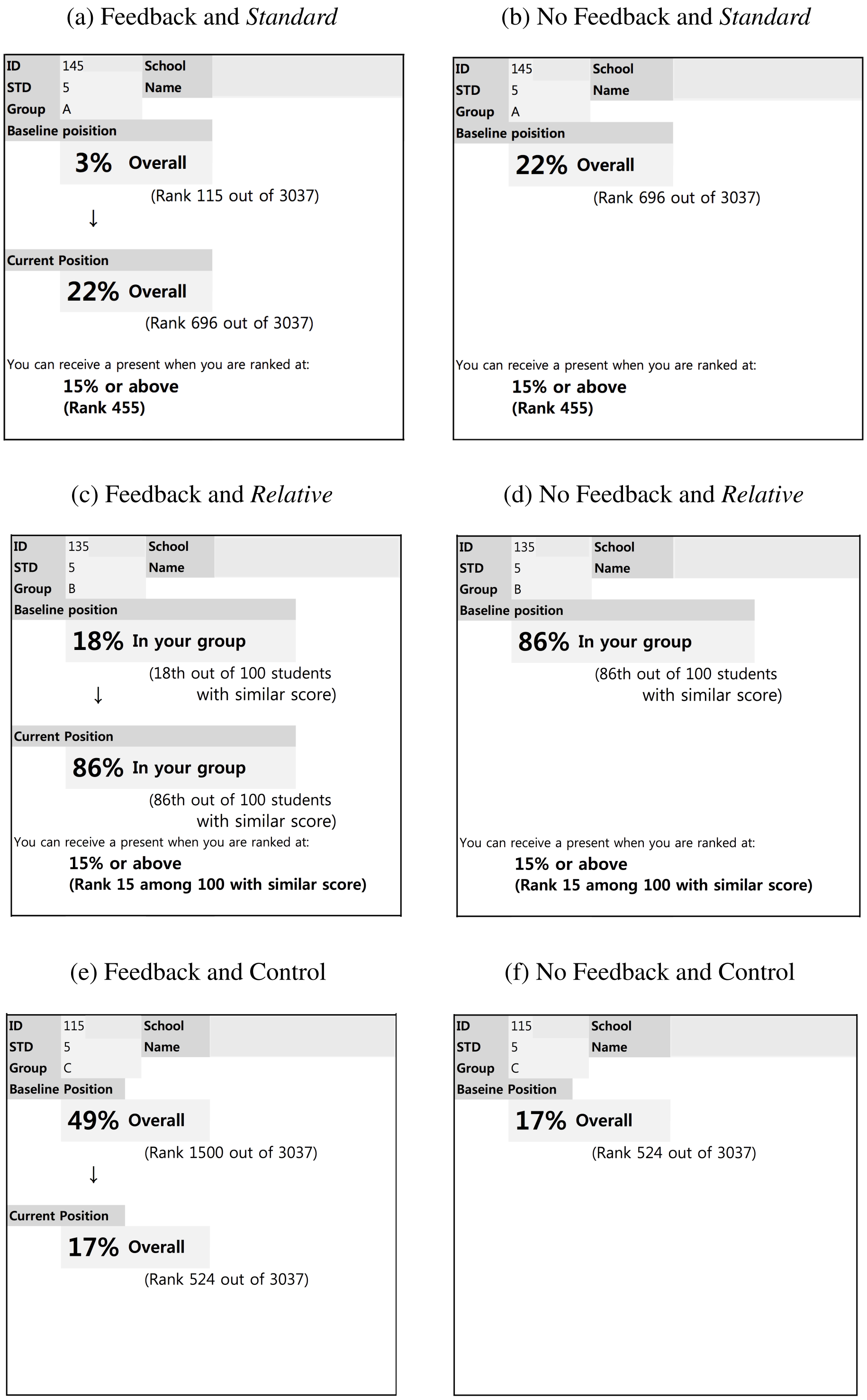
Current Position

74% [1784 out of 2668]

You can receive a present when you are reanked at:

Note: Panels (a), (b), and (c) show the scholarship program announcement notes that were given to students assigned to the *Standard* scholarship group, the *Relative* scholarship group, and the control group, respectively.

Figure 3: Feedback note



(a) Feedback and *Standard*

ID 145 School
STD 5 Name
Group A
Baseline poisition
3% Overall
(Rank 115 out of 3037)
↓
Current Position
22% Overall
(Rank 696 out of 3037)
You can receive a present when you are ranked at:
15% or above
(Rank 455)

(b) No Feedback and *Standard*

ID 145 School
STD 5 Name
Group A
Baseline poisition
22% Overall
(Rank 696 out of 3037)
You can receive a present when you are ranked at:
15% or above
(Rank 455)

(c) Feedback and *Relative*

ID 135 School
STD 5 Name
Group B
Baseline position
18% In your group
(18th out of 100 students with similar score)
↓
Current Position
86% In your group
(86th out of 100 students with similar score)
You can receive a present when you are ranked at:
15% or above
(Rank 15 among 100 with similar score)

(d) No Feedback and *Relative*

ID 135 School
STD 5 Name
Group B
Baseline position
86% In your group
(86th out of 100 students with similar score)
You can receive a present when you are ranked at:
15% or above
(Rank 15 among 100 with similar score)

(e) Feedback and Control

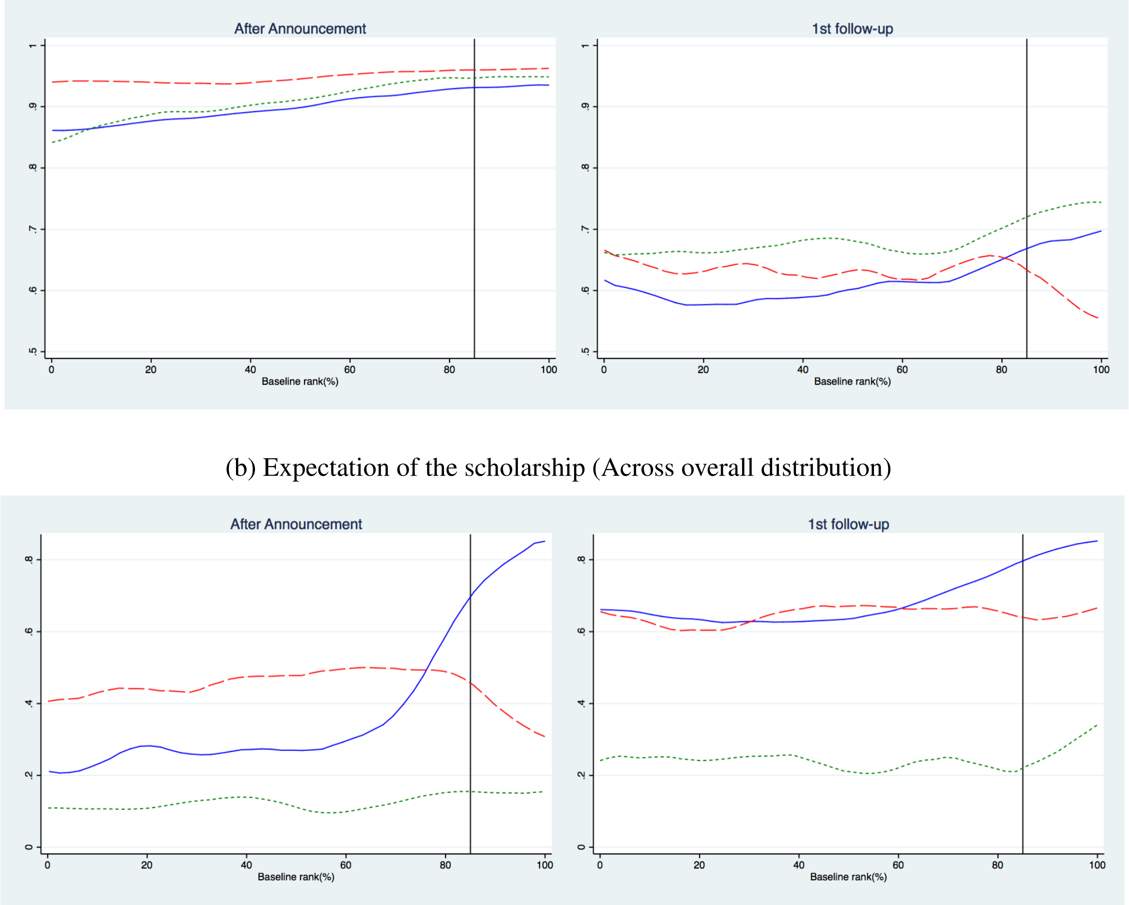
ID 115 School
STD 5 Name
Group C
Baseline Position
49% Overall
(Rank 1500 out of 3037)
↓
Current Position
17% Overall
(Rank 524 out of 3037)

(f) No Feedback and Control

ID 115 School
STD 5 Name
Group C
Baseine Position
17% Overall
(Rank 524 out of 3037)

Note: This figure shows feedback notes that students received in the second semester. The left column presents feedback notes given to the feedback treatment group and those in the right column present feedback notes given to the control group. The feedback treatment group received an information on their rank in the baseline and midterm exam while control group received an information only on the baseline exam. Panels (a) and (b), (c) and (d), and (e) and (f) compare the feedback provided for the Standard scholarship group, the Relative scholarship group, and the control group.

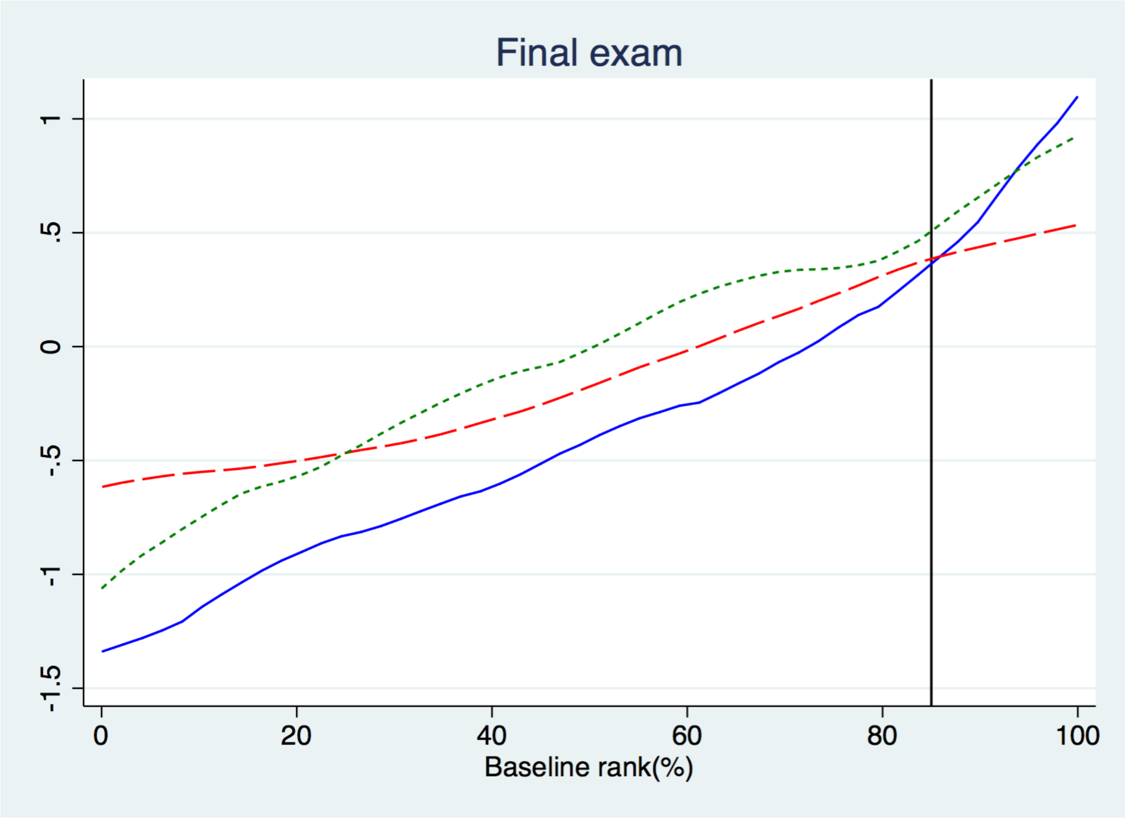
Figure 4: Understanding of the program and Expectation of the scholarship


(a) Understanding of the program

(b) Expectation of the scholarship (Across overall distribution)

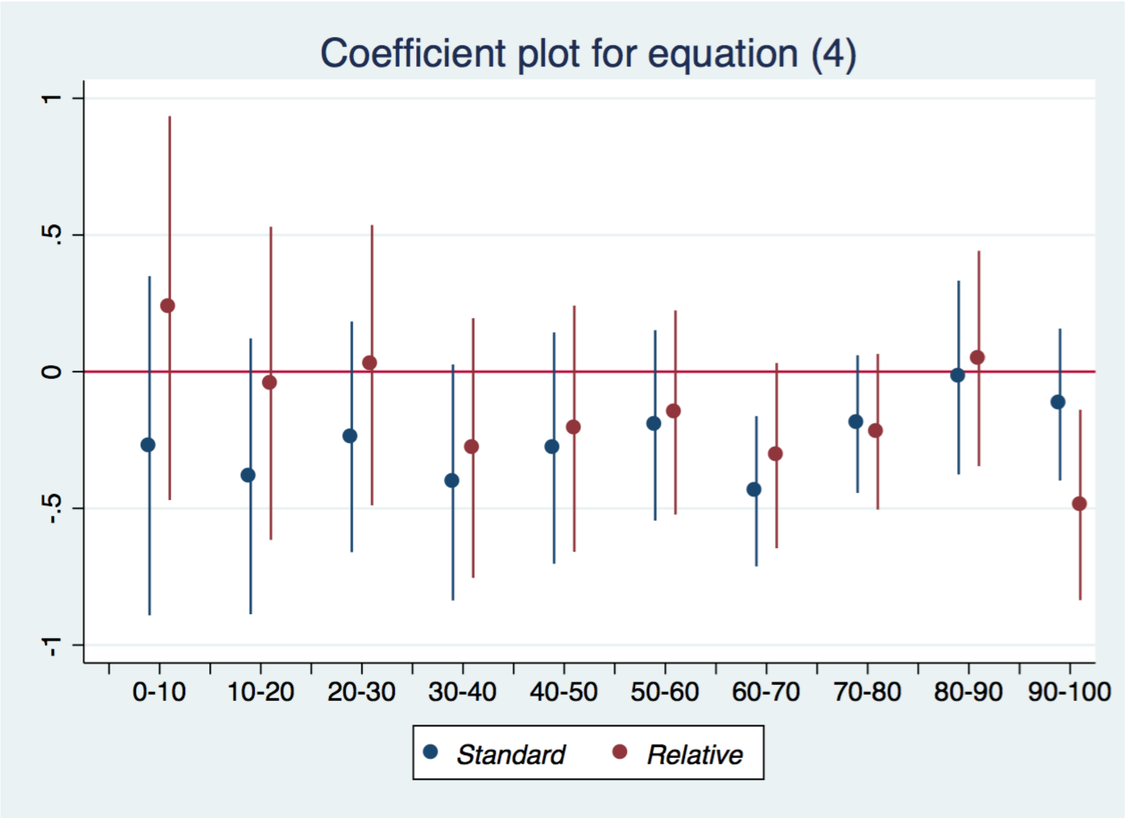
Note: The graph present level of understanding and expectation by baseline rank right after the randomization announcement and at the time of follow-up survey. X-axis presents baseline percentile rank of the students. Grade five to eight are the sample of the graphs. A blue(solid), red(dash), and green(dot) line present distribution among the Standard scholarship group, the Relative scholarship group, and the control group, respectively.

Figure 5: Exam scores at follow-up by Baseline Rank



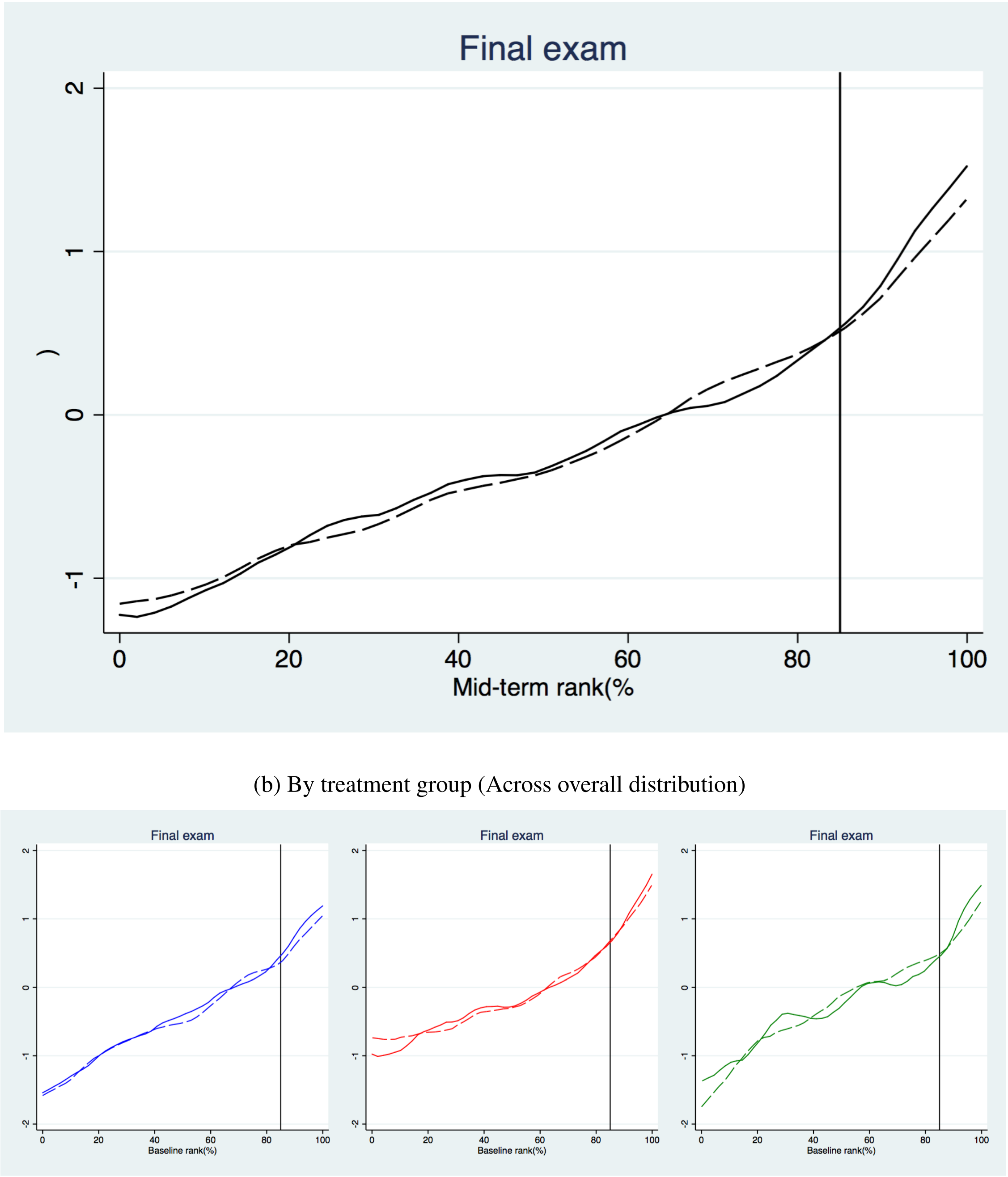
Note: This figure presents follow-up exam scores by baseline rank. X-axis presents baseline percentile rank of the students. A blue(solid), red(dash), and green(dot) line present distribution among the Standard scholarship group, the Relative scholarship group, and the control group, respectively.

Figure 6: Coefficient of scholarship program effect



Note: This figure presents coefficients of interest with 95% confidence interval from equation (4). X-axis presents baseline decile rank of the students. A navy, crimson markers present coefficients of the Standard scholarship effects and the Relative scholarship effect, respectively.

Figure 7: Feedback effect on follow-up exam score by mid-term rank



(a) Whole sample

(b) By treatment group (Across overall distribution)

Note: This figure presents 1st follow-up exam score by mid-term exam rank. A solid and dashed line present results for the feedback treatment group and the feedback control group. Panel A presents the results for whole group and Panel B the results by scholarship treatment status, respectively.

Table 1: Sample Composition by Treatment Category

Panel A: Scholarship Treatment Sample (Grade 5-8)

Scholarship Assignment	Classrooms	Students
Standard merit-based	46	2,830
Relative merit-based	43	2,994
Control	30	1,562
Total	119	7,386

Panel B: Feedback Treatment Sample (Grade 5-8)

Scholarship Assignment	Feedback Assignment	Students
Standard	No Feedback	1,175
merit-based	Feedback	1,195
Relative	No Feedback	1,360
merit-based	Feedback	1,364
Control	No Feedback	510
	Feedback	501
Total		6,105

Panel C: Long-term F/U Sample (Grade 5-6)

Scholarship Assignment	Classrooms	Students
Standard merit-based	19	1,911
Relative merit-based	20	2,022
Control	10	702
Total	49	4,635

Notes: A scholarship assignment was randomized by classroom level with grade stratification. A feedback randomization was done by individual level.

Table 2: Balance of Baseline Variables Across Treatment Groups

		Scholarship Randomization				Feedback Randomization	
	Whole Sample Mean	Control Mean	*Standard* vs. Control	*Relative* vs. Control	N	Feedback vs. No Feedback	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Age	14.2	14.4	-0.366	-0.300	7385	0.199**	6103
	[4.60]	[3.60]	(0.311)	(0.280)		(0.0932)	
Male	0.473	0.486	-0.00358	-0.0275	7385	0.0128	6103
	[0.499]	[0.500]	(0.0195)	(0.0177)		(0.0129)	
Ethnic group: Chewa	0.887	0.914	-0.0329	-0.0360	7358	-0.00274	6077
	[0.317]	[0.280]	(0.0352)	(0.0352)		(0.00641)	
Household size	7.84	7.70	0.164	0.192	7497	0.0782*	6199
	[1.82]	[1.88]	(0.345)	(0.318)		(0.0411)	
Asset index	-0.00396	-0.00919	0.000625	0.0124	7102	-0.0902*	5848
	[1.92]	[1.88]	(0.183)	(0.175)		(0.0510)	
B/L rank(%)	51.2	50.8	-0.0178	1.15	7497	-0.253	6199
	[28.4]	[27.9]	(3.01)	(4.01)		(0.625)	
Baseline score: Total	-0.00998	0.00000	-0.0543	0.0273	7497	-0.00715	6199
	[1.04]	[0.999]	(0.107)	(0.160)		(0.0204)	
Baseline score: Math	0.0250	0.0282	-0.00528	-0.00285	7407	-0.0105	6112
	[0.982]	[0.961]	(0.0797)	(0.0956)		(0.0230)	
Attendance	0.846	0.858	-0.0127	-0.0177	7497	0.00565	6199
	[0.197]	[0.201]	(0.0181)	(0.0179)		(0.00492)	
Study hours per week	16.1	16.8	-1.00	-0.818	7308	0.163	6031
	[16.1]	[16.4]	(0.865)	(0.871)		(0.374)	
Motivation to study [1-5]	4.52	4.53	-0.0541	0.0159	7374	-0.000297	6092
	[0.811]	[0.789]	(0.0650)	(0.0547)		(0.0210)	
Self-esteem [1-4]	2.65	2.67	-0.0273	-0.0188	7368	0.0105	6087
	[0.336]	[0.338]	(0.0232)	(0.0236)		(0.00688)	
Conscientious [1-5]	3.59	3.58	-0.0279	0.0454	7370	0.00236	6089
	[0.586]	[0.600]	(0.0676)	(0.0663)		(0.0154)	
Grit [1-5]	3.18	3.21	-0.0496*	-0.0287	7368	0.0205*	6087
	[0.433]	[0.450]	(0.0256)	(0.0280)		(0.0116)	
Teacher effort index [1-5]	4.03	3.96	0.0661	0.115	7364	0.00157	6083
	[0.537]	[0.567]	(0.0816)	(0.0724)		(0.0132)	
Parental encouragement	4.44	4.47	-0.0528	-0.0362	7281	0.0393**	6024
	[0.801]	[0.754]	(0.0566)	(0.0483)		(0.0187)	

Notes: Columns 1 and 2 reports means of selected baseline variables for the whole sample and for subjects assigned to the control group, respectively. Columns 3 and 4 report mean differences (and significance levels for difference of mean tests) between the scholarship treatment groups and the control group. Column 6 report mean difference between the feedback treatment and the control group. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 3: Understanding and Expectation

	Sample: Grade 5-8			
	Understanding		Expectation	
	After Announcement	1st Follow-up	After Announcement	1st Follow-up
	(1)	(2)	(3)	(4)
Panel A: Average Treatment effect				
Standard	-0.008	-0.019	0.305***	0.442***
	(0.021)	(0.022)	(0.056)	(0.044)
Relative	0.035*	-0.025	0.354***	0.397***
	(0.020)	(0.024)	(0.066)	(0.046)
R-Squared	0.067	0.098	0.115	0.146
P-value of F-test	0.006	0.819	0.411	0.053
Panel B: Heterogeneous treatment effect across overall distribution				
Standard	-0.008	-0.014	0.237***	0.417***
	(0.024)	(0.022)	(0.058)	(0.048)
Relative	0.038*	-0.005	0.381***	0.404***
	(0.022)	(0.025)	(0.065)	(0.049)
Std. x Top 15%	0.001	-0.037	0.452***	0.146***
	(0.024)	(0.036)	(0.077)	(0.049)
Rel. x Top 15%	-0.023	-0.120***	-0.132	-0.042
	(0.021)	(0.032)	(0.080)	(0.052)
Top 15%	0.016	0.084***	-0.065	-0.043
	(0.018)	(0.025)	(0.047)	(0.039)
N	5617	5851	5594	5750
R-Squared	0.068	0.102	0.155	0.150
Mean of Dep. Var.	0.924	0.636	0.356	0.579

Notes: Robust standard errors in brackets. Standard errors clustered at the classroom level. All specification include grade fixed effects, district fixed effects, and demographic controls such as age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 4: Test Score Impacts

	Sample: Grade 5-8					
	1st Follow-up					
	Exam Rank		Exam score (Norm)		Self evaluated performance	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Average Treatment effect						
Standard	-7.876*	-7.607*	-0.280*	-0.271*	-0.081**	-0.081**
	(4.041)	(3.932)	(0.153)	(0.149)	(0.040)	(0.040)
Relative	-4.290	-4.348	-0.127	-0.123	-0.059	-0.059
	(4.573)	(4.464)	(0.182)	(0.178)	(0.042)	(0.042)
R-Squared	0.284	0.302	0.291	0.313	0.099	0.108
P-value of F-test	0.328	0.369	0.321	0.331	0.575	0.584
Panel B: Heterogeneous treatment effect across overall distribution						
Standard	-8.712**	-8.536**	-0.307*	-0.302*	-0.084*	-0.086*
	(4.326)	(4.172)	(0.160)	(0.154)	(0.044)	(0.043)
Relative	-3.911	-4.001	-0.079	-0.079	-0.054	-0.051
	(5.022)	(4.871)	(0.195)	(0.189)	(0.046)	(0.045)
Std. x Top 15%	5.146	5.495	0.162	0.178	0.021	0.029
	(5.531)	(5.297)	(0.231)	(0.223)	(0.063)	(0.065)
Rel. x Top 15%	-2.640	-2.450	-0.279	-0.256	-0.077	-0.086
	(6.083)	(5.936)	(0.269)	(0.260)	(0.065)	(0.067)
Top 15%	2.873	2.889	0.043	0.057	0.245***	0.238***
	(4.329)	(4.183)	(0.189)	(0.185)	(0.042)	(0.043)
R-Squared	0.287	0.306	0.295	0.317	0.114	0.122
Panel C: Hegerogeneous treatment effect within subgroup distribution						
Standard	-6.185	-6.088	-0.224	-0.220	-0.070	-0.071
	(3.862)	(3.795)	(0.143)	(0.140)	(0.044)	(0.044)
Relative	-3.328	-3.578	-0.087	-0.090	-0.052	-0.054
	(4.524)	(4.445)	(0.175)	(0.172)	(0.045)	(0.045)
Std. x Subg. Top 15%	-10.171*	-9.077*	-0.337	-0.305	-0.080	-0.067
	(5.697)	(5.336)	(0.245)	(0.229)	(0.064)	(0.065)
Rel. x Subg. Top 15%	-7.113	-5.772	-0.310	-0.254	-0.045	-0.030
	(5.433)	(4.934)	(0.243)	(0.223)	(0.066)	(0.066)
Subg. Top 15%	-2.644	-3.359	-0.091	-0.114	0.069	0.058
	(5.093)	(4.624)	(0.225)	(0.207)	(0.051)	(0.051)
Demographic cont.	No	Yes	No	Yes	No	Yes
N	6689	6353	6689	6353	6048	5818
R-Squared	0.298	0.316	0.304	0.324	0.100	0.109
Mean of Dep. Var.	51.258	51.550	-0.156	-0.142	3.224	3.224

Notes: Robust standard errors in brackets. Standard errors are clustered at the the classroom level. All specification include grade fixed effects and district fixed effects. Demographic control includes age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 5: Intermediate Outcomes

	Sample: Grade 5-8							
	Student input		Non-cognitive traits			Teachers and parental response		
	Attendance	Study Hours	Motivation to study hard	Self esteem	Conscien tiousness	Teacher effort index	Parental effort index	Parents mentioned scholarship
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Average Treatment effect								
Standard	0.024*	-0.970	-0.071**	-0.030*	-0.045	-0.017	-0.021	0.126**
	(0.013)	(1.036)	(0.035)	(0.017)	(0.032)	(0.045)	(0.046)	(0.064)
Relative	0.009	-1.562	-0.036	-0.028	-0.027	-0.027	0.008	0.087
	(0.015)	(1.158)	(0.039)	(0.017)	(0.034)	(0.040)	(0.044)	(0.071)
R-Squared	0.193	0.076	0.022	0.050	0.080	0.091	0.042	0.038
P-value of F-test	0.253	0.523	0.239	0.911	0.529	0.812	0.299	0.544
Panel B: Heterogeneous treatment effect across overall distribution								
Standard	0.025*	-0.732	-0.093**	-0.034*	-0.053*	-0.020	-0.020	0.080
	(0.014)	(1.097)	(0.038)	(0.018)	(0.031)	(0.046)	(0.048)	(0.069)
Relative	0.008	-1.304	-0.048	-0.026	-0.012	-0.030	0.022	0.107
	(0.016)	(1.205)	(0.043)	(0.019)	(0.031)	(0.043)	(0.047)	(0.069)
Std. x Top 15%	-0.009	-1.214	0.137**	0.025	0.047	0.018	-0.004	0.276**
	(0.023)	(1.698)	(0.062)	(0.035)	(0.083)	(0.055)	(0.057)	(0.111)
Rel. x Top 15%	-0.008	-1.826	0.064	-0.019	-0.091	0.017	-0.080	-0.066
	(0.023)	(1.988)	(0.066)	(0.035)	(0.087)	(0.061)	(0.055)	(0.132)
Top 15%	0.049***	2.988**	-0.024	0.035	0.064	0.019	0.050	-0.236***
	(0.017)	(1.469)	(0.044)	(0.030)	(0.074)	(0.043)	(0.043)	(0.086)
N	7085	5242	5754	5842	5844	5838	5778	5848
R-Squared	0.195	0.078	0.023	0.052	0.083	0.091	0.043	0.042
Mean of Dep. Var.	0.756	14.526	4.298	2.719	3.674	4.006	4.060	3.409

Notes: Robust standard errors in brackets. Standard errors clustered at the classroom level. All specification include grade fixed effects, district fixed effects, baseline value of dependent variable, and demographic controls such as age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 6: Long term Test Score Impacts

	Sample: Grade 5-6			
	1st Follow-up (Norm)		2nd Follow-up (Norm)	
	(1)	(2)	(3)	(4)
Panel A: Average Treatment effect				
Standard	-0.466*	-0.462*	-0.092	-0.091
	(0.271)	(0.255)	(0.208)	(0.210)
Relative	-0.311	-0.327	-0.091	-0.093
	(0.297)	(0.284)	(0.177)	(0.180)
R-Squared	0.288	0.308	0.170	0.213
P-value of F-test	0.467	0.518	0.995	0.991
Panel B: Heterogeneous treatment effect across overall distribution				
Standard	-0.494	-0.498*	-0.132	-0.136
	(0.297)	(0.273)	(0.212)	(0.212)
Relative	-0.281	-0.307	-0.065	-0.071
	(0.330)	(0.310)	(0.173)	(0.175)
Std. x Top 15%	0.211	0.246	0.269	0.296
	(0.357)	(0.344)	(0.189)	(0.183)
Rel. x Top 15%	-0.144	-0.100	-0.095	-0.067
	(0.399)	(0.381)	(0.237)	(0.227)
Top 15%	-0.018	-0.010	-0.115	-0.148
	(0.291)	(0.278)	(0.150)	(0.144)
Demographic cont.	No	Yes	No	Yes
N	4118	3884	2522	2389
R-Squared	0.291	0.311	0.174	0.217
Mean of Dep. Var.	-0.271	-0.255	-0.025	-0.022

Notes: Robust standard errors in brackets. Standard errors are clustered at the the classroom level. All specification include grade fixed effects and district fixed effects. Demographic control includes age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table 7: Long term Intermediate Outcomes

	Sample: Grade 5-6								
	1st Follow-up					2nd Follow-up			
	Student input		Non-cognitive traits			Student input	Non-cognitive traits		
	Attendance	Study Hours	Motivation to study hard	Self esteem	Conscien tious-ness	Study Hours	Motivation to study hard	Self esteem	Conscien tious-ness
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Average Treatment effect									
Standard	0.017	-2.337**	-0.094*	-0.055***	-0.041	2.078*	0.009	-0.014	-0.085
	(0.019)	(1.102)	(0.048)	(0.021)	(0.038)	(1.224)	(0.055)	(0.032)	(0.054)
Relative	0.001	-4.058***	-0.059	-0.060***	-0.030	0.821	0.047	-0.019	-0.131**
	(0.020)	(1.164)	(0.051)	(0.022)	(0.043)	(0.739)	(0.055)	(0.034)	(0.054)
R-Squared	0.188	0.039	0.018	0.046	0.056	0.006	0.020	0.050	0.048
P-value of F-test	0.345	0.095	0.312	0.806	0.769	0.378	0.309	0.801	0.313
Panel B: Heterogeneous treatment effect across overall distribution									
Standard	0.017	-2.009*	-0.111**	-0.066***	-0.045	1.497	-0.005	-0.033	-0.108*
	(0.019)	(1.076)	(0.054)	(0.021)	(0.043)	(1.302)	(0.041)	(0.031)	(0.057)
Relative	0.002	-3.690***	-0.062	-0.064**	-0.008	1.142	0.038	-0.037	-0.125**
	(0.022)	(1.167)	(0.058)	(0.024)	(0.043)	(1.012)	(0.041)	(0.032)	(0.059)
Std. x Top 15%	0.002	-1.632	0.117	0.074*	0.042	3.637	0.088	0.119**	0.141
	(0.033)	(1.942)	(0.094)	(0.041)	(0.073)	(4.775)	(0.177)	(0.052)	(0.101)
Rel. x Top 15%	-0.012	-2.148	0.021	0.021	-0.115	-1.544	0.048	0.097**	-0.024
	(0.032)	(2.133)	(0.088)	(0.042)	(0.087)	(1.756)	(0.176)	(0.048)	(0.113)
Top 15%	0.046*	2.538	-0.034	-0.006	0.060	0.662	-0.031	-0.069**	-0.010
	(0.027)	(1.688)	(0.074)	(0.034)	(0.065)	(1.404)	(0.170)	(0.030)	(0.076)
N	4353	3241	3591	3631	3633	2410	2596	2597	2599
R-Squared	0.191	0.040	0.019	0.048	0.059	0.008	0.021	0.052	0.051
Mean of Dep. Var.	0.728	13.481	4.267	2.708	3.630	7.029	4.255	2.725	3.577

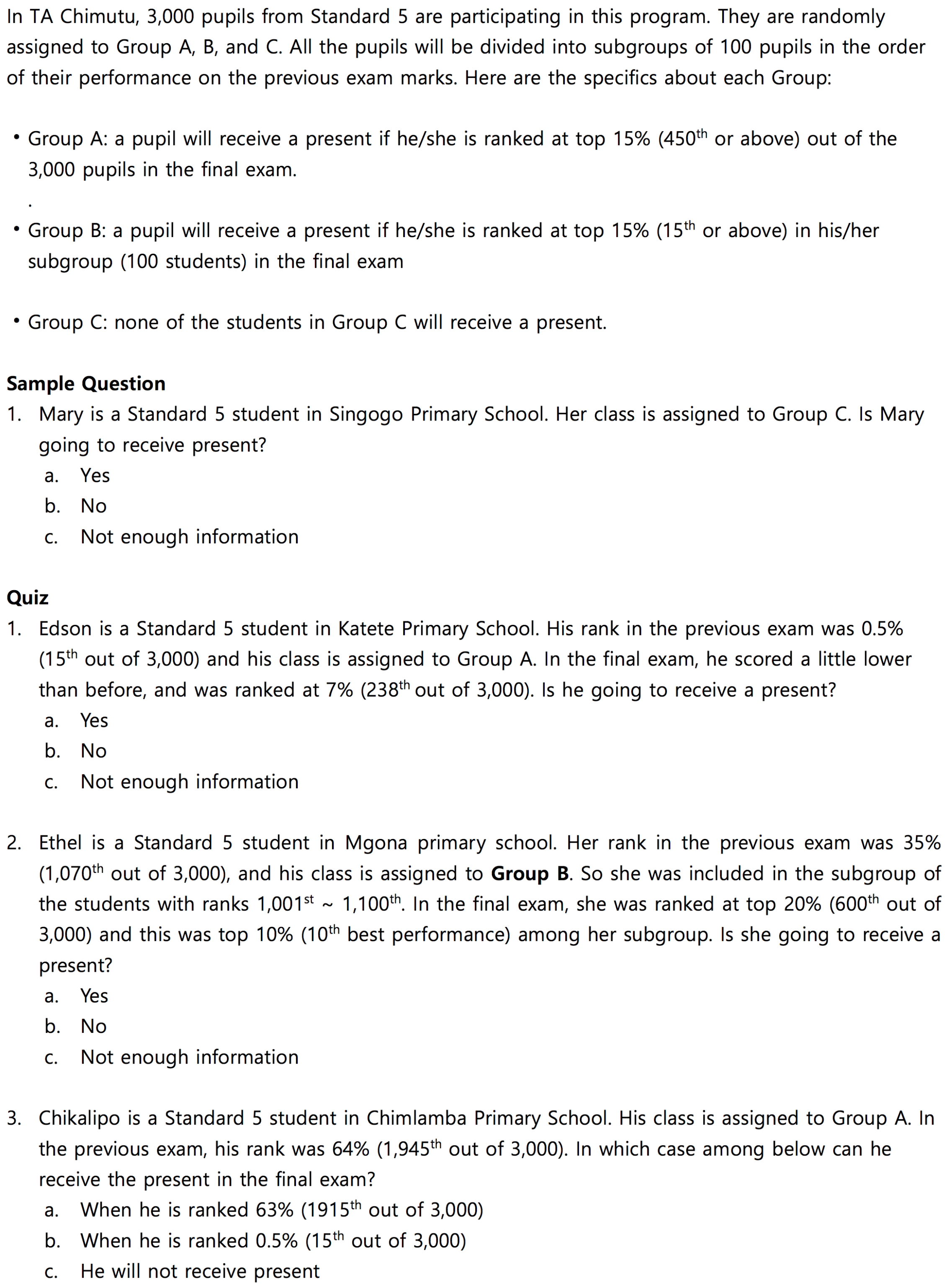
Notes: Robust standard errors in brackets. Standard errors clustered at the classroom level. All specification include grade fixed effects, district fixed effects, baseline value of dependent variable, and demographic controls such as age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 8: Feedback effect: Test Score Impacts

	Sample: Grade 5-7					
	Final exam			Self evaluated performance		
	All	Top 15%	Bot 85%	All	Top 15%	Bot 85%
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Feedback	0.032	0.088	0.030	-0.001	0.044	-0.005
	(0.022)	(0.064)	(0.026)	(0.020)	(0.053)	(0.023)
R-Squared	0.303	0.249	0.228	0.026	0.062	0.024
Panel B						
Feedback	0.058	0.137	0.059	0.015	0.125**	-0.013
	(0.067)	(0.096)	(0.079)	(0.033)	(0.061)	(0.038)
Standard	-0.300	-0.190	-0.246	-0.116**	0.086	-0.146***
	(0.206)	(0.264)	(0.176)	(0.049)	(0.084)	(0.048)
Relative	-0.188	-0.103	-0.085	-0.096*	-0.004	-0.091*
	(0.230)	(0.261)	(0.208)	(0.051)	(0.085)	(0.054)
Std. x FB	-0.028	-0.032	-0.034	-0.029	-0.229**	0.023
	(0.073)	(0.133)	(0.087)	(0.046)	(0.109)	(0.053)
Rel. x FB	-0.033	-0.090	-0.036	-0.009	0.034	-0.001
	(0.073)	(0.162)	(0.086)	(0.046)	(0.097)	(0.052)
N	5188	794	4394	4864	766	4098
R-Squared	0.312	0.255	0.237	0.030	0.072	0.029
Mean of Dep. Var.	-0.180	0.997	-0.393	3.251	3.585	3.189

Notes: Robust standard errors in brackets. Standard errors clustered at the classroom level. All specification include grade fixed effects, district fixed effects, baseline value of dependent variable, and demographic controls such as age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Figure A1: Quiz for program understanding



In TA Chimutu, 3,000 pupils from Standard 5 are participating in this program. They are randomly assigned to Group A, B, and C. All the pupils will be divided into subgroups of 100 pupils in the order of their performance on the previous exam marks. Here are the specifics about each Group:

- Group A: a pupil will receive a present if he/she is ranked at top 15% (450th or above) out of the 3,000 pupils in the final exam.

.

- Group B: a pupil will receive a present if he/she is ranked at top 15% (15th or above) in his/her subgroup (100 students) in the final exam

- Group C: none of the students in Group C will receive a present.

Sample Question

1. Mary is a Standard 5 student in Singogo Primary School. Her class is assigned to Group C. Is Mary going to receive present?
   a. Yes
   b. No
   c. Not enough information

Quiz

1. Edson is a Standard 5 student in Katete Primary School. His rank in the previous exam was 0.5% (15th out of 3,000) and his class is assigned to Group A. In the final exam, he scored a little lower than before, and was ranked at 7% (238th out of 3,000). Is he going to receive a present?
   a. Yes
   b. No
   c. Not enough information

2. Ethel is a Standard 5 student in Mgona primary school. Her rank in the previous exam was 35% (1,070th out of 3,000), and his class is assigned to **Group B**. So she was included in the subgroup of the students with ranks 1,001st ~ 1,100th. In the final exam, she was ranked at top 20% (600th out of 3,000) and this was top 10% (10th best performance) among her subgroup. Is she going to receive a present?
   a. Yes
   b. No
   c. Not enough information

3. Chikalipo is a Standard 5 student in Chimlamba Primary School. His class is assigned to Group A. In the previous exam, his rank was 64% (1,945th out of 3,000). In which case among below can he receive the present in the final exam?
   a. When he is ranked 63% (1915th out of 3,000)
   b. When he is ranked 0.5% (15th out of 3,000)
   c. He will not receive present

4. Enous is a Standard 5 student in Chang'ana Primary School. His class is assigned to Group B. In the previous exam, his rank was 23% (712th out of 3,000), so he was included in the subgroup of students with ranks between 701st ~ 800th. In which scenario will he receive a present in the final exam? (2 answers)
   a. When he is ranked at 10% (315th out of 3,000) and it was top 13% (13rd best performance) within his subgroup
   b. When he is ranked at 23% (710th out of 3,000) and it was top 10% (10th best performance) within his subgroup
   c. When he is ranked at 23% (710th out of 3,000) and it was top 79% (79th best performance) within his subgroup

5. Angella is a Standard 5 student in Phiri Primary School. Her rank in the previous exam was 83% (2,501st. out of 3,000),. In which group will she have the best chance of receiving a present in the final exam?
   a. Group A
   b. Group B
   c. Group C
   d. He has the same chance in Group A and B

Figure A2: Rosenberg self-esteem scale and big five inventory scale

Section VII: Non-Cognitive test

Direction: Here are a number of statements that may or may not apply to you. For the most accurate score, when responding, think of how you compare to most people – not just the people you know well, but most people in the world. There is no right or wrong answer, so just answer honestly! For the following statements, please indicate how often you did the following during the past school year.

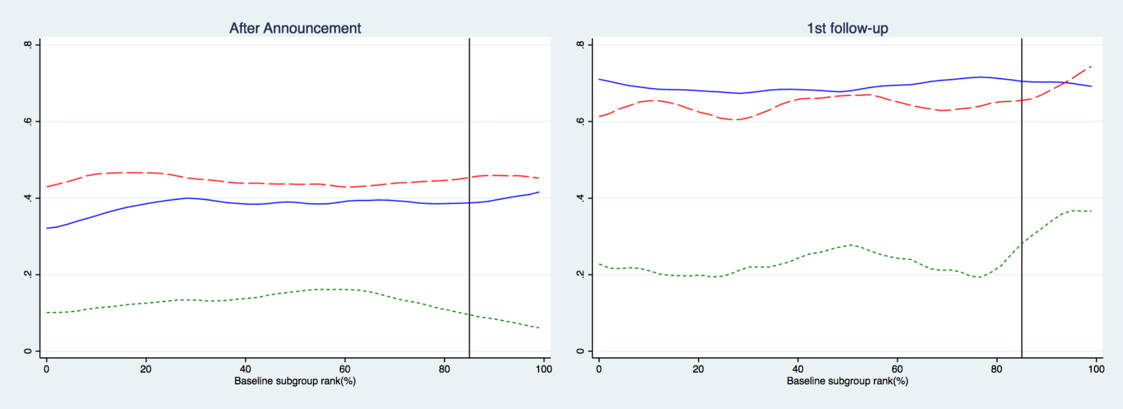
		Strongly disagree	Disagree	Agree	Strongly agree
701.	On the whole, I am satisfied with myself	1	2	3	4
702.	At times I think I am no good at all	1	2	3	4
703.	I feel that I have a number of good qualities	1	2	3	4
704.	I am able to do things as well as most other people.	1	2	3	4
705.	I feel I do not have much to be proud of.	1	2	3	4
706.	I certainly feel useless at times.	1	2	3	4
707.	I feel that I'm a person of worth, at least on an equal plane with others.	1	2	3	4
708.	I wish I could have more respect for myself.	1	2	3	4
709.	All in all, I am inclined to feel that I am a failure.	1	2	3	4
710.	I take a positive attitude toward myself.	1	2	3	4

		Not like me at all	Not much like me	Somewhat like me	Mostly like me	Very much like me
711.	New ideas and projects sometimes distract me from previous ones.	1	2	3	4	5
712.	Setbacks don't discourage me.	1	2	3	4	5
713.	I have been obsessed with a certain idea or project for a short time but later lost interest.	1	2	3	4	5
714.	I am a hard worker.	1	2	3	4	5
715.	I often set a goal but later choose to pursue a different one.	1	2	3	4	5
716.	I have difficulty maintaining my focus on projects that take more than a few months to complete.	1	2	3	4	5
717.	I finish whatever I begin.	1	2	3	4	5
718.	I am diligent.	1	2	3	4	5

	I see Myself as Someone Who...	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
719.	Does a thorough job	1	2	3	4	5
720.	Can be somewhat careless.	1	2	3	4	5
721.	Is a reliable worker.	1	2	3	4	5
722.	Tends to be disorganized.	1	2	3	4	5
723.	Tends to be lazy.	1	2	3	4	5
724.	Perseveres until the task is finished.	1	2	3	4	5
725.	Does things efficiently.	1	2	3	4	5
726.	Makes plans and follows through with them.	1	2	3	4	5
727.	Is easily distracted.	1	2	3	4	5

Figure A3: Expectation of the scholarship

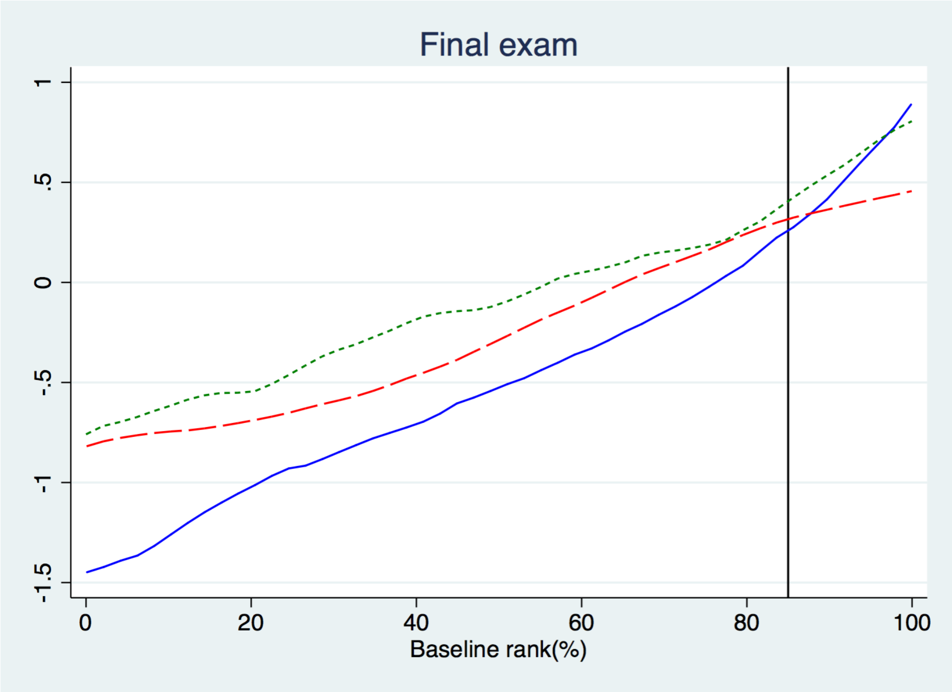
(a) Expectation of the scholarship (Within subgroup distribution)



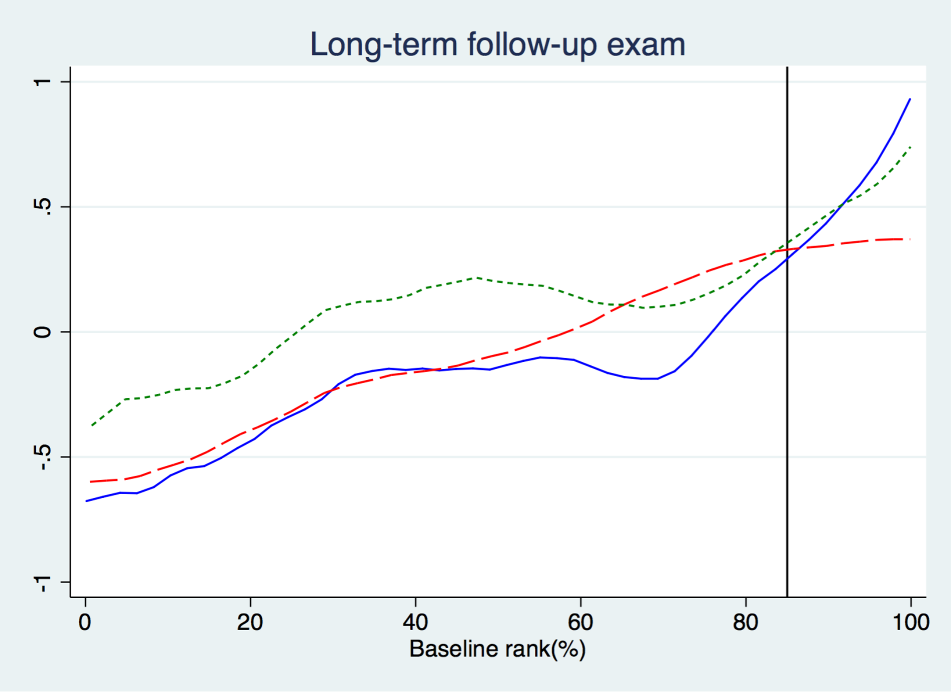
Note: The graph presents level of expectation by baseline rank within subgroup right after the randomization announcement and at the time of follow-up survey. X-axis presents baseline percentile rank of the students. Grade five to eight are the sample of the graphs. A blue(solid), red(dash), and green(dot) line present distribution among the Standard scholarship group, the Relative scholarship group, and the control group, respectively.

Figure A4: Exam scores at follow-up by Baseline Rank -long

(a) 1st follow-up exam, Long-term Follow-up Sample (Grade 5-6)



(b) 2nd follow-up exam, Long-term Follow-up Sample (Grade 5-6)



Note: This figure presents follow-up exam scores by baseline rank. X-axis presents baseline percentile rank of the students. A blue(solid), red(dash), and green(dot) line present distribution among the Standard scholarship group, the Relative scholarship group, and the control group, respectively.

Table A1: Balance of Baseline Variables Across Treatment Groups

		Scholarship Randomization				Feedback Randomization	
	Whole Sample Mean	Control Mean	*Standard* vs. Control	*Relative* vs. Control	N	Feedback vs. No Feedback	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Age	13.6	13.6	-0.0716	0.164	4562	0.308**	4562
	[5.64]	[4.89]	(0.313)	(0.301)		(0.118)	
Male	0.465	0.488	-0.0255	-0.0272	4562	0.00910	4562
	[0.499]	[0.500]	(0.0252)	(0.0223)		(0.0161)	
Ethnic group: Chewa	0.888	0.947	-0.0704	-0.0680*	4541	0.00530	4541
	[0.316]	[0.225]	(0.0426)	(0.0402)		(0.00646)	
Household size	7.77	7.68	-0.0292	0.229	4635	0.0823*	4635
	[1.79]	[1.93]	(0.535)	(0.491)		(0.0480)	
Asset index	-0.00724	-0.225	0.329*	0.190	4365	-0.0410	4365
	[1.92]	[1.76]	(0.179)	(0.160)		(0.0597)	
B/L rank(%)	51.4	54.1	-3.55	-2.93	4635	-0.549	4635
	[28.3]	[27.5]	(4.64)	(5.82)		(0.747)	
Baseline score: Total	-0.0757	0.00000	-0.137	-0.0439	4635	-0.0194	4635
	[1.02]	[0.999]	(0.156)	(0.222)		(0.0233)	
Baseline score: Math	0.0230	0.240	-0.242**	-0.268**	4568	-0.0151	4568
	[0.979]	[0.951]	(0.110)	(0.121)		(0.0269)	
Attendance	0.829	0.829	-0.000472	0.00129	4635	0.00515	4635
	[0.202]	[0.210]	(0.0242)	(0.0227)		(0.00587)	
Study hours per week	15.6	15.4	0.242	0.181	4502	0.295	4502
	[16.4]	[16.4]	(1.09)	(1.04)		(0.449)	
Motivation to study [1-5]	4.47	4.46	-0.0354	0.0646	4552	-0.00890	4552
	[0.853]	[0.817]	(0.0898)	(0.0758)		(0.0252)	
Self-esteem [1-4]	2.63	2.61	0.0148	0.0174	4550	0.0130	4550
	[0.333]	[0.333]	(0.0336)	(0.0322)		(0.00796)	
Conscientious [1-5]	3.52	3.45	0.0358	0.125	4552	0.0190	4552
	[0.584]	[0.591]	(0.107)	(0.103)		(0.0164)	
Grit [1-5]	3.15	3.14	-0.00741	0.0113	4550	0.0225	4550
	[0.423]	[0.432]	(0.0379)	(0.0390)		(0.0138)	
Teacher effort index [1-5]	4.03	3.94	0.0552	0.153	4548	0.0129	4548
	[0.555]	[0.595]	(0.127)	(0.114)		(0.0157)	
Parental encouragement	4.39	4.41	-0.0435	-0.000821	4506	0.0256	4506
	[0.832]	[0.767]	(0.0863)	(0.0738)		(0.0215)	

Notes: Columns 1 and 2 reports means of selected baseline variables for the whole sample and for subjects assigned to the control group, respectively. Columns 3 and 4 report mean differences (and significance levels for difference of mean tests) between the scholarship treatment groups and the control group. Column 6 report mean difference between the feedback treatment and the control group. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A2: Sample Attrition

	Dependent Variable: Participated							
	Sample: Grade 5-8			Sample: Grade 5-6				
	Mid-term	1st Follow-up		Mid-term	1st Follow-up		2nd Follow-up	
	Exam	Survey	Exam	Exam	Survey	Exam	Survey	Exam
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A								
Standard	0.020	-0.016	0.024	0.020	-0.005	0.025	0.032	0.034
	(0.017)	(0.018)	(0.017)	(0.029)	(0.019)	(0.018)	(0.031)	(0.035)
Relative	0.010	-0.027	0.032*	0.001	-0.017	0.030	0.032	0.050
	(0.018)	(0.018)	(0.016)	(0.028)	(0.022)	(0.019)	(0.035)	(0.036)
N	7085	7085	7085	4353	4353	4353	4191	4191
R-Squared	0.148	0.093	0.085	0.153	0.096	0.097	0.113	0.086
Mean of Dep. Var.	0.877	0.827	0.897	0.860	0.837	0.892	0.630	0.570
Panel B								
Feedback		0.002	-0.003		-0.004	-0.003	0.023	0.008
		(0.008)	(0.007)		(0.010)	(0.007)	(0.014)	(0.014)
R-Squared		0.100	0.101		0.096	0.096	0.113	0.085
Panel C								
Feedback		0.002	-0.003		-0.004	-0.003	0.022	0.008
		(0.008)	(0.007)		(0.010)	(0.007)	(0.014)	(0.014)
Standard		-0.012	0.027*		-0.005	0.025	0.032	0.034
		(0.017)	(0.016)		(0.019)	(0.018)	(0.031)	(0.035)
Relative		-0.021	0.030*		-0.017	0.030	0.032	0.049
		(0.019)	(0.016)		(0.022)	(0.019)	(0.035)	(0.036)
N		5832	5832		4353	4353	4191	4191
R-Squared		0.101	0.102		0.096	0.097	0.114	0.087
Mean of Dep. Var.		0.837	0.890		0.837	0.892	0.630	0.570

Notes: Robust standard errors in brackets. Standard errors clustered at the classroom level. All specification include grade fixed effects, district fixed effects, baseline value of dependent variable, and demographic controls such as age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A3: Understanding and Expectation within subgroup distribution

	Sample: Grade 5-8			
	Understanding		Expectation	
	After Announcement	1st Follow-up	After Announcement	1st Follow-up
	(1)	(2)	(3)	(4)
Standard	-0.013	-0.019	0.291***	0.458***
	(0.021)	(0.022)	(0.057)	(0.041)
Relative	0.030	-0.028	0.343***	0.410***
	(0.019)	(0.025)	(0.066)	(0.043)
Std. x Subg. Top 15%	0.034	0.002	0.095*	-0.110*
	(0.021)	(0.041)	(0.056)	(0.064)
Rel. x Subg. Top 15%	0.033*	0.018	0.065	-0.080
	(0.018)	(0.024)	(0.055)	(0.061)
Subg. Top 15%	-0.033*	-0.034*	-0.106***	0.104*
	(0.017)	(0.019)	(0.038)	(0.057)
N	5617	5851	5594	5750
R-Squared	0.068	0.099	0.117	0.147
Mean of Dep. Var.	0.924	0.636	0.356	0.579

Notes: Robust standard errors in brackets. Standard errors clustered at the classroom level. All specification include grade fixed effects, district fixed effects, and demographic controls such as age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A4: Test Score Impacts within subgroup distribution

	Sample: Grade 5-8			
	1st Follow-up			
	Exam Rank		Exam score (Norm)	
	(1)	(2)	(3)	(4)
Standard	-6.185	-6.088	-0.224	-0.220
	(3.862)	(3.795)	(0.143)	(0.140)
Relative	-3.328	-3.578	-0.0874	-0.0903
	(4.524)	(4.445)	(0.175)	(0.172)
Std. x Subg. Top 15%	-10.17*	-9.077*	-0.337	-0.305
	(5.697)	(5.336)	(0.245)	(0.229)
Rel. x Subg. Top 15%	-7.113	-5.772	-0.310	-0.254
	(5.433)	(4.934)	(0.243)	(0.223)
Subg. Top 15%	-2.644	-3.359	-0.0905	-0.114
	(5.093)	(4.624)	(0.225)	(0.207)
Demographic cont.	No	Yes	No	Yes
N	6689	6353	6689	6353
R-Squared	0.298	0.316	0.304	0.324
Mean of Dep. Var.	51.258	51.550	-0.156	-0.142

Notes: Robust standard errors in brackets. Standard errors are clustered at the the classroom level. All specification include grade fixed effects and district fixed effects. Demographic control includes age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table A5: Test score impacts (Noncognitive traits controlled)

	Sample: Grade 5-8			
	Exam Rank		Exam score (Norm)	
	(1)	(2)	(3)	(4)
Panel A: Average Treatment effect				
Standard	-7.161*	-6.996*	-0.248	-0.241
	(4.035)	(3.925)	(0.154)	(0.150)
Relative	-4.466	-4.696	-0.126	-0.129
	(4.623)	(4.484)	(0.185)	(0.180)
R-Squared	0.292	0.310	0.298	0.319
P-value of F-test	0.463		0.439	
Panel B: Heterogeneous treatment effect across overall distribution				
Standard	-7.938*	-7.955*	-0.275*	-0.276*
	(4.352)	(4.196)	(0.162)	(0.155)
Relative	-4.223	-4.567	-0.085	-0.095
	(5.111)	(4.922)	(0.200)	(0.193)
Std. x Top 15%	4.996	5.946	0.177	0.215
	(5.411)	(5.217)	(0.230)	(0.223)
Rel. x Top 15%	-1.618	-1.025	-0.225	-0.187
	(6.047)	(5.889)	(0.268)	(0.258)
Top 15%	2.357	2.298	0.029	0.037
	(4.191)	(4.077)	(0.185)	(0.182)
Demographic cont.	No	Yes	No	Yes
N	5910	5634	5910	5634
R-Squared	0.294	0.313	0.301	0.323
Mean of Dep. Var.	52.326	52.505	-0.122	-0.113

Notes: Robust standard errors in brackets. Standard errors are clustered at the the classroom level. All specification include grade fixed effects and district fixed effects. Demographic control includes age, race, household size, and a household asset index. Noncognitive traits include self esteem, grit scale and conscientiousness. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table A6: Classroom environment

	Sample: Grade 5-8					
	Smart students help friends better	Willingness to help friends	Received help from friends	Provided help to friends	Asked for help to friends	Classroom competitiveness index
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Average Treatment effect						
Standard	0.075	-0.042	0.081	0.075	0.036	0.035
	(0.103)	(0.063)	(0.061)	(0.066)	(0.066)	(0.050)
Relative	-0.208	0.010	-0.049	0.003	-0.053	-0.024
	(0.135)	(0.061)	(0.064)	(0.067)	(0.065)	(0.050)
R-Squared	0.083	0.018	0.021	0.008	0.008	0.018
P-value of F-test	0.015	0.171	0.006	0.178	0.121	0.038
Panel B: Heterogeneous treatment effect across overall distribution						
Standard	0.141	-0.029	0.077	0.078	0.012	0.032
	(0.107)	(0.075)	(0.056)	(0.065)	(0.079)	(0.053)
Relative	-0.196	0.014	-0.055	0.035	-0.069	-0.021
	(0.150)	(0.073)	(0.067)	(0.063)	(0.071)	(0.053)
Std. x Top 15%	-0.421***	-0.081	0.025	-0.022	0.150	0.020
	(0.148)	(0.129)	(0.134)	(0.146)	(0.186)	(0.087)
Rel. x Top 15%	-0.165	-0.040	0.025	-0.160	0.125	-0.011
	(0.184)	(0.129)	(0.154)	(0.173)	(0.202)	(0.117)
Top 15%	0.409***	0.072	0.006	0.057	-0.189	-0.013
	(0.132)	(0.118)	(0.118)	(0.125)	(0.159)	(0.078)
Demographic cont.	Yes	Yes	Yes	Yes	Yes	Yes
N	2698	2697	2690	2692	2698	2700
R-Squared	0.088	0.019	0.021	0.009	0.010	0.018
Mean of Dep. Var.	3.754	4.072	3.889	3.828	4.096	3.970

Notes: Robust standard errors in brackets. Standard errors are clustered at the the classroom level. All specification include grade fixed effects and district fixed effects. Demographic control includes age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; *** at 0.01.