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

James Berry, Hyuncheol Bryant Kim, and Hyuk Son*

July 2019

Abstract

We study how the structure of tournament incentive schemes in education can influence the level and distribution of student outcomes. Through a field experiment among upper-primary students in Malawi, we evaluate two scholarship programs: a *Standard* scholarship that rewarded overall top performers on an exam and a *Relative* scholarship that rewarded the top performers within smaller groups of students with similar baseline scores. We find that the *Standard* scholarship decreased test scores and motivation to study, especially for those least likely to win. 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 Hanyoun So, Seung Chul Lee, 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, briq/IZA Workshop on Behavioral Economics of Education, the Northeast Universities Development Consortium Conference, Korean Economic Association-Asia Pacific Economic Association International Conference, and Seoul Journal of Economics International conference provided helpful comments and suggestions.

1 Introduction

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 counteract 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 of the most often-studied incentive schemes is an individual tournament in which the top performing students on an exam are provided with a reward. Such a scheme allows for the policy maker to set a fixed budget for the incentives, and has 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.[2] In the same vein, the bottom students who are unlikely to receive the reward may not be motivated to exert effort. These effects could result in increased inequality in academic performance.

In addition, the impact of incentives may depend on the information the students have regarding their academic progress. In particular, if students lack precise information on their likelihood of obtaining the incentive, providing them with feedback on performance may enhance the distributional impacts of financial incentives, encouraging those at the top or discouraging those at the

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

[2] 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 Klaauw, 2010; Bettinger, 2011). However, other studies do not find evidence for such effects (e.g., Kremer, Miguel, and Thornton, 2009).

bottom.[3]

In this paper, we study the impacts of two types of incentive programs, as well as performance feedback, on 5th to 8th graders in 31 primary schools in Malawi. The two incentive programs, framed as scholarship schemes, provided rewards of MWK 4500 (USD 9.70) if the corresponding test score goal was met.[4] The first, which we call the *Standard* merit-based scholarship (hereafter *Standard* scholarship) scheme, provided a scholarship to students in the sample who scored in the top 15 percent on the final end-of-year exam in the sub-district. 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 (hereafter *Relative* 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."[5]

We implemented a randomized trial where 5th to 8th grade classrooms were assigned to *Standard* and *Relative* scholarships or a control group. We interviewed 5th to 8th graders at baseline as well as right before the final exam was administered (a short-term follow-up). 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

[3] Students may also respond to information for reasons unrelated to financial 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).

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

[5] 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).

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. By contrast, the *Relative* merit-based scholarship scheme did not have significant impacts on test score performance or motivation. This suggests that by providing a greater chance for all students to receive the reward, the negative motivational effects of high-powered incentives can be mitigated. Still, the *Relative* scholarship failed to produce positive impacts.

Although the effects of the *Standard* scholarship on the final exam were negative, these effects were localized to the incentivized test and did not persist into the next semester, after the incentive had been removed. We find no detectable effects of the *Standard* scholarship on performance on a math test administered as part of the endline survey, conducted just before the final exam. We also find no evidence for effects on final exams scores in the first semester of 2016-2017, for the subsample of students who were initially in grades 5 to 7. These results suggests that the *Standard* scholarship may have de-motivated students to perform the task that it was incentivizing and did not have meaningful effects outside of this task.

We also find that the feedback intervention largely failed to influence test scores, either in the full sample or within each scholarship treatment group. We provide suggestive evidence that this lack of impact is due to the limited amount of information provided: students knew their baseline test scores and appeared to have additional information on their progress during the semester, aside from the information the feedback was providing.

Taken together, these results are consistent with arguments that financial incentives may crowd out intrinsic motivation (Gneezy, Meier, and Rey-Biel, 2011). In this case, tournament incentives may also de-motivate low-performing students by reminding them of their place in the performance distribution and signalling that high performance is valuable. This is similar to research showing that social identity may affect performance. For example, Hoff and Pandey (2014) find that in mixed-caste classrooms in India, caste revelation significantly lowers the performance of low-caste students.

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 Rey-Biel, 2011, for a review).[6] The work closest to our *Standard* scholarship is that of Kremer, Miguel, and Thornton (2009), who study a merit scholarship program for 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 both boys and girls. A second key difference is that in our setting, students are aware of their initial test score and percentile rank. 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 for achievement (Hirshleifer, 2017), incentives targeted to parents and to children (Berry, 2015), and incentives for students and 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 school students, Bettinger (2011) finds that incentives for exam performance did not decrease survey-based intrinsic motivation, while Visaria et al. (2016)

[6] Within the developed-country literature, of particular note is Leuven, Oosterbeek, and 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.

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

2 Context, Programs, and Study 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.[7]

The academic calendar, starting in 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

[7] 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.

typically must pay a fee of about USD 0.5 to 1 to take the exam, to cover printing costs 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 Study Design

The study was conducted in TA Chimutu, a rural sub-district with three school zones located about 15 km from the capital city of Lilongwe.[8] The scholarship programs were conducted in grades 5 to 8 in 31 public primary schools in the sub-district. There is a total of 118 school-grades since several schools do not have upper grades. The scholarships were implemented by the Africa Future Foundation (AFF), an international NGO focused on health and education programs in Malawi and several other countries in Africa.

2.2.1 Study design

The project chronology and study design are 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). Baseline exams were conducted twice, at the end of the first semester (December 2014) and the beginning of the second semester (January 2015).[9] The midterm exam data used for the feedback intervention was implemented at the end of the 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 of the 2014-2015 academic year, in June 2015. Lastly, for students initially in the 5th and 6th grades, we collected sub-district-level exam scores in March of 2016, nine months after the scholarship programs ended.

In February 2015, we stratified the 118 school-grades by grade and randomly assigned school-

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

[9] 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 scores are similar: 11.5 (3.2) and 11.5 (3.4), respectively.

grades into three groups: the *Standard* scholarship, the *Relative* scholarship, or the control group. The results of the scholarship randomization were announced in the middle of the second semester. At the time of the randomization announcement, each student was provided an individualized note describing his or her treatment assignment. Figure 2 provides examples of notes for each treatment group, as well as the control group. For the *Standard* scholarship group, information on the student's overall sub-district rank (hereafter overall rank) as well as the scholarship eligibility condition (top 15 percent) was provided. For the *Relative* scholarship group, information on overall rank and rank within bin (hereafter bin rank) as well as the scholarship eligibility condition (top 15 percent within bin) was provided. For the control group, only information on the student's overall rank was provided.

The feedback intervention provided rank information on the midterm exam, administered at the end of the second semester (March 2015), to a random set of students. Specifically, across all three scholarship study groups, students in grades 5 to 7 were individually randomized into a "feedback" or "no-feedback" group.[10]

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 5th and 6th graders at baseline were administered nine months after the experiment was completed (March 2016). Table 1 displays the sample composition in each treatment category.

2.2.2 Scholarship Programs

Under the *Standard* scholarship scheme, within each grade, students scoring in the top 15 percent in the sub-district on the final exam were eligible to receive the award.[11] Under the *Relative* scholarship scheme, students were grouped into bins of 100 students by baseline test score, and the

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

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

top 15 percent of each bin in the final exam were eligible to receive the award.

The awards for *Standard* and *Relative* 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.[12] This represents a significant amount considering that Malawi GDP per capita was only around USD 362.7 in 2014 (The World Bank, 2015).

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. At the end of the session, students were informed of their treatment and control assignments, and took a short 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 overall and bin rank in the final exam. To measure expectations of winning a scholarship, 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 sub-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 sub-district.[13] 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. At the beginning of the third semester (March of 2015), each student received a note providing their ranking as of the midterm exam privately in a separated place and encouraged

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

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

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 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 bin rankings in the midterm.

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 scholarship eligibility. There is potential complementarity between feedback on relative performance and test scores 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, which could make the feedback effect less effective.

2.3 Data

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

Our main source of data is student performance on the sub-district-level exams. The main outcome variables are test scores and students' overall rank in these tests.[14] In addition to the exams, we measured students' school attendance through unannounced checks. These checks were conducted every month between April 2014 and June 2015, four times before the scholarship announcement and four times after.

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 –

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

including self esteem, conscientiousness, and grit – and motivation. Our measure of self esteem is based on the Rosenberg self-esteem scale, which measures both positive and negative feelings about oneself (Rosenberg, 1965). Conscientiousness was measured using questions based on the Big Five Inventory scale (John and Srivastava, 1999). To measure grit, we used the Short Grit Scale from Duckworth and Quinn (2009).[15] Finally, motivation was measured by asking how strongly the students agree with the statement "I am motivated to study hard" on a five-point scale, with one being strongly disagree and five being strongly agree.[16] To measure impacts on overall non-cognitive skills, we aggregate all four measures into an index, following the method of Kling, Liebman, and Katz (2007).[17]

In addition, 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, 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 constructed our sample by first collecting a list of all enrolled students in grades 5 to 8 in participating schools. Among these 9,419 students, 7,638 (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 for the scholarship and feedback randomizations. Column 1 displays 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 school attendance rate of the students was 85 percent, and the average study hours per week was 16.1.

[15] Survey questions used to measure self-esteem, grit, and conscientiousness are shown in Appendix Figure A2. Grit and conscientiousness questions were measured on a five-point scale, and self-esteem questions were measured on a four-point scale. We take the simple average of scores for all questions in a category to form our measures.

[16] Our measure of motivation captures general motivation to study, which includes both intrinsic motivation (often defined as studying for the joy of learning, see, e.g., Bettinger (2011)) as well as extrinsic motivation to study in order to receive the scholarship.

[17] The index is constructed by taking the average of the standardized measures, where the mean and standard deviation in the control group is used in the standardization. The resulting index is also standardized relative to the control group, so that it has a mean of 0 and standard deviation of 1.

Columns (2) and (3) of Table 2 show tests of differences in means between the scholarship groups and the control group, and Column (5) 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 compared with the no-feedback group, a difference of 0.64 percent.[18]

Table A1 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, respectively. 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. In Appendix A.1, we present additional analysis of attrition by scholarship treatment. The analysis shows that scholarship treatment effects, as well as interactions with between treatment and baseline test score, are unlikely to be substantively affected by differential attrition.

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:

$$Y_{igsz1} = \beta_0 + \beta_1 Standard_{gsz} + \beta_2 Relative_{gsz} + Y_{igsz0} + X_{igsz} + \eta_g + \gamma_z + \epsilon_{igsz} \tag{1}$$

where Y_{igsz1} is the outcome of interest for student i of grade g in school s at school zone z.

[18] We find similar result for students in grades 5 and 6, as shown in Table A4.

Standard and *Relative* are indicators for being *Standard* and *Relative* scholarship groups, respectively. Y_{igsz0} is the outcome measured at baseline. η_g is a grade fixed effect and γ is a fixed effect for zone. In some specifications, we include X_{igsz}, a set of student-level controls, including age, race, household size, and a household asset index. Standard errors are clustered at the the school-grade 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 students' initial rank. First, we present nonparametric plots to show impacts across sub-district baseline rank as well as bin rank used for the *Relative* scholarship. For the corresponding regressions, we interact the treatment groups with an indicator for whether the student's baseline overall rank was in the top 15 percent. We select the top 15% because students' responses to the scholarships might differ based on whether they are above or below the cutoff for scholarship eligibility. This implies the following regression:

$$\begin{aligned} Y_{igsz1} =& \beta_0 + \beta_1 Standard_{gsz} + \beta_2 Relative_{gsz} + \beta_3 Top15_{igsz0} \\ &+ \beta_4 Standard_{gsz} * Top15_{igsz0} + \beta_5 Relative_{gsz} * Top15_{igsz0} + Y_{igsz0} + \eta_g + \gamma_z + X_{igsz} + \epsilon_{igsz} \end{aligned} \tag{2}$$

where $Top15_{igsz0}$ is an indicator for being within the top 15 percent as of 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* scholarship group between the top 15 and bottom 85 percent of 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:

$$Y_{igsz1} =\beta_0 + \beta_1 Feedback_{igsz} + Y_{igsz0} + \eta_g + \gamma_z + X_{igsz} + \epsilon_{igsz} \tag{3}$$

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:

$$\begin{aligned} Y_{igsz1} =& \beta_0 + \beta_1 Standard_{gsz} + \beta_2 Relative_{gsz} + \beta_3 Feedback_{igsz} + \beta_4 Standard_{gsz} * Feedback_{igsz} \\ &+ \beta_5 Relative_{gsz} * Feedback_{igsz} + Y_{igsz0} + \eta_g + \gamma_z + X_{igsz} + \epsilon_{igsz} \end{aligned} \tag{4}$$

In these specifications, β_3 shows how feedback affects 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 expectations that they would receive the scholarship. As described in Section 2.2, students' understanding and expectations were elicited at the time of the program announcement, and again during the follow-up survey before the final exam. The results confirm that students generally understood the scholarship scheme and had expectations consistent with their assigned groups.

Figure 4 presents graphs of percent of questions answered correctly on the test for understanding of the scholarship schemes (y-axis) by baseline overall rank (x-axis) and by scholarship treatment group. Columns (1) and (2) of Table 3 present 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. Understanding was fairly similar across groups. Panel A of Table 3 shows that there are no significant differences in students' understanding between the scholarship and control groups either right after the program announcement or right before the endline exam. As shown in Figure 4 and Panel B of Table 3, there is some evidence of heterogeneity by baseline test score. Understanding of the programs is slightly higher for the top 15 percent of students relative to the bottom 85 percent of students in both the *Standard* scholarship group and the control group.

Panel A of Figure 5 displays students' expectations of winning the scholarship by baseline overall rank.[19] For students in the *Standard* scholarship group, expectations of receiving the scholarship should increase with baseline overall rank; for students in the *Relative* scholarship group, expectations should not be related to overall rank; and for students in the control group, expectations should be close to zero. Figure 5 generally confirms this pattern, particularly at the time of program announcement. Corresponding regression results in Columns (3) and (4) of Panel B in Table 3 show that students in the scholarship groups were 29-35 percentage points more likely to expect the scholarship. Examining differences across baseline overall rank, those in the top 15 percent in the *Standard* 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.

Panel B of Figure 5 shows students' expectations of winning the scholarship by distribution within bin where baseline bin rank is on the *x*-axis. Columns (3) and (4) of Panel C in Table 3 present corresponding regression results. Immediately after the announcement, both the *Standard* and *Relative* scholarship groups have higher expectations of receiving the scholarship than the control group. In addition, expectations increase with baseline bin rank only for the *Relative* scholarship group, as expected. By the first follow-up, students in both scholarship groups have higher expectations than the control group. However, expectations in both groups are relatively flat across baseline bin rank.

[19] We code a student as expecting the scholarship if he or she answered "very likely" or "likely" to the following question: Based on your current position how much do you think you have a chance of receiving a gift?, or zero otherwise.

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 overall rank (Columns (1) and (2)) as well as normalized test scores (Columns (3) and (4)).[20] 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 also negative in sign, but they were much smaller in magnitude, ranging from -0.04 to -0.12 standard deviations. Although these effects are insignificantly different from zero, we cannot reject that the impacts of the *Standard* and *Relative* scholarships are equal, with p-values of the test for equality of 0.20 and 0.32 for the specifications excluding and including controls, respectively.

Panel A of Figure 6 presents nonparametric plots of final exam scores in each treatment group by baseline overall rank. As shown in the figure, the negative impacts of the *Standard* scholarship are concentrated among those with low baseline rank, 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 decrease 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 overall 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* scholarship is negative and significant, and that on the interaction between *Standard* 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

[20] For each outcome, we present two specifications with and without control variables, but the results are robust to other variations in the set of control variables (available upon request).

following linear regressions are estimated:

$$\begin{aligned}Y_{igsz1} =&\beta_0 + \beta_1 Standard_{gsz} + \beta_2 Relative_{gsz} + \sum_{l=1}^{10} \gamma 1_l Topl + \sum_{l=1}^{10} \gamma 2_l Standard_{gsz} Topl \\ &+ \sum_{l=1}^{10} \gamma 3_l Relative_{gsz} Topl + Y_{igsz0} + \eta_g + \zeta X_{igsz} + \epsilon_{igsz}\end{aligned} \tag{5}$$

Figure 7 presents estimates of ɣ2 and ɣ3, the relative impacts of the *Standard* and *Relative* scholarship treatments 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 some estimates are not statistically significant).

Finally, we examine whether the impacts vary by bin rank – that is, the ranking within the 100-student subgroups used in the *Relative* merit-based scholarship. In Panel B of of Figure 6, we plot performance for the two scholarship groups and control groups across the distribution of bin rank. We do not observe differential impacts for those with higher ranks within these bins, even for the *Relative* scholarship scheme. These results are confirmed in Panel C of Table 4, where we run regressions interacting the treatment groups with being in the top 15 percent of the subgroup at baseline: there is no evidence of heterogeneity by bin rank.

4.3 Intermediate Outcomes

In this subsection we analyze intermediate outcomes in order to explore the 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 overall rank in Panel 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, we do find changes that generally correspond to the overall test score results presented in the previous section (Columns (3) to (7) of Table 5). As shown in Panel A, the point estimates for the *Standard* scholarship program are negative for all four measures, with statistically significant impacts on motivation and self esteem. Column (7) displays impacts on the aggregate index of all four non-cognitive skill measures. The impact of the *Standard* scholarship was -0.14 standard deviations, significant at the 1 percent level. The *Relative* scholarship program also had negative effects on each of the individual measures, although these impacts were smaller and not statistically significant. However, the impact on the index of all four measures is -0.10 standard deviations and is significant at the 10 percent level.

Turning to heterogeneity by baseline score, Panel B of Table 5 shows that the negative impacts of the *Standard* scholarship on non-cognitive skills were concentrated among the bottom 85 percent of students. By contrast, the *Relative* scholarship program had the largest impacts on the top 15 percent of students, although the estimates are not statistically significant.

The impacts of the *Standard* scholarship on these intermediate outcomes correspond to the argument that financial incentives may crowd out intrinsic motivation. We find negative effects on overall motivation and self esteem, with these effects concentrated among those least likely to win the scholarship. There is some suggestive evidence that the *Relative* scholarship had negative effects on non-cognitive skills, but these effects were smaller and did not not appear to be greater among the lowest-performing students.

Columns (8) to (10) 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 efforts.

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.[21] A lack of capacity and resources may explain the null impacts of parental effort. However, the results in Column (10) 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 Long-term impacts

As discussed previously, the *Standard* scholarship program resulted in large negative impacts on non-cognitive skills as well as the score in the incentivized test. In this section, we analyze impacts on test scores in the next semester, 9 months after the incentivized final exam, and show that these impacts did not persist after the incentive 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 presents 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 faded substantially: the average long-term impacts (Columns (3) and (4)) are much smaller in absolute value than the short-term impacts (Columns (1) and (2)) and are no longer statistically significant. In contrast, within this sample, the *Relative* scholarship shows negative impacts of about -0.2 to -0.3 standard deviations in both the short and long term. However, these results are imprecise, reaching marginal significance in only one of the four specifications, and we are hesitant to draw strong conclusions from these somewhat surprising point estimates.

Table A5 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

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

negative effects of the *Standard* scholarship on non-cognitive skills in the short-term, we do not find persistent changes in the long-term, which corresponds to the absence of long-term effects on test scores.

To further examine whether the negative effects were isolated to the incentivized test, we present estimates of the scholarship programs on a simple multiple-choice mathematics test that was included in the follow-up survey. These results are displayed in Table A6. For comparison, Column (1) presents the treatment effects estimates on the math section of the final exam. Although the results suffer from imprecision, they generally correspond with the effects on the full exam from Table 4, particularly the negative effect of the *Standard* scholarship treatment. However, the impacts of this treatment on the endline survey math test, while still negative, are substantially smaller in magnitude.

Together, these results suggest that the negative impacts of the *Standard* scholarship program were largely isolated to the incentivized test.

4.5 Discussion

The previous sections 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. Although the impacts on non-cognitive skills generally correspond to the test score results, we can perform a suggestive analysis to quantify the amount of test score impacts that are driven by changes in non-cognitive skills. We do this by adding follow-up measures of non-cognitive skills into the test score regressions. Of course, these non-cognitive measures were taken as of the follow-up survey and are therefore endogenous. Thus, this analysis should be treated as speculative. As shown in Table A7, we find that test scores are explained at least partially by these control variables: controlling for these variables reduces the impacts on test scores by about 11%. However, much of the impacts remain even after controlling for these variables. This could imply imprecision in our non-cognitive measures; for example, the test score impacts could have been driven by specific types of motivation that our somewhat coarse measure does not capture.

Several other potential mechanisms are worth exploring. First, the incentives could have affected the classroom environment (even though it was a competition across the sub-district, 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 study. Our follow-up survey collected student reports of the classroom environment, allowing us to test for this possibility. As shown in Table A8, 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. While we received no reports of cheating, the final exams were monitored by school officials, not our enumerators. In order to drive the negative results, cheating would have had to occur *less* in the scholarship classrooms, which we view as unlikely. Still, one possibility is that students or teachers in the scholarship classrooms were more likely to prevent cheating, which could explain the decrease in test scores in these classrooms. 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, significantly decreased students' achievement.

4.6 Impacts of Feedback

We now turn to the impacts of the feedback intervention. Panel A, Column (1) of Table 7, presents estimates of the average impacts of feedback on all scholarship groups. The estimated effect is small (about 0.03 standard deviations) and not statistically significant. As shown in Panel B, Column (1), there is no evidence of an effect within either scholarship group, implying that the feedback treatment did not motivate students within these groups.

Because feedback was provided on the students' rank on the midterm exam, we focus our analysis of heterogeneity on the distribution of impacts across midterm exam scores. Panel A of Figure 8 plots final exam score by midterm exam overall rank for the feedback and no-feedback groups. Performance in each group was similar across most of the distribution of midterm scores, although those in the top 15 percent performed slightly better in the feedback group. Panel B repeats these plots for each of the 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 group and the control group. However, as shown in Column (3) of Table 7, Panel B, the impacts in the top 15 percent are not significant for either scholarship group or for the control group.

In Figure A4, we display feedback impacts by bin rank. Although there is some noise across the distribution, within the *Relative* scholarship group there is a small difference between the feedback and no-feedback groups for those with high bin rank. Again, however, this small difference is not statistically significant, as shown in Table A9.

These results imply limited, if any, impacts of the feedback intervention. We present several additional analyses to explore how students reacted to the information provided by the feedback. First, we examine how feedback influenced students' perceptions of performance and expectations of winning the scholarship. In the follow-up survey, we collected students' perceptions of their performance within their classes. Responses were on a scale of 1 to 5, ranging from ”very bad (0-20%)” to ”very good (81-100%)”. Table A10 presents the results of a regression of this perception on baseline scores, midterm scores, and the interactions of scores with the feedback treatment. If feedback changed perceptions, we would expect a stronger correlation between this perception and midterm scores within the feedback treatment. We find that while students' perceptions are related to both baseline and midterm test scores, the relationships do not substantially differ between the feedback and no-feedback groups. This implies that feedback treatment may have conveyed little additional information beyond what students received at the announcement of the scholarship (Figure 2) or outside of the experiment.[22]

Second, we examine whether feedback may have been more valuable when it carried a stronger signal about student progress. All students were told their rankings as of the baseline test, and those with a larger difference between midterm and baseline may have responded more strongly to the feedback. Table A12 presents regressions of final test scores on a dummy for the feedback treatment interacted with the difference between midterm and baseline test scores. We find no evidence that a larger difference between baseline and midterm scores was associated with a larger impact of feedback.

[22] In Table A11, we show that expectations of receiving the scholarship are not related to the feedback treatment.

Taken together, these results suggest that the feedback treatment provided little additional information beyond what the students already knew about their baseline and midterm exam scores. Thus, in environments where there is already a high level of information on performance, additional precise information may have little marginal effect. In the *Relative* treatment, when the feedback contained within-bin information that the students could not readily obtain from their test scores, some students did revise their expectations of winning the scholarship in line with their ranking. However, this did not translate into a substantial improvement in performance. This is consistent with the results from the previous section that showed that the *Relative* scholarship did not motivate students to learn, even among those likely to win.

5 Conclusion

Understanding if, when, and how financial incentives can promote educational achievement remains an important topic of research. While these incentives have been shown to work in some contexts, in others they may not, whether through negative psychological effects, or by otherwise failing to induce productive effort on the part of students.

In this paper we study the impacts of incentives in rural Malawi, a context with low educational achievement and few other learning resources. We evaluate two incentive schemes: a *Standard* scholarship program that provided scholarships for students whose test scores were within the top 15 percent with a novel *Relative* scholarship scheme that provided scholarships for the top students within smaller groups with similar baseline scores. Using an additional randomized intervention, we also estimate 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* 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. We do not find such negative impacts among the *Relative* scholarship group: the point estimates of the impacts are closer to zero and statistically significant, although still negative. We find limited evidence that feedback on ranking influences test scores. We provide suggestive evidence that students did not react to the information provided through the feedback intervention, plausibly because they already had information from their baseline test scores and other results during the semester.

Our results suggest caution in using tournament incentive schemes as a policy to promote learning on contexts such as ours: we find that in the short term, not only did the *Standard* scholarship decrease test scores on average; it also increased inequality by concentrating these decreases on the lowest performing students. Fortunately, these negative effects appear largely isolated to the incentivized test and dissipate in the longer term. These findings, along with our results on non-cognitive skills, correspond to the literature that incentives may not work due to psychological effects (Bénabou and Tirole, 2006; Gneezy, Meier, and Rey-Biel, 2011; Hoff and Pandey, 2014).

The negative distributional effects of tournament incentives may be especially pronounced in environments such as ours, in which students know 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. 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), 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, 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), 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, 13–25.

Barlevy, Gadi and Derek Neal (2012). "Pay for Percentile." *American Economic Review* 102(5), 1805–1831.

Behrman, Jere R. et al. (2015). "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." *Journal of Political Economy* 123(2), 325–364.

Bénabou, Roland and Jean Tirole (2006). "Incentives and Prosocial Behavior." *American Economic Review* 96(5), 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), 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), 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), 90–109.

Cameron, Judy and W. David Pierce (1994). "Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis." *Review of Educational Research* 64(3), 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), 627–668.

Duckworth, Angela Lee and Patrick D. Quinn (2009). "Development and validation of the short grit scale (grit-s)." *Journal of Personality Assessment* 91(2), 166–174.

Fryer, Roland G. (2011). "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *The Quarterly Journal of Economics* 126(4), 1755–1798.

Gilligan, Daniel O et al. (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), 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.

Hoff, Karla and Priyanka Pandey (2014). “Making up people—The effect of identity on performance in a modernizing society.” *Journal of Development Economics* 106, 118–131.

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), 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), 102–138.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz (2007). “Experimental Analysis of Neighborhood Effects.” *Econometrica* 75(1), 83–119.

Kremer, Michael, Edward Miguel, and Rebecca Thornton (2009). “Incentives to Learn.” *Review of Economics and Statistics* 91(3), 437–456.

Lazear, Edward P. and Sherwin Rosen (1981). “Rank-Order Tournaments as Optimum Labor Contracts.” *Journal of Political Economy* 89(5), 841–864.

Lee, David S. (2009). “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” *The Review of Economic Studies* 76(3), 1071–1102.

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), 1243–1265.

Levitt, Steven D. et al. (2012). “The behavioralist goes to school: Leveraging behavioral economics to improve educational performance.” National Bureau of Economic Research.

Li, Tao et al. (2014). “Encouraging classroom peer interactions: Evidence from Chinese migrant schools.” *Journal of Public Economics* 111, 29–45.

Loyalka, Prashant Kumar et al. (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), 804–804.

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

The World Bank (2015). “World development indicators 2015.” 95682. The World Bank, 1–171.

Tran, Anh and Richard Zeckhauser (2012). “Rank as an inherent incentive: Evidence from a field experiment.” *Journal of Public Economics* 96(9), 645–650.

Visaria, Sujata et al. (2016). “Unintended consequences of rewards for student attendance: Results from a field experiment in Indian classrooms.” *Economics of Education Review* 54, 173–184.

Table 1: Sample Composition by Treatment Category

Panel A: Scholarship Treatment (Grade 5-8)

Scholarship Assignment	School-Grades	Students
Standard scholarship	46	2830
Relative scholarship	42	2993
Control	30	1562
Total	118	7385

Panel B: Scholarship Treatment (Grade 5-6 with long-term follow-up)

Scholarship Assignment	School-Grades	Students
Standard scholarship	24	1869
Relative scholarship	24	2000
Control	13	693
Total	61	4562

Panel C: Feedback Treatment (Grade 5-7)

Scholarship Assignment	Feedback Assignment	Students
Standard scholarship	No Feedback	1175
	Feedback	1195
Relative scholarship	No Feedback	1360
	Feedback	1362
Control	No Feedback	510
	Feedback	501
Total		6103

Notes: The scholarship assignment was randomized at the school-grade level with stratification by grade. The feedback assignment was randomized at the individual level.

Table 2: Balance of Baseline Variables Across Treatment Groups

	Scholarship Randomization				Feedback Randomization	
	Control	*Standard* vs.	*Relative* vs.		Feedback vs.	
	Mean	Control	Control	N	No Feedback	N
	(1)	(2)	(3)	(4)	(5)	(6)
Age	14.4	-0.366	-0.300	7385	0.199**	6103
	[3.60]	(0.311)	(0.280)		(0.093)	
Male	0.486	-0.004	-0.028	7385	0.013	6103
	[0.500]	(0.019)	(0.018)		(0.013)	
Ethnic group: Chewa	0.914	-0.033	-0.036	7358	-0.003	6077
	[0.280]	(0.035)	(0.035)		(0.006)	
Household size	7.81	0.228	0.157	7385	0.038	6103
	[1.66]	(0.361)	(0.328)		(0.032)	
Asset index	-0.009	0.0006	0.012	7102	-0.090*	5848
	[1.88]	(0.183)	(0.175)		(0.051)	
Baseline rank(%)	51.5	-0.284	1.89	7342	-0.246	6061
	[27.3]	(3.05)	(3.90)		(0.591)	
Baseline Score	51.5	-1.78	-1.93	7342	-0.075	6061
	[8.59]	(1.28)	(1.56)		(0.154)	
Attendance	0.863	-0.011	-0.021	7385	0.005	6103
	[0.196]	(0.018)	(0.018)		(0.005)	
Study hours per week	16.8	-1.00	-0.818	7308	0.163	6031
	[16.4]	(0.865)	(0.871)		(0.374)	
Motivation to study	4.53	-0.054	0.016	7374	-0.0003	6092
	[0.789]	(0.065)	(0.055)		(0.021)	
Self-esteem	2.67	-0.027	-0.019	7368	0.011	6087
	[0.338]	(0.023)	(0.024)		(0.007)	
Conscientious	3.58	-0.028	0.045	7370	0.002	6089
	[0.600]	(0.068)	(0.066)		(0.015)	
Grit	3.21	-0.050*	-0.029	7368	0.021*	6087
	[0.450]	(0.026)	(0.028)		(0.012)	
Teacher effort index	-0.003	0.112	0.202	7364	0.002	6083
	[1.000]	(0.144)	(0.127)		(0.023)	
Parental Effort Index	0.001	-0.070	-0.050	7281	0.052**	6024
	[1.00]	(0.076)	(0.065)		(0.025)	

Notes: Column 1 reports means of baseline variables for subjects assigned to the control group. Columns 2 and 3 report mean differences between the scholarship treatment groups and the control group. Column 5 reports the mean difference between the feedback treatment and the control group. Standard deviations are in brackets, and standard errors, clustered at the school-grade level, are in parentheses. The asset index is constructed as the 1st principal component of variables indicating the ownership of 26 assets. Teacher and parental effort indeces are aggregates of the seven and four measures, respectively. Both teacher and parental effort isgenerated by taking the average of the standardized measures where the mean and standard deviation in the control group is used in the standardization. The resulting index is also standardized relative to the control group, so that it has a mean of 0 and a standard deviation of 1. Grit and conscientiousness questions were measured on a five-point scale, and self-esteem questions were measured on a four-point scale. We take the simple average of scores for all questions in a category to form our measures. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 3: Understanding and Expectations

	Sample: Grade 5-8			
	Understanding		Expectation	
	After Announcement	1st Follow-up	After Announcement	1st Follow-up
	(1)	(2)	(3)	(4)
Panel A: Average Treatment effects				
Standard	-0.009	-0.021	0.301***	0.442***
	(0.023)	(0.023)	(0.057)	(0.043)
Relative	0.036	-0.028	0.358***	0.405***
	(0.022)	(0.024)	(0.066)	(0.044)
R-Squared	0.038	0.092	0.097	0.135
P-value: *Std* = *Rel*	0.007	0.800	0.330	0.112
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-0.007	-0.019	0.231***	0.407***
	(0.026)	(0.023)	(0.059)	(0.046)
Relative	0.041*	-0.011	0.386***	0.409***
	(0.024)	(0.025)	(0.066)	(0.046)
Std. x Top 15%	-0.015	-0.018	0.485***	0.212***
	(0.025)	(0.035)	(0.084)	(0.045)
Rel. x Top 15%	-0.040*	-0.107***	-0.135	-0.028
	(0.022)	(0.029)	(0.083)	(0.054)
Top 15%	0.056***	0.091***	0.046	0.013
	(0.020)	(0.019)	(0.042)	(0.037)
R-Squared	0.047	0.098	0.157	0.145
Panel C: Hegerogeneous treatment effects by bin rank				
Standard	-0.011	-0.025	0.290***	0.443***
	(0.023)	(0.023)	(0.057)	(0.044)
Relative	0.033	-0.030	0.294***	0.394***
	(0.021)	(0.024)	(0.066)	(0.045)
Std. x Subg. Top 15%	0.008	0.025	0.080*	-0.004
	(0.017)	(0.026)	(0.044)	(0.041)
Rel. x Subg. Top 15%	0.015	0.017	0.394***	0.067
	(0.016)	(0.025)	(0.063)	(0.042)
Controls	Yes	Yes	Yes	Yes
N	5617	5851	5594	5750
R-Squared	0.038	0.092	0.136	0.136
Mean of Dep. Var.	0.924	0.636	0.356	0.579

Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects, zone fixed effects, age, ethnic group, household size, and a household asset index. Expectation is a dummy variable equal to one if a student answered very likely or likely to the following question: Based on your current position, how much do you think you have a chance of receiving a gift?, or zero otherwise * 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)	
	(1)	(2)	(3)	(4)
Panel A: Average Treatment effects				
Standard	-7.402**	-7.368*	-0.265*	-0.266*
	(3.620)	(3.868)	(0.135)	(0.145)
Relative	-2.516	-4.730	-0.046	-0.129
	(4.668)	(4.404)	(0.186)	(0.174)
R-Squared	0.234	0.305	0.251	0.323
P-value: *Std = Rel*	0.250	0.447	0.211	0.344
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-8.961**	-8.682**	-0.315**	-0.308**
	(3.833)	(4.138)	(0.138)	(0.152)
Relative	-1.543	-4.016	0.013	-0.080
	(4.987)	(4.769)	(0.192)	(0.183)
Std. x Top 15%	9.697*	7.507	0.315	0.238
	(5.540)	(5.316)	(0.248)	(0.236)
Rel. x Top 15%	-5.696	-4.348	-0.337	-0.275
	(7.370)	(6.057)	(0.302)	(0.260)
Top 15%	2.777	3.847	0.075	0.109
	(5.111)	(4.730)	(0.231)	(0.216)
R-Squared	0.244	0.312	0.260	0.329
Panel C: Hegerogeneous treatment effects by bin rank				
Standard	-7.404**	-7.360*	-0.266*	-0.266*
	(3.727)	(3.982)	(0.139)	(0.151)
Relative	-2.234	-4.423	-0.030	-0.111
	(4.761)	(4.527)	(0.190)	(0.180)
Std. x Subg. Top 15%	0.069	0.038	0.008	0.001
	(2.201)	(2.270)	(0.087)	(0.089)
Rel. x Subg. Top 15%	-1.731	-1.877	-0.099	-0.104
	(2.166)	(2.227)	(0.087)	(0.088)
Controls	No	Yes	No	Yes
N	6586	6323	6586	6323
R-Squared	0.234	0.305	0.251	0.323
Mean of Dep. Var.	51.346	51.489	-0.154	-0.146

Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects and the baseline value of the outcome variable. Additional controls include zone fixed effects, age, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 5: Intermediate Outcomes

	Sample: Grade 5-8									
	Student input		Non-cognitive skills					Teacher and parental response		
	Attendance	Study Hours	Motivation to study hard	Self esteem	Grit	Conscien tiousness	Non-cognitive skill index	Teacher effort index	Parental effort	Parents mentioned scholarship
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Average Treatment effects										
Standard	0.024*	-0.970	-0.071**	-0.030*	-0.034	-0.045	-0.137***	-0.017	-0.037	0.126**
	(0.013)	(1.036)	(0.035)	(0.017)	(0.023)	(0.032)	(0.052)	(0.045)	(0.085)	(0.064)
Relative	0.009	-1.562	-0.036	-0.028	-0.027	-0.027	-0.101*	-0.027	0.022	0.087
	(0.015)	(1.158)	(0.039)	(0.017)	(0.023)	(0.034)	(0.055)	(0.040)	(0.083)	(0.071)
R-Squared	0.193	0.076	0.022	0.050	0.049	0.080	0.116	0.091	0.044	0.038
P-value: *Std* = *Rel*	0.253	0.523	0.239	0.911	0.724	0.529	0.492	0.812	0.246	0.544
Panel B: Heterogeneous treatment effects by overall rank										
Standard	0.024*	-0.961	-0.090**	-0.035*	-0.039*	-0.059*	-0.175***	-0.017	-0.043	0.081
	(0.013)	(1.121)	(0.038)	(0.018)	(0.023)	(0.031)	(0.055)	(0.047)	(0.090)	(0.067)
Relative	0.010	-1.432	-0.049	-0.026	-0.011	-0.021	-0.089	-0.026	0.047	0.107
	(0.016)	(1.237)	(0.042)	(0.018)	(0.024)	(0.031)	(0.057)	(0.042)	(0.087)	(0.069)
Std. x Top 15%	-0.008	0.093	0.116*	0.032	0.029	0.083	0.227*	-0.001	0.030	0.278**
	(0.023)	(1.721)	(0.063)	(0.039)	(0.051)	(0.094)	(0.135)	(0.056)	(0.111)	(0.108)
Rel. x Top 15%	-0.021	-0.977	0.067	-0.016	-0.098**	-0.034	-0.082	-0.011	-0.157	-0.054
	(0.027)	(2.049)	(0.066)	(0.034)	(0.040)	(0.092)	(0.122)	(0.056)	(0.111)	(0.120)
Top 15%	0.043***	1.511	-0.004	0.024	0.090***	0.026	0.087	0.027	0.131	-0.230***
	(0.016)	(1.526)	(0.050)	(0.030)	(0.029)	(0.083)	(0.102)	(0.042)	(0.093)	(0.086)
N	7085	5242	5754	5842	5842	5844	5850	5838	5778	5848
R-Squared	0.194	0.076	0.023	0.052	0.054	0.083	0.121	0.091	0.046	0.042
Mean of Dep. Var.	0.756	14.526	4.298	2.719	3.259	3.674	-0.131	4.006	-0.026	3.409

Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects and the baseline value of the outcome variable. Additional controls include zone fixed effects, age, ethnic group, household size, and a household asset index. Parental and teacher effort indices are constructed by the way explained in Table 2. * 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 effects				
Standard	-0.463**	-0.517**	-0.238	-0.226
	(0.193)	(0.247)	(0.159)	(0.143)
Relative	-0.191	-0.375	-0.295	-0.319**
	(0.245)	(0.277)	(0.187)	(0.158)
R-Squared	0.038	0.317	0.014	0.201
P-value: *Std* = *Rel*	0.243	0.471	0.700	0.513
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-0.473**	-0.547**	-0.233	-0.217
	(0.224)	(0.266)	(0.154)	(0.148)
Relative	-0.132	-0.324	-0.303*	-0.326**
	(0.292)	(0.297)	(0.170)	(0.158)
Std. x Top 15%	0.209	0.182	-0.063	-0.072
	(0.273)	(0.294)	(0.286)	(0.265)
Rel. x Top 15%	-0.441	-0.275	0.016	-0.003
	(0.355)	(0.325)	(0.318)	(0.278)
Top 15%	0.124	0.129	0.097	0.160
	(0.249)	(0.262)	(0.213)	(0.197)
Controls	No	Yes	No	Yes
N	4040	3860	2615	2505
R-Squared	0.240	0.322	0.154	0.202
Mean of Dep. Var.	-0.272	-0.264	0.039	0.045

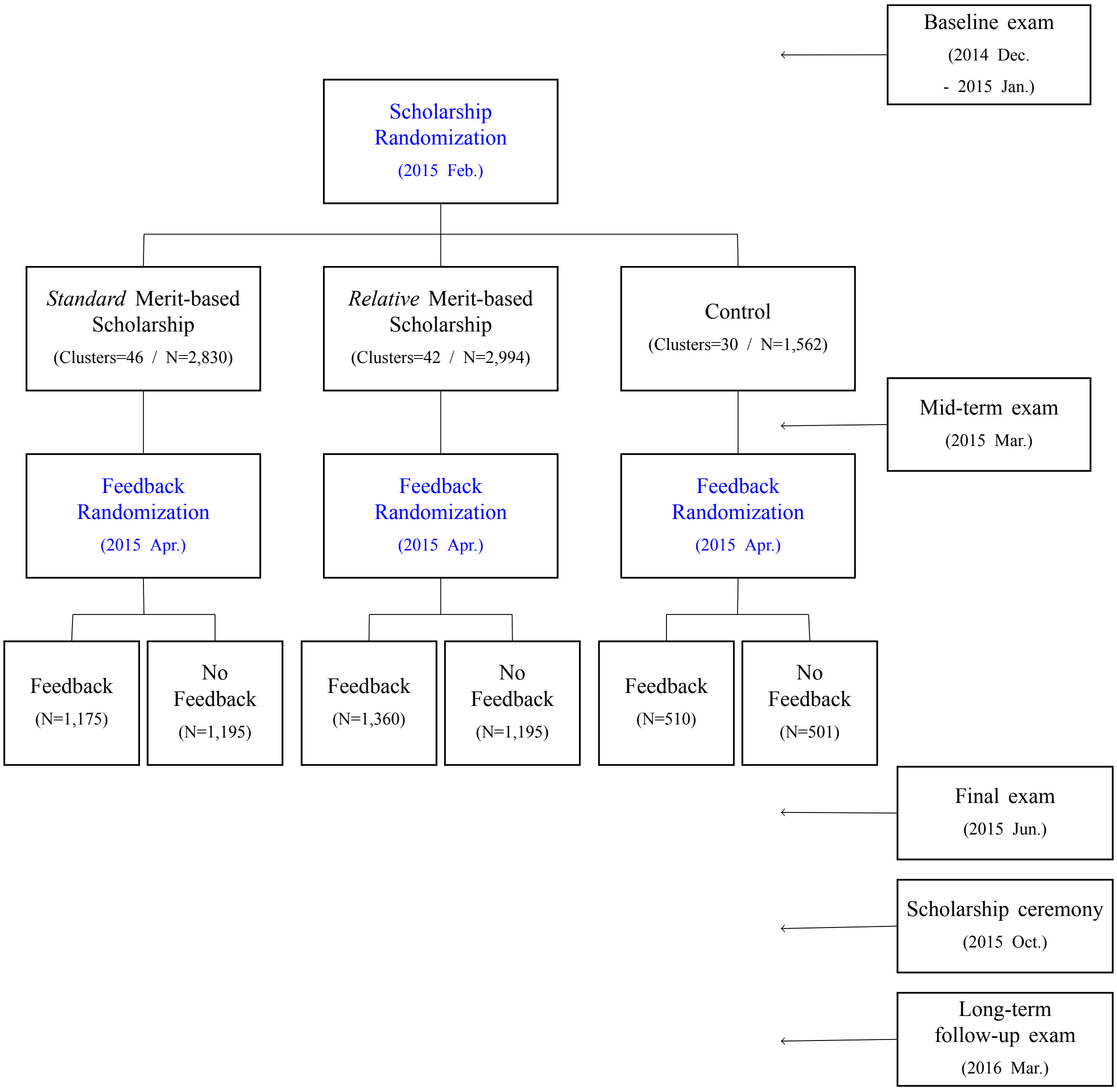
Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects and the baseline value of the outcome variable. Additional controls include zone fixed effects, age, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 7: Feedback effect: Test Score Impacts

	Sample: Grade 5-7		
	Final exam		
	All	Mid-term Top 15%	Mid-term Bot 85%
	(1)	(2)	(3)
Panel A			
Feedback	0.028	0.065	0.014
	(0.023)	(0.052)	(0.028)
R-Squared	0.308	0.242	0.220
Panel B			
Feedback	0.046	0.084	0.035
	(0.064)	(0.080)	(0.081)
Standard	-0.324	-0.185	-0.277
	(0.202)	(0.242)	(0.174)
Relative	-0.207	-0.109	-0.108
	(0.227)	(0.248)	(0.211)
Std. x FB	-0.015	-0.006	-0.018
	(0.072)	(0.102)	(0.090)
Rel. x FB	-0.027	-0.036	-0.031
	(0.073)	(0.134)	(0.090)
N	5159	1057	4102
R-Squared	0.317	0.247	0.230
Mean of Dep. Var.	-0.186	0.846	-0.452

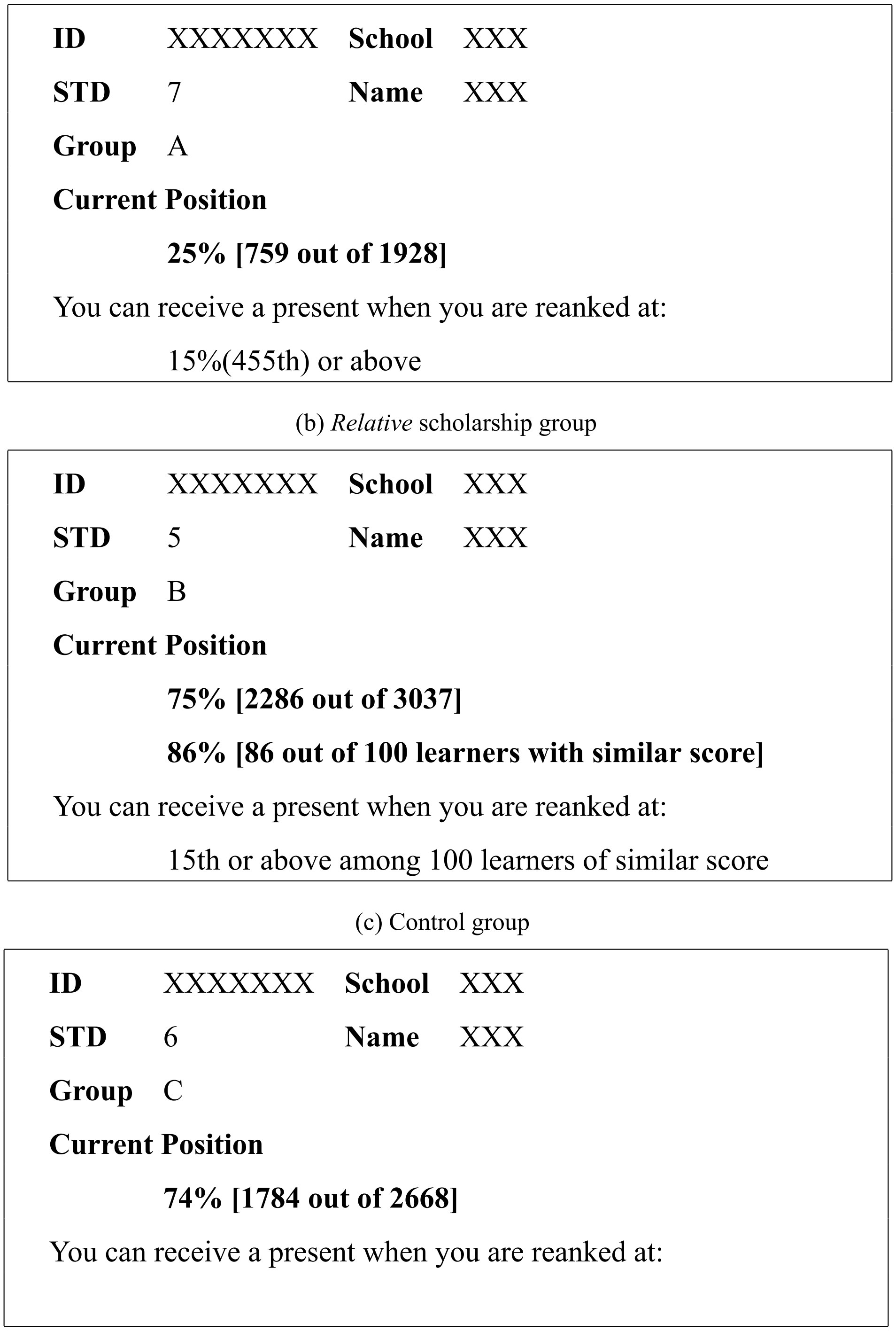
Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects and the baseline value of the outcome variable. Additional controls include zone fixed effects, age, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

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 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 figure.

Figure 2: Scholarship Randomization result announcement note



(a) *Standard* 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* 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		

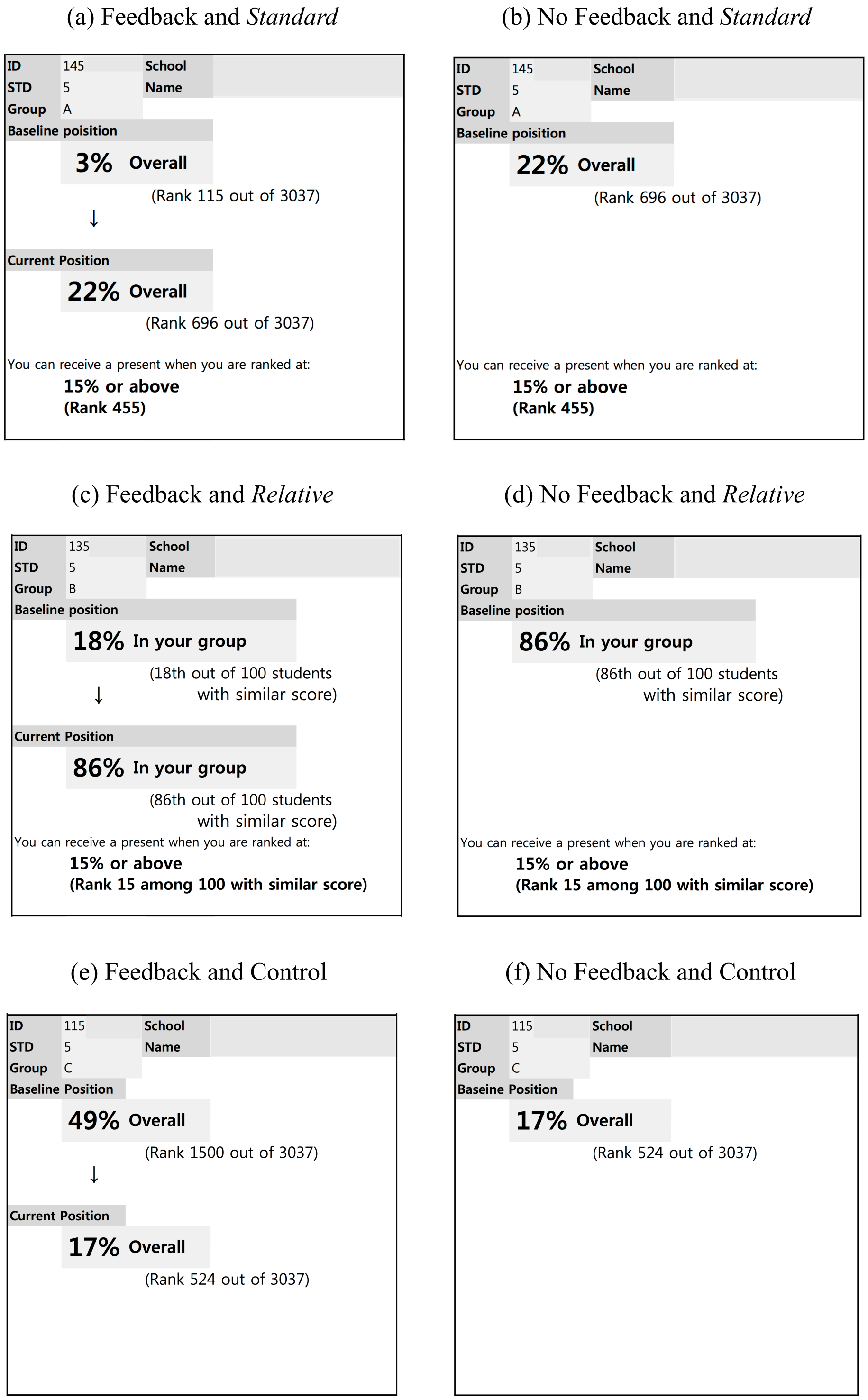
Current Position

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

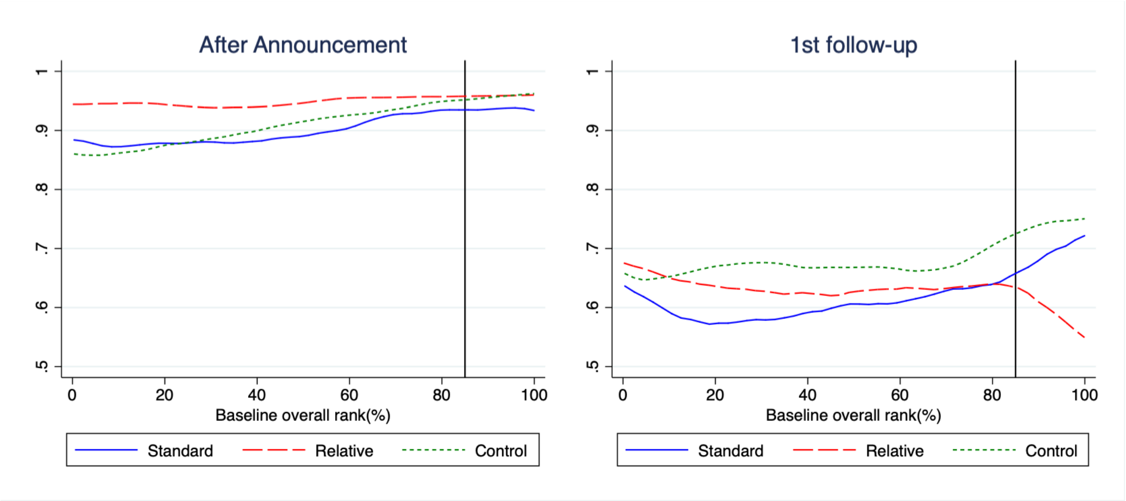
ID 115 **School**
STD 5 **Name**
Group C
Baseline Position
49% Overall
(Rank 1500 out of 3037)
↓
Current Position
17% Overall
(Rank 524 out of 3037)

(f) No Feedback and Control

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

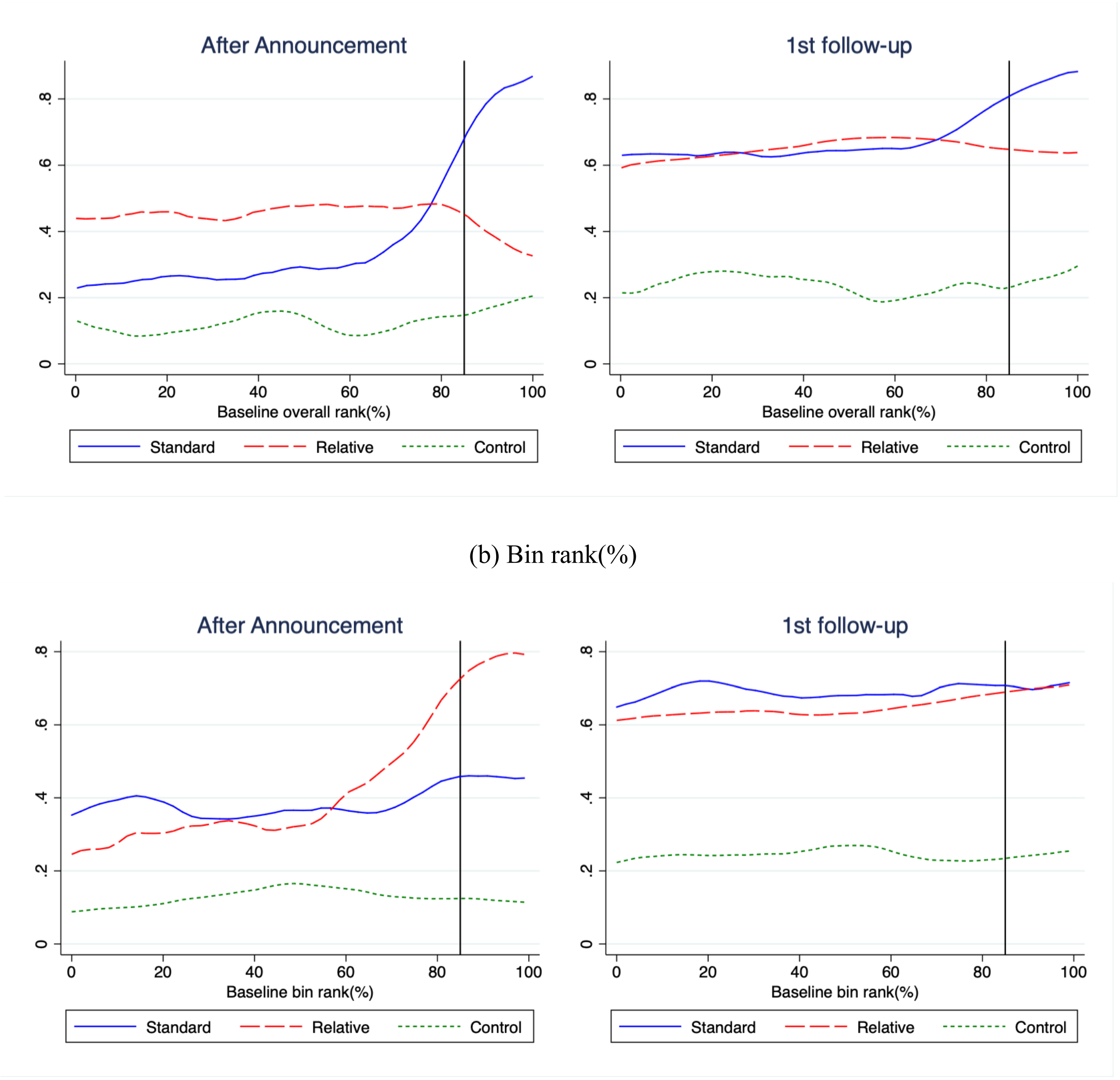
Note: This figure displays the feedback notes that students received in the second semester. The left column presents feedback notes given to the feedback treatment group and the right column presents feedback notes given to the control group. The feedback treatment group received information on their rank in the baseline and midterm exam while the control group received information only on the baseline exam. Panels A and B, C and D, and E and F display the feedback provided for the *Standard* scholarship group, the *Relative* scholarship group, and the control group, respectively.

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



Note: This figure presents students' levels of understanding measured by the percent of questions answered correctly on quizzes immediately after the scholarship announcements and at the time of the follow-up surveys.

Figure 5: Expectation of the scholarship



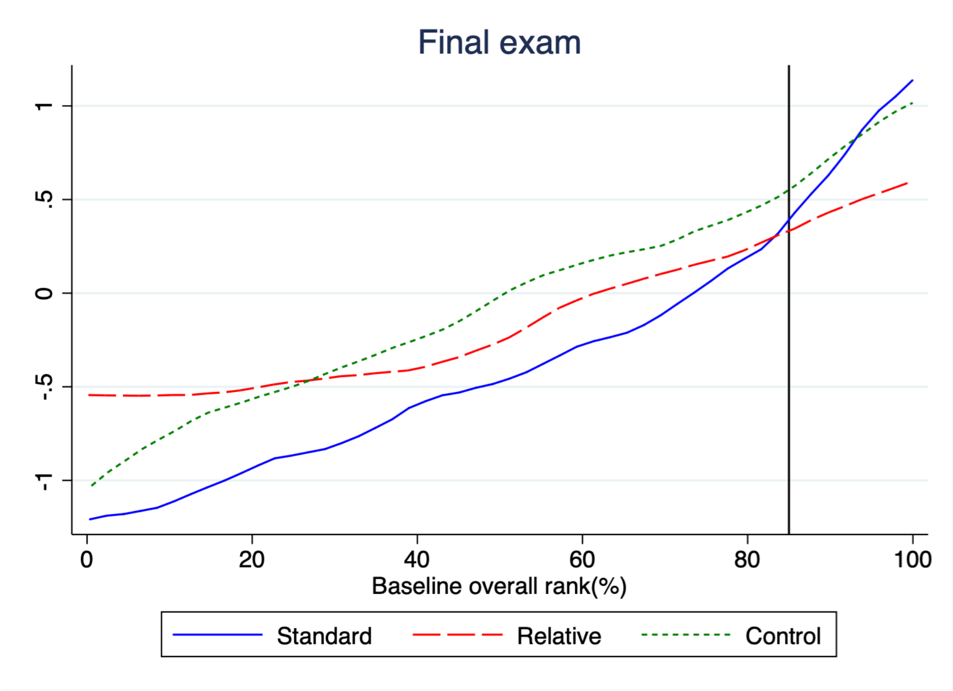
(a) Overall rank(%)

(b) Bin rank(%)

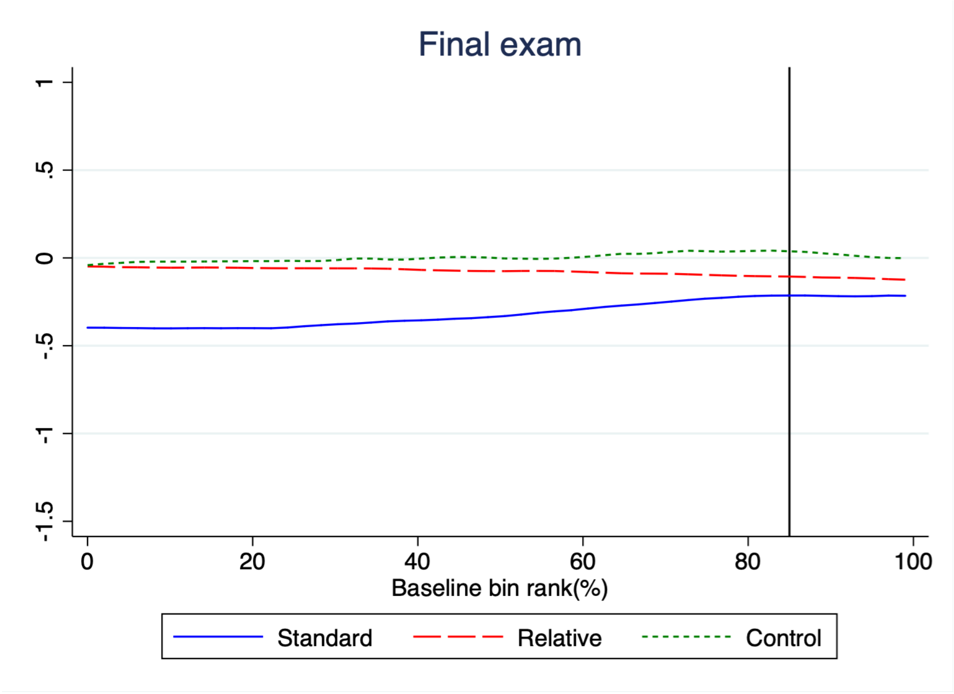
Note: This figure presents students' expectations of winning the scholarship immediately after the scholarship announcements and at the time of the follow-up surveys.

Figure 6: Exam scores at follow-up by Baseline Rank

(a) Overall rank(%)

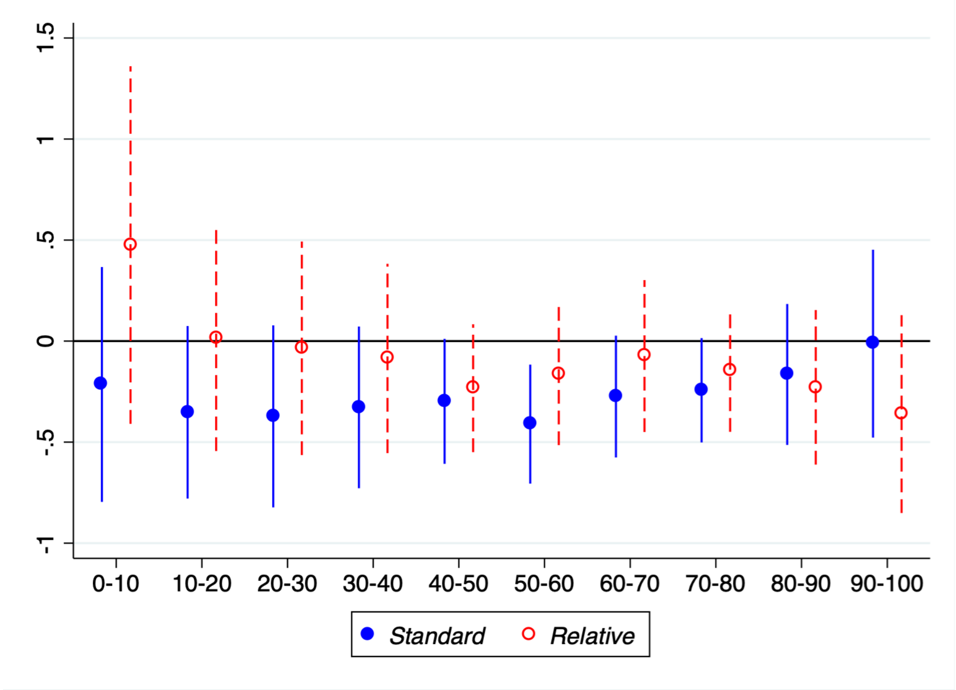


(b) Bin rank(%)



Note: This figure presents average follow-up exam scores by baseline rank for each study group.

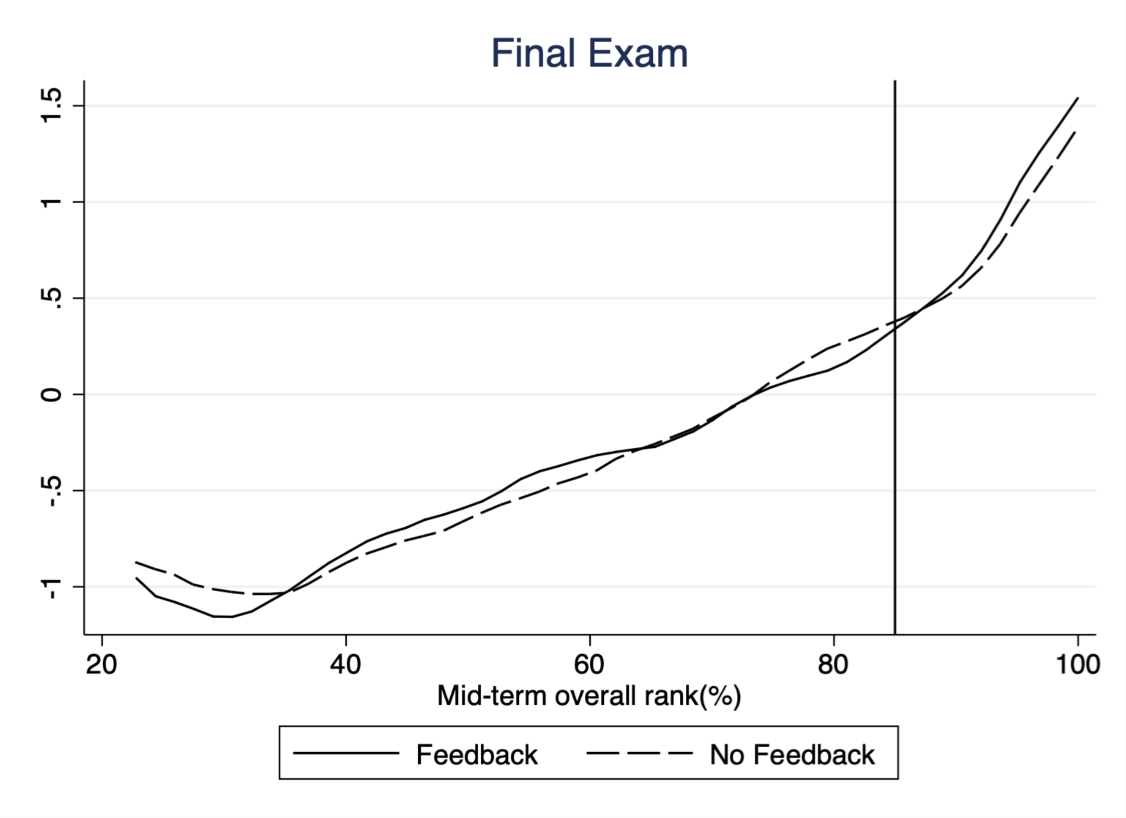
Figure 7: Coefficient of scholarship program effect



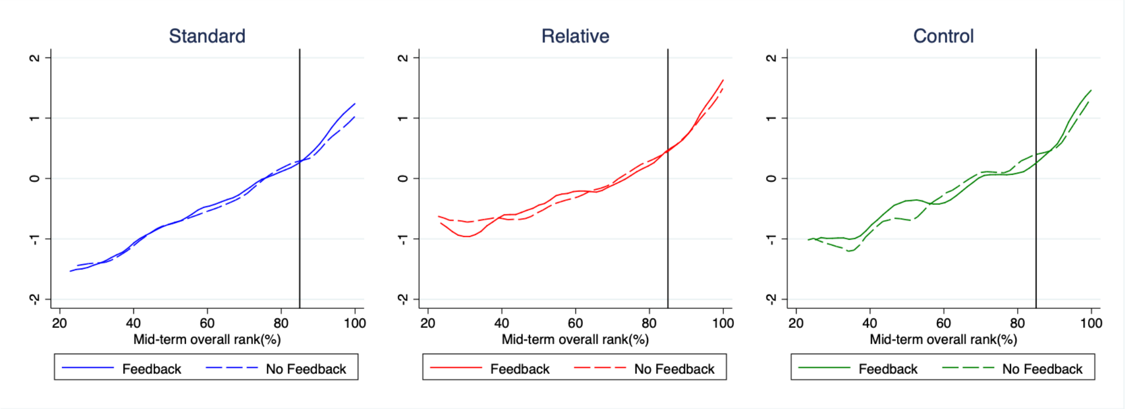
Note: This figure presents coefficients with 95% confidence intervals from equation (5). The x-axis presents the baseline decile rank of the students. Navy and crimson markers present coefficients of the *Standard* scholarship and *Relative* scholarship effects, respectively.

Figure 8: Feedback effect on follow-up exam score by mid-term rank

(a) Whole sample



(b) By treatment group (Overall rank(%)



Note: This figure presents average follow-up exam scores by mid-term Overall rank. Panel A presents the results for all students, while Panel B presents the results by scholarship treatment status.

Appendices

A Appendix Tables and Figures

A.1 Attrition (Tables A1 - A3)

This section presents additional analysis of attrition. As discussed in the main text, although attrition was largely balanced across treatment groups in the follow-up survey and second final exam, there is some evidence of differential attrition as of the first final exam: those in the *Relative* scholarship group were 2.9 percent more likely to take the final exam, relative to 88.4 percent in the control group. Here we focus on this differential attrition and its potential to influence our treatment effect estimates.

We first construct bounds following the method of D. S. Lee (2009). Because both the Standard and Relative scholarship groups had lower attrition than the control group, we trim these groups by the fraction of ”excess” observations in these groups. The lower (upper) bound is constructed by trimming the highest (lowest) final exam scores and running the impact regressions. As shown in Table A2, these bounds are relatively tight. For the *Relative* scholarship group, where we observed a significant difference in attrition, the impacts on exam rank are -3.24 to -0.97 percentage points, and the impacts on normalized exam scores are -0.12 to 0.01 standard deviations. None of these estimates are statistically significant.

Because heterogeneity by baseline exam score is a key part of our analysis, we also examine whether attriters in each scholarship treatment group have different baseline test scores. We examine this by regressing attrition as of the final exam on the scholarship treatment groups, the baseline score (either a continuous variable or an indicator for the top 15 percent), and the interaction of the scholarship treatment groups and the baseline score. The results of these regressions are shown in Table A3. As shown in Columns (3) and (5), there is no evidence that attriters in the scholarship treatment groups had different baseline scores than those in the control group.

Table A1: 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	-0.019	0.022	0.017	-0.012	0.023	0.043	0.036
	(0.021)	(0.017)	(0.015)	(0.032)	(0.019)	(0.014)	(0.032)	(0.039)
Relative	0.008	-0.025	0.029**	-0.003	-0.014	0.027*	0.025	0.043
	(0.023)	(0.017)	(0.014)	(0.035)	(0.021)	(0.015)	(0.034)	(0.033)
N	7385	7385	7385	4562	4562	4562	4393	4393
R-Squared	0.006	0.003	0.004	0.002	0.000	0.001	0.001	0.001
Mean of Dep. Var.	0.877	0.827	0.896	0.859	0.836	0.891	0.629	0.568
Panel B								
Feedback		0.003	-0.003		-0.005	-0.006	0.015	0.005
		(0.008)	(0.007)		(0.010)	(0.007)	(0.013)	(0.014)
N		6103	6103		4562	4562	4393	4393
R-Squared		0.000	0.000		0.000	0.000	0.001	0.000
Mean of Dep. Var.		0.836	0.889		0.836	0.891	0.629	0.568

Notes: Standard errors, in parentheses, are clustered at the school-grade level. All specifications include grade fixed effects. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A2: Lee (2009) Bounds of Main Test Score Estimates

	Exam Rank			Exam Score (Norm)		
	Main	Lower Bound	Upper Bound	Main	Lower Bound	Upper Bound
	(1)	(2)	(3)	(4)	(5)	(6)
Merit	-7.402**	-8.321**	-6.707*	-0.265*	-0.303**	-0.232*
	(3.671)	(3.671)	(3.671)	(0.138)	(0.138)	(0.138)
Relative merit	-2.516	-3.724	-1.374	-0.046	-0.123	0.003
	(4.730)	(4.730)	(4.730)	(0.187)	(0.187)	(0.187)
N	6586	6586	6586	6586	6586	6586

Notes: Lower (upper) bounds are computed by trimming the highest (lowest) observations in the scholarship treatment groups. The fraction of trimmed observations equals the relative difference in attrition, computed from Column 3 of Table A2. Standard errors are in parentheses and are constructed using 500 bootstrap samples, where classes are sampled to account for clustering. All specifications include grade fixed effects and the baseline value of the outcome variable. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A3: Attrition by Scholarship Treatment and Baseline Test Score

		Baseline Variable			
		Baseline Score		Top 15 percent	
	(1)	(2)	(3)	(4)	(5)
Merit	0.022		0.020		0.023
	(0.015)		(0.101)		(0.015)
Relative merit	0.029**		-0.027		0.026*
	(0.014)		(0.106)		(0.014)
Baseline		0.004***	0.004*	0.036***	0.034
		(0.001)	(0.002)	(0.009)	(0.022)
Baseline*Standard			0.000		-0.008
			(0.002)		(0.026)
Baseline*Relative			0.001		0.010
			(0.002)		(0.027)
N	7385	7342	7342	7385	7385

Notes: Lower (upper) bounds are computed by trimming the highest (lowest) observations in the scholarship treatment groups. The fraction of trimmed observations equals the relative difference in attrition, computed from Column 3 of Table A2. Standard errors are in parentheses and are constructed using 500 bootstrap samples, where classes are sampled to account for clustering. All specifications include grade fixed effects and the baseline value of the outcome variable. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A4: Balance of Baseline Variables Across Treatment Groups: Long-term Sample (Grades 5 and 6)

		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	-0.072	0.164	4562	0.308**	4562
	[5.64]	[4.89]	(0.313)	(0.301)		(0.118)	
Male	0.465	0.488	-0.025	-0.027	4562	0.009	4562
	[0.499]	[0.500]	(0.025)	(0.022)		(0.016)	
Ethnic group: Chewa	0.888	0.947	-0.070	-0.068*	4541	0.005	4541
	[0.316]	[0.225]	(0.043)	(0.040)		(0.006)	
Household size	7.89	7.78	0.043	0.216	4562	0.031	4562
	[1.51]	[1.73]	(0.558)	(0.507)		(0.035)	
Asset index	-0.007	-0.225	0.329*	0.190	4365	-0.041	4365
	[1.92]	[1.76]	(0.179)	(0.160)		(0.060)	
Baseline rank(%)	52.6	53.0	-0.792	-0.122	4528	-0.585	4528
	[27.9]	[27.0]	(4.43)	(5.66)		(0.692)	
Baseline Score	48.3	48.2	0.049	0.072	4528	-0.170	4528
	[7.86]	[7.29]	(1.22)	(1.74)		(0.171)	
Attendance	0.833	0.830	0.006	0.001	4562	0.003	4562
	[0.198]	[0.210]	(0.024)	(0.023)		(0.006)	
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	4.47	4.46	-0.035	0.065	4552	-0.009	4552
	[0.853]	[0.817]	(0.090)	(0.076)		(0.025)	
Self-esteem	2.63	2.61	0.015	0.017	4550	0.013	4550
	[0.333]	[0.333]	(0.034)	(0.032)		(0.008)	
Conscientious	3.52	3.45	0.036	0.125	4552	0.019	4552
	[0.584]	[0.591]	(0.107)	(0.103)		(0.016)	
Grit	3.15	3.14	-0.007	0.011	4550	0.022	4550
	[0.423]	[0.432]	(0.038)	(0.039)		(0.014)	
Teacher effort index	0.124	-0.034	0.096	0.270	4548	0.022	4548
	[0.981]	[1.05]	(0.226)	(0.202)		(0.028)	
Parental Effort Index	-0.107	-0.079	-0.061	-0.007	4506	0.037	4506
	[1.11]	[1.02]	(0.116)	(0.101)		(0.028)	

Notes: Columns 1 and 2 report 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 reports the mean difference between the feedback treatment and the control group. The indices for parental and teacher effort and non-cognitive skills are constructed by the way explained in Table 2. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A5: Long term Intermediate Outcomes

	Sample: Grade 5-6								
	1st Follow-up					2nd Follow-up			
	Student input		Non-cognitive skills			Student input	Non-cognitive traits		
	Attendance	Study Hours	Motivation to study hard	Self esteem	Conscientious-ness	Study Hours	Motivation to study hard	Self esteem	Conscientious-ness
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Average Treatment effects									
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: *Std = Rel*	0.345	0.095	0.312	0.806	0.769	0.378	0.309	0.801	0.313
Panel B: Heterogeneous treatment effects by overall rank									
Standard	0.016	-2.321**	-0.107*	-0.067***	-0.053	1.469	-0.028	-0.028	-0.096*
	(0.019)	(1.121)	(0.054)	(0.022)	(0.044)	(1.288)	(0.045)	(0.033)	(0.056)
Relative	0.005	-3.853***	-0.065	-0.064***	-0.023	1.168	0.024	-0.034	-0.111*
	(0.020)	(1.181)	(0.059)	(0.024)	(0.044)	(1.014)	(0.045)	(0.034)	(0.056)
Std. x Top 15%	0.005	0.013	0.084	0.077*	0.075	3.506	0.232	0.086	0.061
	(0.032)	(2.474)	(0.093)	(0.044)	(0.079)	(4.465)	(0.166)	(0.054)	(0.108)
Rel. x Top 15%	-0.028	-1.237	0.032	0.022	-0.034	-1.616	0.155	0.074	-0.101
	(0.038)	(2.606)	(0.094)	(0.045)	(0.086)	(1.893)	(0.165)	(0.050)	(0.117)
Top 15%	0.037	0.952	-0.009	-0.018	0.010	0.447	-0.164	-0.031	0.050
	(0.027)	(2.257)	(0.082)	(0.039)	(0.068)	(1.605)	(0.160)	(0.031)	(0.082)
N	4353	3241	3591	3631	3633	2410	2596	2597	2599
R-Squared	0.190	0.039	0.019	0.048	0.057	0.008	0.022	0.053	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 parentheses. Standard errors clustered at the school-grade level. All specifications include grade fixed effects, zone fixed effects, baseline value of dependent variables, 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 A6: Impact on scores with no incentive

	Sample: Grade 5-8	
	Math scores	
	Final Exam	Survey
	(1)	(2)
Panel A: Average Treatment effects		
Standard	-0.207	-0.021
	(0.139)	(0.019)
Relative	0.106	-0.006
	(0.157)	(0.020)
R-Squared	0.078	0.321
P-value: *Std = Rel*	0.008	0.311
Panel B: Heterogeneous treatment effects by overall rank		
Standard	-0.203	-0.019
	(0.154)	(0.019)
Relative	0.102	-0.009
	(0.173)	(0.021)
Std. x Top 15%	-0.078	-0.014
	(0.176)	(0.025)
Rel. x Top 15%	-0.168	-0.021
	(0.245)	(0.027)
Top 15%	0.554***	0.148***
	(0.148)	(0.018)
N	6317	5857
R-Squared	0.104	0.359
Mean of Dep. Var.	0.032	0.548

Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include baseline final exam score, grade fixed effects, zone fixed effects, age, ethnic group, household size, and a household asset index. The survey-based math test was not conducted at baseline. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A7: Test score impacts (Noncognitive skills controlled)

	Sample: Grade 5-8			
	Exam Rank		Exam score (Norm)	
	(1)	(2)	(3)	(4)
Panel A: Average Treatment effects				
Standard	-6.739*	-7.010*	-0.234*	-0.245*
	(3.639)	(3.856)	(0.137)	(0.146)
Relative	-2.774	-5.208	-0.049	-0.139
	(4.717)	(4.416)	(0.190)	(0.177)
R-Squared	0.253	0.317	0.269	0.334
P-value: *Std* = *Rel*	0.349		0.294	
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-8.554**	-8.554**	-0.298**	-0.299*
	(3.881)	(4.133)	(0.141)	(0.153)
Relative	-2.066	-4.677	-0.001	-0.098
	(5.075)	(4.790)	(0.198)	(0.185)
Std. x Top 15%	10.724**	8.587*	0.369	0.294
	(5.095)	(5.071)	(0.224)	(0.222)
Rel. x Top 15%	-3.790	-2.573	-0.250	-0.192
	(6.687)	(5.620)	(0.273)	(0.243)
Top 15%	-31.711*	-42.303***	-1.705**	-2.089***
	(17.202)	(14.314)	(0.708)	(0.622)
Demographic cont.	No	Yes	No	Yes
N	5829	5596	5829	5596
R-Squared	0.267	0.324	0.284	0.342
Mean of Dep. Var.	52.316	52.437	-0.123	-0.117

Notes: Notes: Standard errors, in parentheses, are clustered at the school-grade level. All specifications include grade fixed effects. Demographic controls include age, race, household size, and a household asset index. Noncognitive traits include motivation, self esteem, grit, and conscientiousness. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table A8: 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 from friends	Classroom competi-tiveness index
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Average Treatment effects						
Standard	0.072	-0.038	0.084	0.078	0.040	0.064
	(0.101)	(0.062)	(0.061)	(0.065)	(0.067)	(0.080)
Relative	-0.220	0.013	-0.051	0.008	-0.053	-0.087
	(0.134)	(0.060)	(0.064)	(0.067)	(0.066)	(0.080)
R-Squared	0.086	0.018	0.022	0.008	0.009	0.038
P-value: *Std = Rel*	0.012	0.164	0.005	0.184	0.115	0.002
Panel B: Heterogeneous treatment effects by overall rank						
Standard	0.118	-0.046	0.084	0.098	0.039	0.079
	(0.109)	(0.072)	(0.052)	(0.065)	(0.077)	(0.084)
Relative	-0.250*	-0.008	-0.047	0.048	-0.064	-0.093
	(0.148)	(0.070)	(0.060)	(0.065)	(0.068)	(0.086)
Std. x Top 15%	-0.300*	0.057	-0.001	-0.129	0.020	-0.094
	(0.162)	(0.101)	(0.144)	(0.157)	(0.166)	(0.128)
Rel. x Top 15%	0.078	0.113	-0.021	-0.223	0.087	0.016
	(0.180)	(0.111)	(0.151)	(0.191)	(0.196)	(0.155)
Top 15%	0.239	-0.052	0.003	0.117	-0.151	0.035
	(0.155)	(0.100)	(0.136)	(0.153)	(0.137)	(0.123)
Baseline Score	-0.002	0.000	-0.001	0.002	0.001	0.000
	(0.006)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Demographic cont.	Yes	Yes	Yes	Yes	Yes	Yes
N	2680	2679	2672	2674	2680	2682
R-Squared	0.090	0.019	0.022	0.009	0.010	0.038
Mean of Dep. Var.	3.755	4.074	3.888	3.829	4.095	-0.010

Notes: Robust standard errors in parentheses. Standard errors are clustered at the the school-grade level. All specifications include grade fixed effects and zone fixed effects. Demographic controls include age, race, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; *** at 0.01.

Table A9: Feedback effect: Test Score Impacts (Bin rank(%))

	Sample: Grade 5-7		
	Final exam		
	All	Mid-term Subgroup Top 15%	Mid-term Subgroup Bot 85%
	(1)	(2)	(3)
Panel A			
Feedback	0.028	0.112	0.015
	(0.023)	(0.081)	(0.024)
R-Squared	0.308	0.365	0.300
Panel B			
Feedback	0.046	0.007	0.054
	(0.064)	(0.152)	(0.075)
Standard	-0.324	-0.443**	-0.308
	(0.202)	(0.203)	(0.208)
Relative	-0.207	-0.359	-0.189
	(0.227)	(0.232)	(0.232)
Std. x FB	-0.015	0.093	-0.035
	(0.072)	(0.180)	(0.083)
Rel. x FB	-0.027	0.122	-0.052
	(0.073)	(0.220)	(0.083)
N	5159	722	4437
R-Squared	0.317	0.376	0.310
Mean of Dep. Var.	-0.186	-0.070	-0.205

Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects and the baseline value of the outcome variable. Additional controls include zone fixed effects, age, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A10: Perceptions of school performance by feedback status

	Sample: Grade 5-7	
	Self-evaluated performance	
	(1)	(2)
Mid-term score	0.015***	0.015***
	(0.002)	(0.002)
Baseline score	0.005**	0.005*
	(0.003)	(0.003)
Feedback		-0.049
		(0.140)
Mid × Feedback		0.001
		(0.003)
Base × Feedback		0.000
		(0.003)
N	4597	4597
R-Squared	0.066	0.066
Mean of Dep. Var.	3.255	3.255

Notes: Robust standard errors in parentheses. Standard errors clustered at the school-grade level. All specifications include grade fixed effects, zone fixed effects, baseline value of dependent variables, 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 A11: Expectations and Feedback

	Sample: Grade 5-7		
	Expectation		
	All	Top 15%	Bot 85%
	(1)	(2)	(3)
Feedback	-0.016	-0.030	-0.014
	(0.015)	(0.034)	(0.019)
Mid-Base	0.002***	-0.001	0.002**
	(0.001)	(0.003)	(0.001)
Feedback * (Mid-Base)	-0.000	-0.000	-0.000
	(0.001)	(0.002)	(0.001)
Baseline Score	0.009***	0.003	0.006**
	(0.002)	(0.004)	(0.003)
N	3792	795	2997
R-Squared	0.015	0.007	0.004
Mean of Dep. Var.	0.669	0.772	0.641

Notes: Robust standard errors in parentheses. Standard errors clustered at the school-grade level. All specifications include grade fixed effects, zone fixed effects, baseline value of dependent variables, and demographic controls such as age, race, household size, and a household asset index. Mid-Base is a difference of percentile ranks between the midterm and baseline exam. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table A12: Feedback effect: Feedback intensity

	Sample: Grade 5-7		
	Final exam		
	All	Top 15%	Bot 85%
	(1)	(2)	(3)
Feedback	0.029	0.034	0.021
	(0.025)	(0.074)	(0.026)
Mid-Base	0.023***	0.016**	0.020***
	(0.003)	(0.007)	(0.003)
Feedback * (Mid-Base)	-0.001	0.002	-0.002
	(0.001)	(0.003)	(0.001)
N	4688	1057	3631
R-Squared	0.439	0.270	0.303
Mean of Dep. Var.	-0.146	0.846	-0.435

Notes: Robust standard errors in parentheses. Standard errors clustered at the school-grade level. All specifications include grade fixed effects, zone fixed effects, baseline value of dependent variables, and demographic controls such as age, race, household size, and a household asset index. Mid-Base is a difference of percentile ranks between midterm and baseline exam. * 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: Measures of Self-esteem, Grit, etc

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.

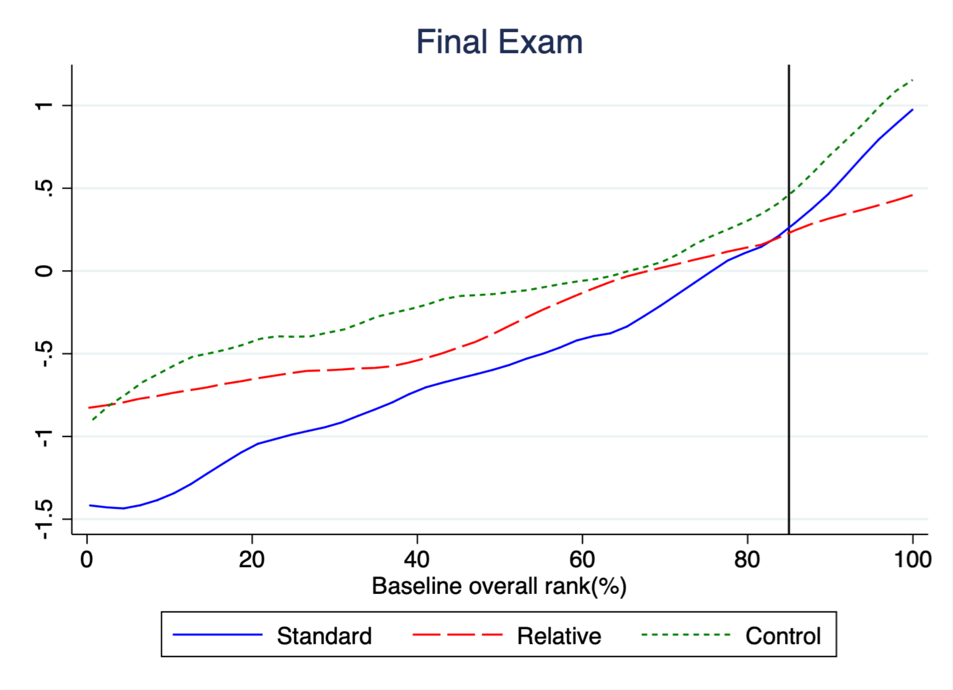
	Self-Esteem	Strongly disagree	Disagree	Agree	Strongly agree
701.	On the whole, I am satisfied with myself	1	2	3	4
702.	At times I think I am no good at all	1	2	3	4
703.	I feel that I have a number of good qualities	1	2	3	4
704.	I am able to do things as well as most other people.	1	2	3	4
705.	I feel 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

	Grit	Not like me at all	Not much like me	Some-what 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

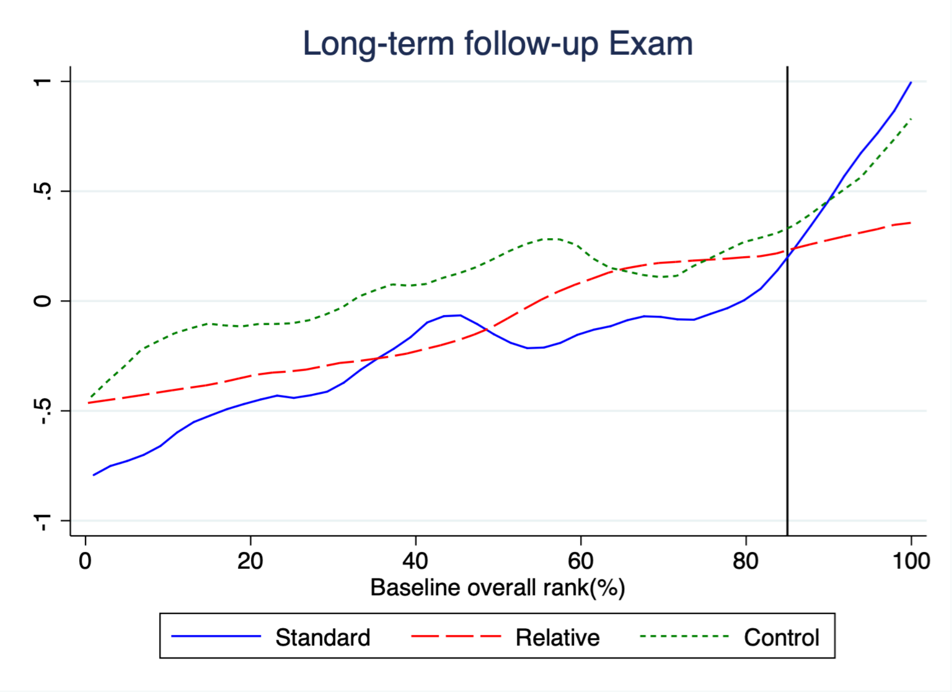
	Conscientiousness					
	I see Myself as Someone Who...	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
719.	Does a thorough job	1	2	3	4	5
720.	Can be somewhat careless.	1	2	3	4	5
721.	Is a reliable worker.	1	2	3	4	5
722.	Tends to be disorganized.	1	2	3	4	5
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: Exam scores at follow-up by Baseline Rank, Long-term Follow-up Sample

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

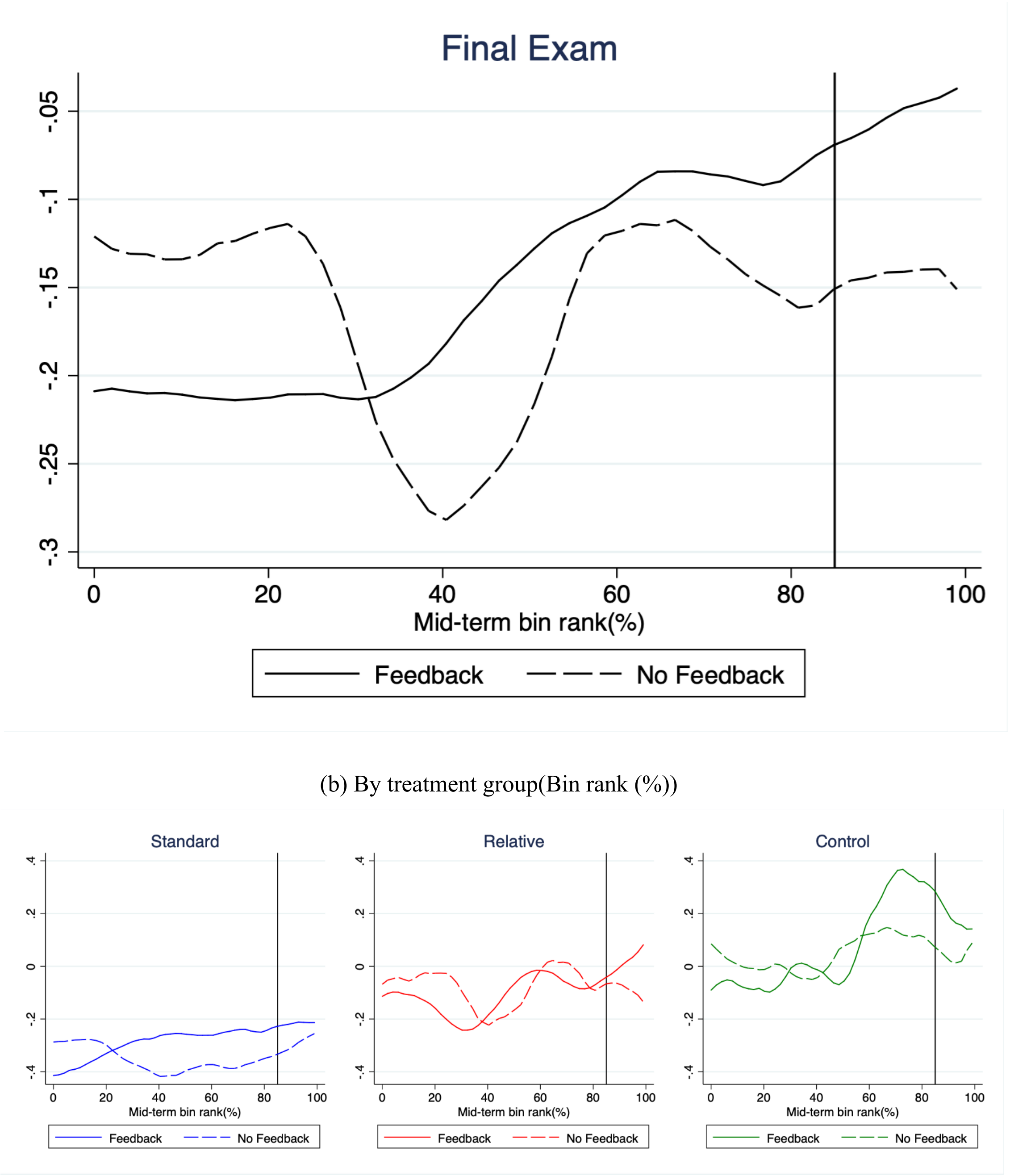


(b) 2nd follow-up exam, Long-term Follow-up Sample (Grade 5-6)



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

Figure A4: Feedback effect on follow-up exam score by mid-term bin rank(%)



(a) Whole sample

(b) By treatment group(Bin rank (%))

Note: This figure presents average final exam scores by mid-term bin rank(%). Panel A presents the results for all students, while Panel B presents the results by scholarship treatment status.