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

James Berry, Hyuncheol Bryant Kim, and Hyuk Harry Son*

February 2021

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 scholarships, field experiments

*Berry: University of Delaware, jimberry@udel.edu; Kim: Hong Kong University of Science and Technology and 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, and generous funding support from Viatron Technology. We thank Miguel Urquiola, Cristian Pop-Eleches, Jonah Rockoff, and seminar participants at the American Economic Association Annual Meeting, briq/IZA Workshop on Behavioral Economics of Education, International Food Policy Research Institute, Korean Economic Association-Asia Pacific Economic Association International Conference, National University of Singapore, New York University, Northeast Universities Development Consortium Conference, Syracuse University, Seoul Journal of Economics International Conference, and Yonsei University for 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., 2016; 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 dont 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 this paper, we study the impacts of two types of incentive programs on 5th to 8th graders in 31 primary schools in Malawi. The two incentive programs, presented as scholarship schemes, provided rewards of MWK 4500 (USD 9.70) if the corresponding test score goal was met.[3] The first, which we call the *Standard* scholarship scheme, provided a schol-

[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).

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

arship 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* 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."[4]

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 (the first follow-up). In addition, for students in 5th and 6th grade at baseline, we conducted a second follow-up survey and exam six months after the experiment was completed. The second follow-up survey and exam allow us to understand the impacts of and behavioral responses to the incentive for students after the incentives disappeared.

Our main finding is that the *Standard* scholarship scheme reduced final exam scores by 0.27 standard deviations across the full sample, with the largest negative impacts on students with low 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, with small and negative

[4]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).

point estimates. Although our study lacks power to detect statistically significant differences between the impacts of the *Standard* and *Relative* scholarships on average, point estimates suggest that in the *Relative* group performed better than those in the *Standard* group, especially among the bottom performers at baseline. 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. In addition, using an additional round of data collection, we show short-term negative impacts of the *Standard* scholarship were diminished in the next term, after the incentive had been removed.

Taken together, these results suggest that tournament incentives may de-motivate students—particularly low-performing students—by reminding them of their place in the performance distribution and signalling that high performance is valuable. This is related to research on stereotype threat, in which revealing one's social identity can lead individuals to confirm to negative steroetypes. For example, Hoff and Pandey (2006) 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 two primary dimensions. First, it contributes to the growing literature on incentives to learn 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).[5] 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 was structured similarly, although it applied to both boys and girls. A key difference is that in our setting, students are aware of their initial test score

[5]Within the developed-country literature, two studies are of particular note. Leuven, Oosterbeek, and Klaauw (2010) 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. In a second study of university students in Spain, Campos-Mercade and Wengström (2020) provide incentives for passing a GPA threshold and find effects only for students whose prior GPA was just below the threshold.

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 incentives vary across studies, many study a single incentive scheme. A smaller but growing literature evaluates the structure of incentives 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.

Second, 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) find that incentives for attendance among primary students in India decreased intrinsic motivation.

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 and discusses the results. We conclude in Section 5.

2 Context, Programs, and Study Design

2.1 Primary education in Malawi

Similar to other countries in Sub-Saharan Africa, the government of Malawi abolished primary school fees in the early 1990s, leading to near-universal enrollment in grades 1 to 8. However, like many countries in the developing world, learning outcomes among Malawian primary students are low. Even within 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.[6]

The academic calendar, starting in September, consists of three terms. At the end of each term, students in primary school take exams in six subjects: Chichewa (the vernacular language), English, mathematics, primary science, social studies, and art and life skills. Students typically must pay a fee 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 term 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.[7] The scholarship programs were conducted in grades 5 to 8 in 31 public primary schools in the sub-district. 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.

[6]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.

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

2.2.1 Study design

The project chronology is summarized in Figure 1. The baseline survey and baseline exams were implemented during the first term of the 2014-2015 academic year (December 2014 to January 2015).[8] The final exam and surveys were conducted at the end of third term of the 2014-2015 academic year, in June 2015. Lastly, for students initially in the 5th and 6th grades, we collected exam scores in March of 2016, nine months after the scholarship programs ended.[9]

Table 1 displays the sample composition in each treatment category. In February 2015, we stratified the 118 school-grades by grade and randomly assigned school-grades into three groups: the *Standard* scholarship, the *Relative* scholarship, or the control group.[10] The results of the scholarship randomization were announced in the middle of the second term. 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 students 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 students overall rank was provided.

The first follow-up survey and final exams took place at the end of the third term (June 2015).[11] The final exam determined eligibility for the scholarships. Awards were distributed

[8]Baseline exams were conducted twice, at the end of the first term (December 2014) and the beginning of the second term (January 2015). 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. If the student took both tests, we use the average score. Otherwise, we use the score of the test the student took.

[9]After the March 2016 exam, we conducted a second-year trial in which we randomly assigned students to the *Relative* scholarship or to a tutoring program. Both years' evaluations are described on the projects Social Science Registry website, https://www.socialscienceregistry.org/trials/1119. The results of the second-year trial are in progress.

[10]Several schools did not have upper grades, resulting in 118 grades between 5 and 8 in our 31 study schools.

[11]As we describe in the next section, we used the PSLCE exam for eighth graders, which took place in

in an area-wide awards ceremony that took place after the experiment was completed (October 2015). Finally, the second follow-up exams and surveys for 5th and 6th graders at baseline were administered nine months after the experiment was completed (March 2016).

2.2.2 Interventions

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. Under the *Relative* scholarship scheme, students were grouped into bins of 100 students by sub-district level baseline test score, and the top 15 percent of each bin in the final exam were eligible to receive the award.[12]

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.[13,14] This represents a significant amount considering that Malawi GDP per capita was only around USD 362.7 in 2014 (Bank, 2015).

To ensure that students fully understood the scholarship programs (particularly the *Relative* scholarship scheme) 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

May 2015.

[12]Specifically, the 7386 students in our study sample were grouped into 74 bins by baseline test score. Seventy-three bins contained 100 students, and the last (bottom) bin contained 86 students.

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

[14]The value of the award is comparable to that of Kremer, Miguel, and Thornton (2009) and Blimpo (2014), whose awards were valued at USD 6.4 and 10, respectively.

individualized announcements.

For fifth, sixth, and seventh graders, exams used in this study were developed by a sub-district-level exam committee to ensure uniformity across schools.[15] 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. For eighth graders, the study utilized the PSLCE national exam instead of the sub-district-level final exam.

In addition to the scholarship programs, the study design included a feedback intervention which provided rank information on a midterm exam, administered at the end of the second term (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.

Unfortunately, there were issues with the calculation of the midterm ranks that resulted in students receiving incorrect or overstated information on their midterm performance. We discuss these issues and analyze the impacts of the feedback interventions, as implemented, in Appendix B. Table B5 in Appendix B also presents our main estimates of the impacts of the scholarship programs on only the students who were randomly assigned to the no-feedback group. As we show, our conclusions are unchanged if we restrict analysis to these students.

2.3 Data

We use several sources of data: standardized test score data (the baseline, final exam, and longer-term follow-up exams), school attendance checks, 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' ranks in these tests.[16] In addition to

[15]Prior to this study, each school created its own end-of-term 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. The exam committee consisted of eight teachers, one vice-principal, and one principal (head teacher) of the schools within the sub-district.

[16]For 8th graders who took the PSLCE instead of the regular final exam, we were able to obtain letter

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 – 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).[17] 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.[18] 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).[19]

In addition, the surveys collected students' reports on their own effort, as well as that of teachers and parents. Student effort was measured through self reports of weekly study hours and monthly unannounced checks of 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

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.

[17]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.

[18]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.

[19]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.

to 8 in participating schools. Among these 9,581 students, 7,637 (79.7 percent) completed the baseline survey and 8,597 (89.7 percent) participated in the baseline exam. The final study sample consists of 7,385 students (77.1 percent) who participated in both the baseline survey and baseline exam.

Table 2 presents baseline characteristics and balance checks for the scholarship randomization. Column 1 displays summary statistics of key variables for the control group. The average age is 14.4, and 48.6 percent of the sample are males. At the time of the baseline survey, the school attendance rate of the students was 86 percent, and the average study hours per week was 16.8.

Columns (2) and (3) of Table 2 show tests of differences in means between the scholarship groups and the control group. 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.

Table A1 displays sample attrition across treatment groups. On average 83 and 90 percent of the study sample participated in the follow-up survey and final exam, respectively. For the longer-term follow-up survey and exam, 63 and 57 percent of baseline 5th and 6th graders 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 2.9 percentage points more likely to take the final exam (significant at the 5 percent level). In Appendix C, we present additional analysis of attrition by scholarship treatment, including bounds on our main treatment effects following Lee (2009). The analysis shows that scholarship treatment effects, as well as interactions between treatment and baseline test score, are unlikely to be substantively affected by differential attrition.

3 Estimating Equations

To estimate the average impacts of the *Standard* and *Relative* scholarship programs, we use the following equation:

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

where Y_{igsz1} is the outcome of interest for student i of grade g in school s at school zone z. *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_z is a fixed effect for zone. In some specifications, we include X_{igsz}, a set of student-level controls, including age, gender, 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 overall baseline rank was in the top 15 percent. We select the top 15 percent because students responses to the scholarships might differ based on whether they are above or below the cutoff for scholarship eligibility at baseline. This implies the following regression:

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

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 bins used in the *Relative* scholarship scheme.

To examine the heterogeneous impacts by students' initial rank in more detail, we interact the treatment group dummies with a series of indicators for whether the student's overall baseline rank was in each quintile. To implement this, we estimate following regression:

$$Y_{igsz1} = \sum_{k=1,2,3,4,5} \theta_k \cdot k^{th}\_Quintile_{igsz0} + \sum_{k=1,2,3,4,5} \theta_k^s \cdot Standard_{gsz} \cdot k^{th}\_Quintile_{igsz0} \quad (3)$$
$$+ \sum_{k=1,2,3,4,5} \theta_k^r \cdot Relative_{gsz} \cdot k^{th}\_Quintile_{igsz0} + Y_{igsz0} + \eta_g + \gamma_z + X_{igsz} + \epsilon_{igsz}$$

where $k^{th}\_Quintile_{igsz0}$ for $k = \{1, 2, 3, 4, 5\}$ are binary variables equal to one if a student is ranked in each quintile at baseline. In this specification, we omit indicators for *Standard* and *Relative* scholarship treatment, so that the coefficients θ_k^s and θ_k^r capture the impacts of the *Standard* and *Relative* scholarships on the students whose ranks were in quintile k at baseline.

4 Results

4.1 Understanding and Expectation

Before turning to the main impact results, we first discuss students' understanding of the program and expectations that they would receive a scholarship. As described in Section 2.2, AFF provided one-hour introduction sessions to all students to ensure students fully understood the scholarship schemes. We measured students understanding and expectations 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 schemes

and had expectations consistent with their assigned groups.

Figure 3 presents graphs of the percent of questions answered correctly on the test for understanding of the scholarship schemes (y-axis) by overall baseline 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.

Panel A of Figure 4 displays students expectations of winning the scholarship by overall baseline rank.[20] For students in the *Standard* scholarship group, expectations of receiving the scholarship should increase with overall baseline 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 4 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 30-44 percentage points more likely to expect the scholarship. Examining differences across overall baseline rank, those in the top 15 percent in the *Standard* scholarship group were significantly more likely to expect the scholarship, 49 and 21 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 4 shows students' expectations of winning the scholarship by baseline bin rank. Columns (3) and (4) of Panel C in Table 3 present corresponding regression results. Immediately after the announcement, expectations increase with baseline bin rank only for the *Relative* scholarship group, as expected. However, by the first follow-up, expectations in

[20]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?"

both scholarship groups are relatively flat across baseline bin rank.

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)).[21] The *Standard* scholarship had substantial negative impacts on student performance: students performed 0.27 standard deviations worse than those in the control group (significant at the 10 percent level). The effects of the *Relative* scholarship were not statistically significant, with negative point estimates ranging from -0.05 to -0.13 standard deviations. Although the point estimates suggest a substantially larger negative reaction to the *standard* scholarship compared to the *Relative* scholarship, we cannot reject that the impacts are equal, with p-values of the test for equality of 0.21 and 0.34 for the specifications excluding and including controls, respectively.

Panel A of Figure 5 presents nonparametric plots of final exam scores in each treatment group by overall baseline rank. The figure shows that 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.[22]

Panel B of Table 4 presents an additional analysis of heterogeneity by overall baseline rank by interacting the treatment with an indicator for being in the top 15 percent of baseline test scores, as per Equation (2). These results confirm that the decrease in academic achievement in the *Standard* treatment is driven by students with initial test scores in the bottom 85 percent: the coefficient on *Standard* scholarship is negative and significant, and that on the interaction between *Standard* scholarship and being in the top 15 percent at

[21] 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).

[22] The negative impacts of the *Standard* scholarship were significantly larger among girls than boys (See Appendix Table A3).

baseline is of opposite sign and larger than the coefficient on the *Standard* scholarship, 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 cannot reject that the impacts of the *Standard* and *Relative* scholarships are equal in each initial performance level, with p-values of the test for equality of 0.13 and 0.12 for the bottom 85 percent and the top 15 percent, respectively.

Table 5 presents analysis of heterogeneity in by students' initial ranks in more detail, using the a series of indicators for being in each quintile as of the baseline test instead of in the top 15%, following Equation (3). The results confirm that the negative impacts of the *Standard* scholarship program are concentrated in the lower quintiles: coefficients on the interaction of the *Standard* treatment (θ_k^s) and each quintile are larger in magnitude in the lower quintiles, although some of these coefficients are not statistically significant (Column 1). On the other hand, as shown in Column (2), the *Relative* scholarship program had positive impacts on the lowest-performing students and negative impacts on the highest-performing students, although none of the estimates is statistically significant. Lastly, Column (3) provides an estimate and standard error of the difference between the two impacts ($\theta_k^s - \theta_k^r$), which is the largest in the lowest quintile (0.51 standard deviations, significant at the 10 percent level).

Finally, we examine whether the impacts vary by bin rank – that is, the ranking within the 100-student subgroups used to award the *Relative* scholarship. In Panel B of of Figure 5, 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.[23]

[23]Table A2 presents the analysis of Table 4 by subject. Results are largely similar across the subjects.

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 survey responses of students, including school attendance, time spent studying, motivation to study, self-esteem, and conscientiousness. These results are presented in Columns (1) to (7) of Table 6, with average impacts in Panel A and heterogeneity by overall baseline rank in Panel B.

We find few impacts on observed and self-reported student effort. As shown in Column (1) of Table 6, there is a small marginally significant 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 first follow-up survey (Column (2)), but point estimates suggest slightly less study effort in both scholarship treatment groups on average (Panel A), and we do not find meaningful heterogeneity by baseline score (Panel B).

Turning to impacts on non-cognitive measures, we find impacts that generally correspond to the overall test score results presented in the previous section (Columns (3) to (7) of Table 6). 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 standardized 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.

In terms of heterogeneity by baseline score, Panel B of Table 6 shows that the negative impacts of the *Standard* scholarship on non-cognitive skills were concentrated among the bottom 85 percent of students: as shown in Column 7, the impact on the non-cognitive skill index among this group is -0.18 standard deviations and is significant at the 1 percent level. The impact on the top 15 percent is 0.23 standard deviations higher than the bottom 85

percent (significant at the 10 percent level). By contrast, we do not find similar evidence of heterogeneity for the *Relative* scholarship group. These findings suggest that, by signalling that high performance was valuable, the scholarships may have de-motivated students, particularly those in the *Standard* scholarship program that were the least likely to receive the scholarship.

Columns (8) to (10) of Table 6 present impacts on students' perceptions of teacher and parental effort. We do not find evidence for changes in teacher effort as a result of either scholarship program. We do find that parents mentioned the scholarship program more often in the standard scholarship group, with effects concentrated among children with the highest baseline test scores. However, even though parents of the *Standard* scholarship group mentioned the opportunity more, it did not appear to translate into actual parental effort.

It is worth noting that a large portion of parents in our sample had little or no education and therefore may not have had the skills to effectively help their children at home.[24] 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 small attendance impacts of the *Standard* scholarship may therefore have been partially a result of parental encouragement to attend school.

4.4 Longer-term Impacts

As discussed previously, the *Standard* scholarship program resulted in large negative impacts on non-cognitive skills as well as the score on the final exam, an incentivized test. In this section, we analyze impacts on the scores of the test administered in the following term, 9 months after the incentivized final exam.

As described in Section 2.2, longer-term follow-up tests were conducted in the school year after the scholarship programs took place. The participants for these longer-term follow-up exams were the students who were 5th and 6th graders at the baseline. When presenting our longer-term follow-up results, we also display final exam results of the sub-sample of 5th

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

and 6th graders to confirm that the results presented in the previous subsections hold for the sample that was followed into the next school year.

Table 7 presents the longer-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 longer-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. We note, however, that these estimates are imprecise, with confidence intervals admitting fairly large negative impacts. We also find smaller – and still statistically insignificant – negative effects of the *Relative* scholarship program in the longer-term, although again the estimates lack precision to draw stronger conclusions.

Table A4 presents corresponding short- and longer-term results on attendance, self-reported student effort, and non-cognitive skills for 5th and 6th graders at the baseline. Even though there were negative effects of the *Standard* scholarship on non-cognitive skills in the short-term, we do not find persistent changes in the longer-term, which corresponds to a reduced longer-term impact on test scores.

Although our estimates are imprecise, these results suggest that the negative short-term impacts of the *Standard* scholarship program diminished over time. This brings up the possibility that the short-term test score impacts could be due to test-taking effort on the final exams rather than learning over the course of the term. While we cannot rule out the possibility of test-day effort on the final exams, the impacts on survey-based measures of non-cognitive outcomes are consistent with the short-term test-score impacts, suggesting that the incentives had broader effects on students during the term.

4.5 Discussion

This section provides additional discussion of our results. In Section 4.2 we showed that the *Standard* scholarship resulted in a significant decrease in test scores, especially for those with lower scores at baseline. We also find negative impacts of the *Standard* scholarship on motivation to study and other non-cognitive skills, again with larger effects on those who

are unlikely to win the reward. These findings suggest that the *Standard* scholarship may have decreased non-cognitive skills, and subsequently exam performance, by highlighting a goal that was difficult to achieve, particularly for the lowest-performing students.

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, because these non-cognitive measures were taken as of the follow-up survey and are therefore endogenous, this analysis should be treated as speculative. As shown in Table A5, we find that test scores are explained at least partially by these control variables (about 11% (7.368-7.010)/7.368)). The original negative impacts is 7.368 (Column(4) of Table 4) and it reduces to 7.010 after controlling for non-cognitive skills. However, much of the impacts remain even after controlling for these variables. This could imply that our non-cognitive measures are not comprehensive enough; for example, the test score impacts could have been driven by specific types of motivation that our somewhat coarse measure does not capture.

In contrast with the *Standard* scholarship, we did not find negative impacts in the *Relative* scholarship group. This suggests that by providing a greater chance to achieve the incentive, the negative effects of the *Standard* scholarship were mitigated. However, the comparison of the two treatment groups is limited by power: we cannot reject the hypothesis of equal impacts of the two scholarship schemes, despite differences of about 0.15 to 0.2 standard deviations. This suggests that some caution is warranted in interpreting the results of the *Relative* scholarship program and in comparing the two schemes.

In the remainder of this section we consider several explanations for the effects (and lack of effects) we observe. First, students may not have fully understood their scholarship scheme. However, as we showed in Section 4.1, students did appear to understand and had expectations in line with their assigned groups. While understanding was not perfect, the amount of misunderstanding was unlikely to have negated positive effects, and particularly would not have resulted in negative impacts of the *Standard* scholarship.

The second possibility is that the power of incentives was not great enough to induce

effort and may have been muted by the other stakes within the exams. The cash incentive was USD 9.70, which is substantial relative to Malawi's annual GDP per capita of USD 380. Nonetheless, the end-of-year exams do nominally determine progression to the next grade, and therefore they carry their own incentives. One way to check this is to compare the results of 5th to 7th grades with those of 8th grade. While the exams at all levels are used for grade progression, the 8th grade exam additionally conveys the primary school leaving certificate credential. As shown in Panel A Table A6, we actually find that the smallest (i.e., most negative) effects were for grades 5 to 7, where the outside incentives on the exam were lowest. This suggests that incentives outside of the experiment did not dampen student effort and drive down our estimated impacts.

Instead, a key factor for negative impacts of the *Standard* scholarship and the lack of impacts of the *Relative* scholarship may have been the context in which the incentives were provided. These contextual differences may explain the contrasting results in Kremer, Miguel, and Thornton (2009), who also worked in rural schools in sub-Saharan Africa and whose design the *Standard* scholarship was based upon. First, because students knew their ranking at baseline, the difficulty in in achieving the *Standard* scholarship may have been particularly salient, especially for students with the lowest baseline scores in our setting. This contrasts with Kremer, Miguel, and Thornton (2009), where no such information was provided. Second, as noted in Section 2.1, there were approximately 85 students for each teacher in these schools. Although Kremer, Miguel, and Thornton (2009) operated in a similarly under-resourced environment, the intervention increased teacher effort, which they note may have contributed substantially to their impacts. As shown in Table 6, there do not appear to have been an increase in teacher effort in our study.[25] Within this environment, the scholarships may have been particularly de-motivating for students who have little chance of reaching the goal. This could explain why we see the most negative impacts for the initially lowest performing students in the *Standard* scholarship group, but not in the *Relative* scholarship group.

[25]The incentives could also have affected the classroom environment by making students in the scholarship classrooms more competitive and less likely to help each other study. However, using students in our our longer-term follow-up survey, we do not find evidence that either scholarship group changed the classroom environment (see Table A7).

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.

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 not statistically significant, although they are still negative.

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. 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, 2006). In our context, the negative effects appear largely isolated to the incentivized test and dissipate in the longer term.

The negative distributional effects of tournament incentives may be especially pronounced in environments such as ours, in which students have relatively few education inputs in schools or at home. The information provided on student ranking may also have made the difficulty in achieving the incentives more salient. We speculate that this may explain

the differences between our results and those of Kremer, Miguel, and Thornton (2009), but future work is needed to more rigorously estimate the factors that contribute to the success (or failure) of such schemes.

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.

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

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

Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin (2015). "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." *Journal of Political Economy* 123(2), 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.

Campos-Mercade, Pol and Erik Wengström (2020). "Threshold Incentives and Academic Performance."

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

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel (2011). "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25(4), 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 (2006). "Discrimination, Social Identity, and Durable Inequalities." *American Economic Review* 96(2), 206–211.

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., John A. List, Susanne Neckermann, and Sally Sadoff (2016). "The behavioralist goes to school: Leveraging behavioral economics to improve educational performance." *American Economic Journal: Economic Policy* 8(4), 183–219.

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

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

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

Rosenberg, Morris (1965). "Society and the Adolescent Self-Image." *Science* 148(3671), 804–804.

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

Visaria, Sujata, Rajeev Dehejia, Melody M. Chao, and Anirban Mukhopadhyay (2016). "Unintended consequences of rewards for student attendance: Results from a field experiment in Indian classrooms." *Economics of Education Review* 54, 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 longer-term follow-up)

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

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

Table 2: Balance of Baseline Variables Across Treatment Groups

	Control Mean	*Standard* vs. Control	*Relative* vs. Control	N
	(1)	(2)	(3)	(4)
Age	14.4	0.052	0.159	7385
	[3.60]	(0.178)	(0.187)	
Male	0.486	0.005	-0.017	7385
	[0.500]	(0.020)	(0.018)	
Ethnic group: Chewa	0.914	-0.038	-0.041	7358
	[0.280]	(0.034)	(0.033)	
Household size	7.81	0.305	0.263	7385
	[1.66]	(0.354)	(0.344)	
Asset index	-0.009	0.011	0.037	7102
	[1.88]	(0.175)	(0.176)	
Baseline rank(%)	51.5	-0.625	1.47	7342
	[27.3]	(3.28)	(3.99)	
Baseline Score	0.00000	-0.021	0.066	7342
	[0.999]	(0.127)	(0.159)	
Attendance	0.863	0.003	-0.004	7385
	[0.196]	(0.016)	(0.016)	
Study hours per week	16.8	-0.507	-0.210	7308
	[16.4]	(0.792)	(0.796)	
Motivation to study	4.53	-0.014	0.053	7374
	[0.789]	(0.058)	(0.052)	
Self-esteem	2.67	-0.013	-0.006	7368
	[0.338]	(0.022)	(0.022)	
Conscientiousness	3.58	0.028	0.094	7370
	[0.600]	(0.060)	(0.060)	
Grit	3.21	-0.026	-0.011	7368
	[0.450]	(0.020)	(0.024)	
Teacher Index	-0.003	0.151	0.250*	7364
	[1.000]	(0.140)	(0.133)	
Parental Effort Index	0.001	-0.024	-0.010	7281
	[1.00]	(0.071)	(0.064)	

Notes: Column 1 reports means for subjects assigned to the control group. Columns 2 and 3 report mean differences between the scholarship treatment groups and the control group.Standard deviations are in brackets, and standard errors, clustered at the school-grade level, are in parentheses. Test scores are normalized using the control group mean and standard deviation. The asset index is constructed as the 1st principal component of variables indicating the ownership of 26 assets. Teacher and parental effort indices are aggregates of the seven and four measures, respectively. Indices are generated 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 then standardized relative to the control group. Self-esteem, grit, and conscientiousness measures are simple averages of questions measured on a four-point scale (self-esteem) or five-point scale (grit and conscientiousness). * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 3: Understanding and Expectation

	Sample: Grade 5-8			
	Understanding		Expectation	
	After Announcement	1st Follow-up	After Announcement	1st Follow-up
	(1)	(2)	(3)	(4)
Panel A: Average treatment 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: In Columns 1 and 2, the dependent variable reflects the percent of questions answered correctly on the test of understanding of the scholarship schemes. In Columns 3 and 4, the dependent variable is an indicator that equals 1 if the student answered [very likely] or [likely] to the question: [Based on your current position, how much do you think you have a chance of receiving the gift?] Standard errors, clustered at the classroom level, are in parentheses. All specifications include grade fixed effects, district 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 4: Test Score Impacts

	Sample: Grade 5-8			
	1st Follow-up			
	Exam Rank		Exam score	
	(1)	(2)	(3)	(4)
Panel A: Average treatment effects				
Standard	-7.402**	-7.368*	-0.265*	-0.266*
	(3.620)	(3.868)	(0.135)	(0.146)
Relative	-2.516	-4.730	-0.045	-0.126
	(4.668)	(4.404)	(0.186)	(0.174)
R-Squared	0.234	0.305	0.252	0.324
P-value: *Std = Rel*	0.250	0.447	0.207	0.337
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-8.961**	-8.682**	-0.313**	-0.305**
	(3.833)	(4.138)	(0.139)	(0.153)
Relative	-1.543	-4.016	0.018	-0.073
	(4.987)	(4.769)	(0.193)	(0.184)
Std. x Top 15%	9.697*	7.507	0.301	0.224
	(5.540)	(5.316)	(0.241)	(0.230)
Rel. x Top 15%	-5.696	-4.348	-0.359	-0.299
	(7.370)	(6.057)	(0.294)	(0.253)
Top 15%	2.777	3.847	0.081	0.118
	(5.111)	(4.730)	(0.223)	(0.209)
R-Squared	0.244	0.312	0.262	0.330
P-value: *Std = Rel* at Bot. 85%	0.095	0.211	0.063	0.124
P-value: *Std = Rel* at Top 15%	0.169	0.086	0.174	0.125
Panel C: Hegerogeneous treatment effects by bin rank				
Standard	-7.404**	-7.360*	-0.267*	-0.266*
	(3.727)	(3.982)	(0.140)	(0.151)
Relative	-2.234	-4.423	-0.029	-0.109
	(4.761)	(4.527)	(0.190)	(0.180)
Std. x Subg. Top 15%	0.069	0.038	0.010	0.003
	(2.201)	(2.270)	(0.088)	(0.090)
Rel. x Subg. Top 15%	-1.731	-1.877	-0.100	-0.106
	(2.166)	(2.227)	(0.087)	(0.088)
Additional controls	No	Yes	No	Yes
N	6586	6323	6586	6323
R-Squared	0.234	0.305	0.252	0.324
Mean of Dep. Var.	51.346	51.489	-0.154	-0.146
P-value: *Std = Rel* at Bot. 85%	0.231	0.406	0.181	0.289
P-value: *Std = Rel* at Top 15%	0.419	0.770	0.448	0.737

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: Heterogeneity by Quintile of Baseline Test Score

	Sample Grade: 5-8		
	Exam Score		
	Standard-Control θ_k^s	*Relative*-Control θ_k^r	*Standard*-*Relative* $\theta_k^s - \theta_k^r$
	(1)	(2)	(3)
Fifth Quintile (Lowest)	-0.310	0.210	-0.521*
	(0.239)	(0.336)	(0.269)
Fourth Quintile	-0.334	-0.062	-0.272
	(0.212)	(0.239)	(0.173)
Third Quintile	-0.350**	-0.203	-0.147
	(0.144)	(0.159)	(0.130)
Second Quintile	-0.263**	-0.121	-0.142
	(0.119)	(0.157)	(0.153)
First Quintile (Highest)	-0.083	-0.314	0.232
	(0.195)	(0.215)	(0.176)

Notes: Standard errors, clustered at the school-grade level, are in parentheses. Controls include baseline test score, grade and 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 6: Intermediate Outcomes

	Sample: Grade 5-8									
	Student input		Non-cognitive skills					Teacher and parental response		
	Atten-dance	Study Hours	Motiva-tion to study hard	Self esteem	Grit	Conscien tiousness	Non-cognitive skill index	Teacher effort index	Parental effort	Parents mentioned scholar-ship
	(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***	-0.046	-0.037	0.126**
	(0.013)	(1.036)	(0.035)	(0.017)	(0.023)	(0.032)	(0.052)	(0.102)	(0.085)	(0.064)
Relative	0.009	-1.562	-0.036	-0.028	-0.027	-0.027	-0.101*	-0.060	0.022	0.087
	(0.015)	(1.158)	(0.039)	(0.017)	(0.023)	(0.034)	(0.055)	(0.088)	(0.083)	(0.071)
R-Squared	0.193	0.076	0.022	0.050	0.049	0.080	0.116	0.086	0.044	0.038
P-value: *Std = Rel*	0.253	0.523	0.239	0.911	0.724	0.529	0.492	0.877	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***	-0.049	-0.043	0.081
	(0.013)	(1.121)	(0.038)	(0.018)	(0.023)	(0.031)	(0.055)	(0.106)	(0.090)	(0.067)
Relative	0.010	-1.432	-0.049	-0.026	-0.011	-0.021	-0.089	-0.058	0.047	0.107
	(0.016)	(1.237)	(0.042)	(0.018)	(0.024)	(0.031)	(0.057)	(0.093)	(0.087)	(0.069)
Std. x Top 15%	-0.008	0.093	0.116*	0.032	0.029	0.083	0.227*	0.019	0.030	0.278**
	(0.023)	(1.721)	(0.063)	(0.039)	(0.051)	(0.094)	(0.135)	(0.129)	(0.111)	(0.108)
Rel. x Top 15%	-0.021	-0.977	0.067	-0.016	-0.098**	-0.034	-0.082	-0.023	-0.157	-0.054
	(0.027)	(2.049)	(0.066)	(0.034)	(0.040)	(0.092)	(0.122)	(0.125)	(0.111)	(0.120)
Top 15%	0.043***	1.511	-0.004	0.024	0.090***	0.026	0.087	0.065	0.131	-0.230***
	(0.016)	(1.526)	(0.050)	(0.030)	(0.029)	(0.083)	(0.102)	(0.098)	(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	0.086	0.046	0.042
Mean of Dep. Var.	0.756	14.526	4.298	2.719	3.259	3.674	-0.131	-0.013	-0.026	3.409

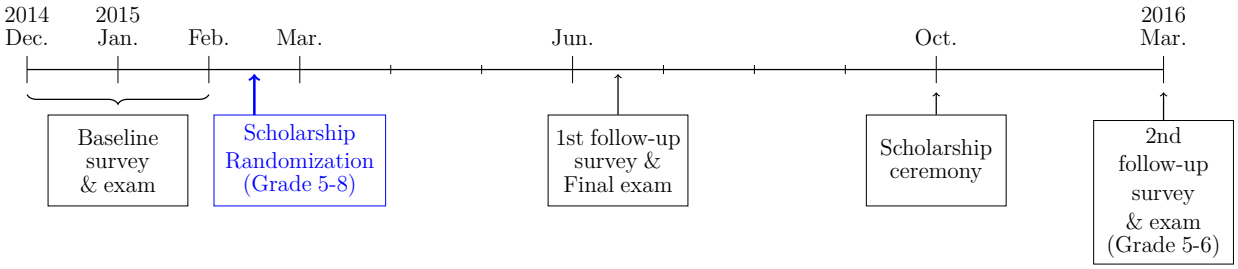
Notes: Standard errors, clustered at the school-grade level, are in parentheses. Controls include baseline test score, grade and zone fixed effects, age, ethnic group, household size, and a household asset index. Teacher and parental effort indices are aggregates of the seven and four measures, respectively. Indices are generated 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 then standardized relative to the control group. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 7: Longer Term Test Score Impacts

	Sample: Grade 5-6			
	1st Follow-up		2nd Follow-up	
	(1)	(2)	(3)	(4)
Panel A: Average treatment effects				
Standard	-0.463**	-0.518**	-0.242	-0.124
	(0.193)	(0.248)	(0.161)	(0.201)
Relative	-0.191	-0.374	-0.190	-0.104
	(0.245)	(0.277)	(0.128)	(0.169)
R-Squared	0.038	0.318	0.007	0.211
P-value: *Std = Rel*	0.243	0.465	0.763	0.907
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-0.474**	-0.547**	-0.305*	-0.186
	(0.226)	(0.267)	(0.158)	(0.201)
Relative	-0.131	-0.322	-0.161	-0.058
	(0.293)	(0.298)	(0.121)	(0.165)
Std. x Top 15%	0.212	0.185	0.355	0.343
	(0.275)	(0.296)	(0.245)	(0.221)
Rel. x Top 15%	-0.442	-0.277	-0.163	-0.221
	(0.353)	(0.326)	(0.278)	(0.237)
Top 15%	0.118	0.123	-0.007	0.025
	(0.249)	(0.263)	(0.173)	(0.169)
Controls	No	Yes	No	Yes
N	4040	3860	2476	2371
R-Squared	0.241	0.323	0.112	0.222
Mean of Dep. Var.	-0.272	-0.264	-0.018	-0.015

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.

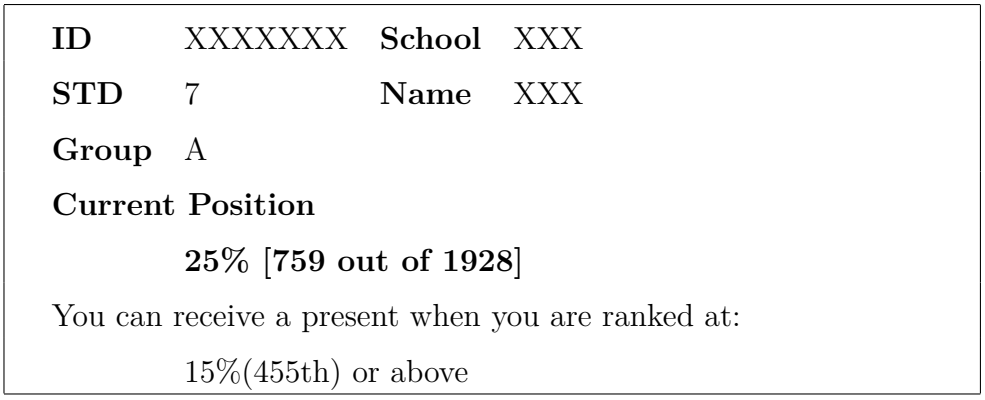
Figure 1: Experimental Timeline



Notes: Eighth graders took the PSLCE, a national-level exam to obtain secondary school admission, instead of the final exam, in May 2015.

Figure 2: Scholarship Randomization Result Announcement Note

(a) *Standard* scholarship group



ID XXXXXXX **School** XXX

STD 7 **Name** XXX

Group A

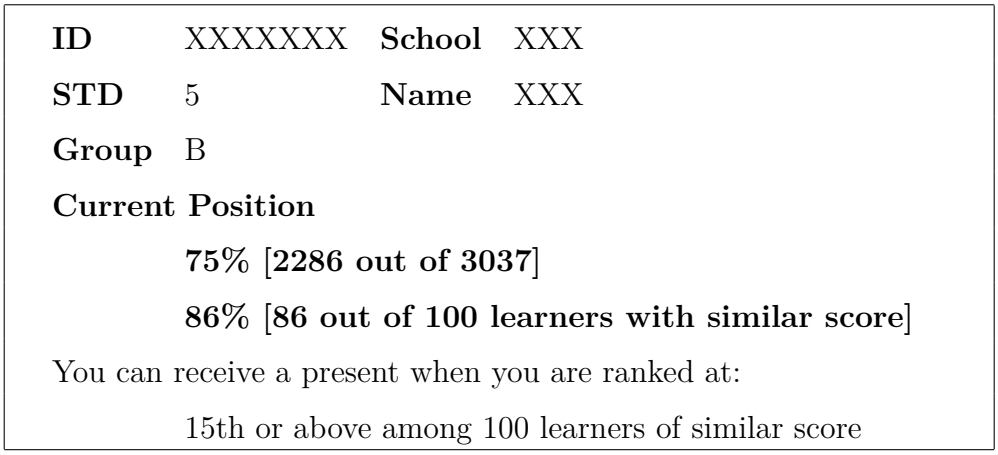
Current Position

25% [759 out of 1928]

You can receive a present when you are ranked at:

15%(455th) or above

(b) *Relative* scholarship group



ID XXXXXXX **School** XXX

STD 5 **Name** XXX

Group B

Current Position

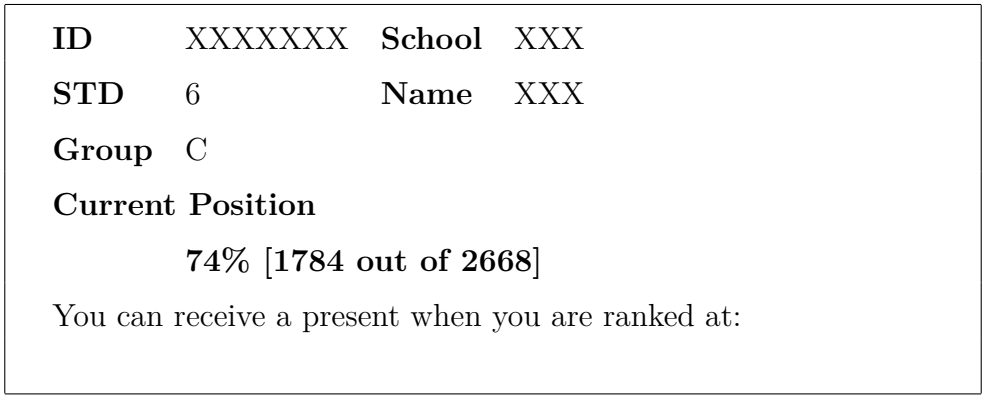
75% [2286 out of 3037]

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

You can receive a present when you are ranked at:

15th or above among 100 learners of similar score

(c) Control group



ID XXXXXXX **School** XXX

STD 6 **Name** XXX

Group C

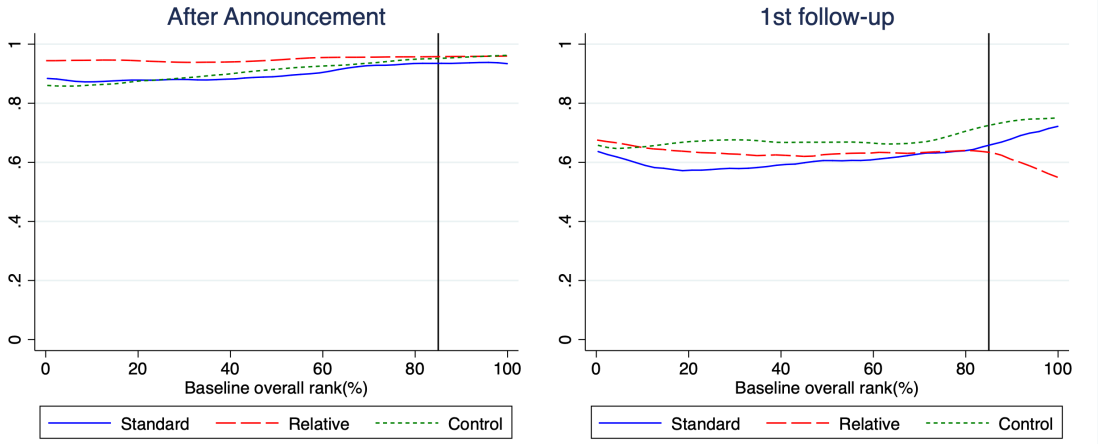
Current Position

74% [1784 out of 2668]

You can receive a present when you are ranked 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.

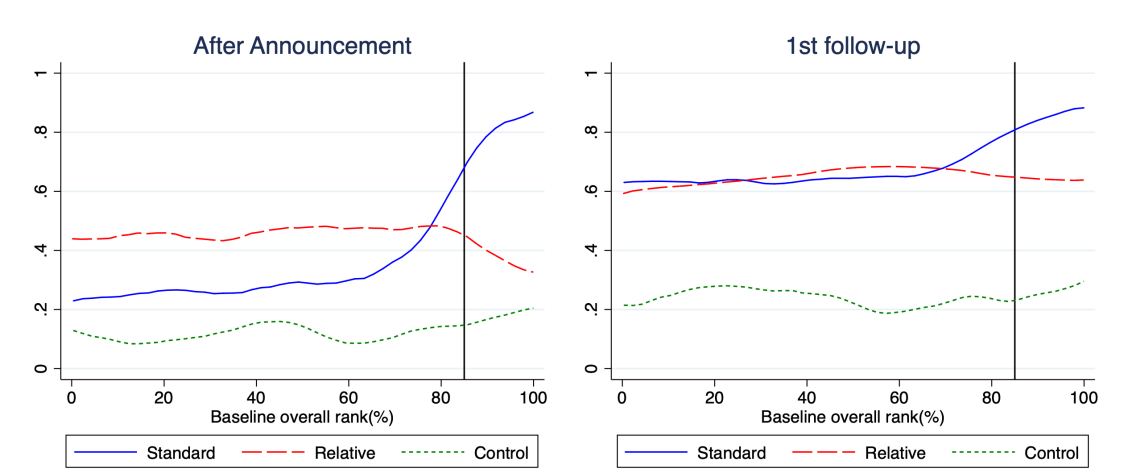
Figure 3: Understanding of the Program



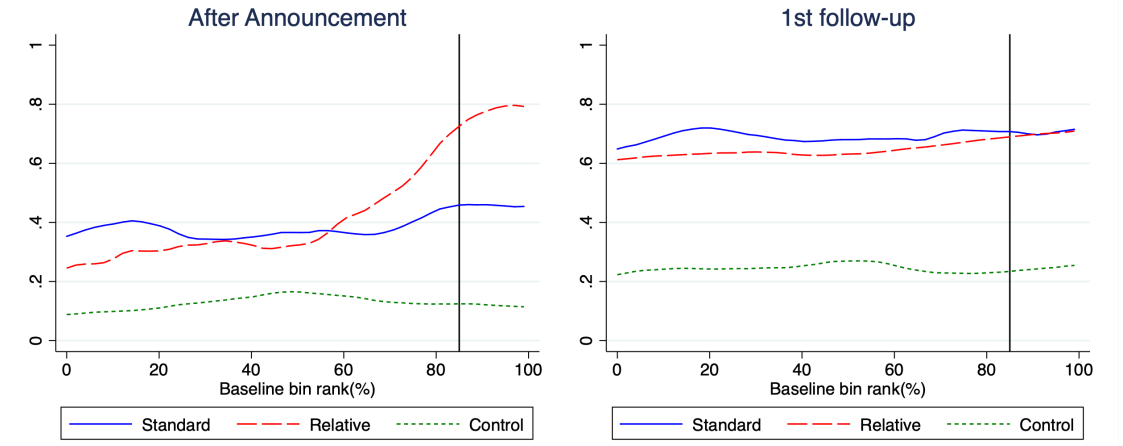
Notes: This figure presents students' levels of understanding measured by the percent of questions answered correctly on quizzes by baseline rank for each study group, immediately after the scholarship announcements and at the time of the first follow-up surveys.

Figure 4: Expectation of the Scholarship

(a) Overall rank(%)



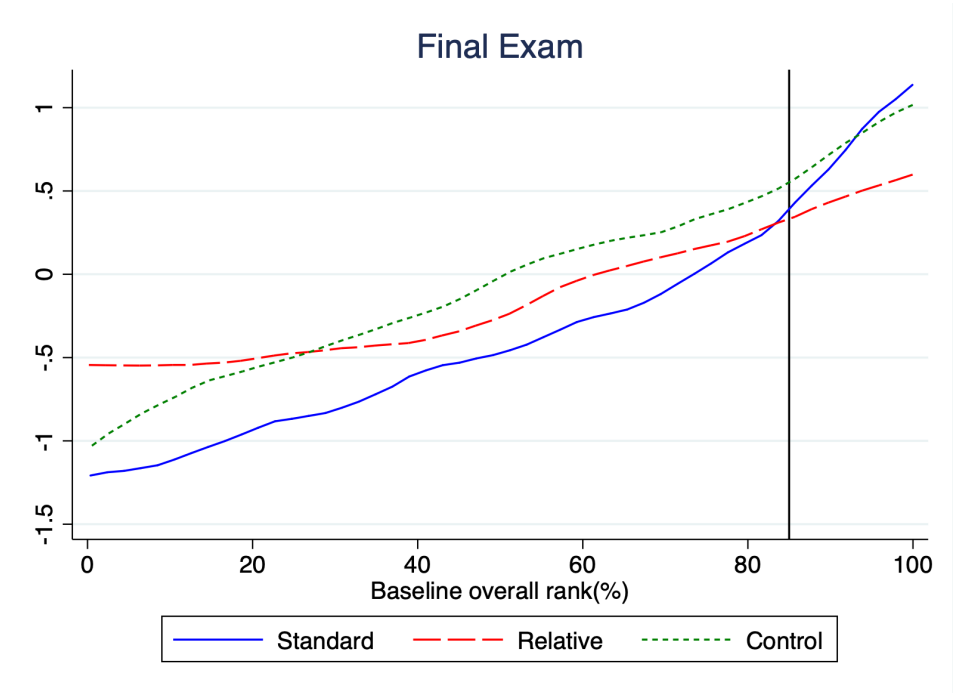
(b) Bin rank(%)



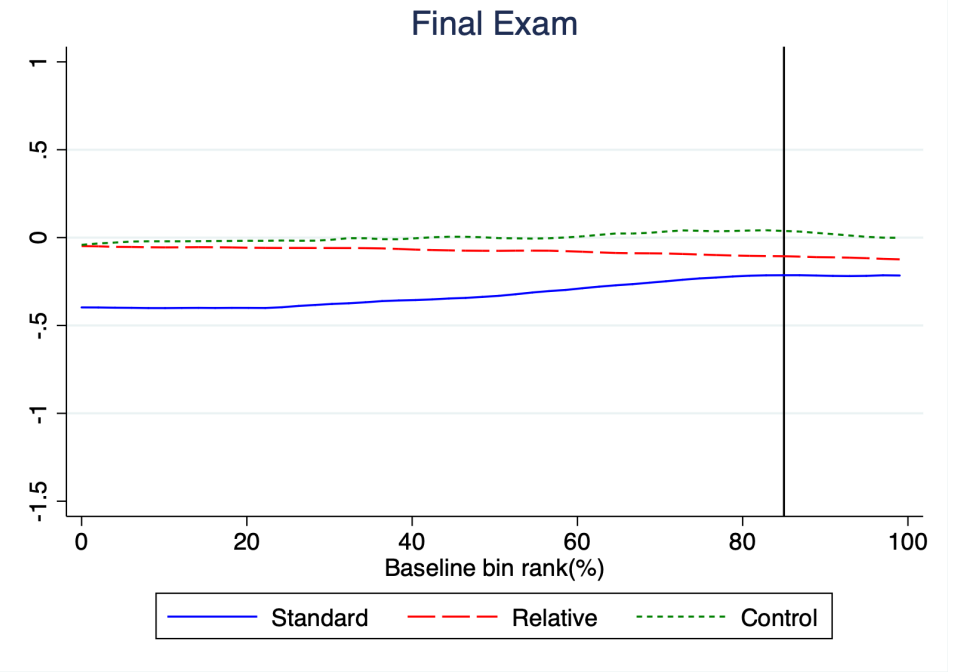
Notes: This figure presents students' expectations of winning the scholarship by baseline rank for each study group, immediately after the scholarship announcements and at the time of the first follow-up survey.

Figure 5: Final Exam Scores by Baseline Rank

(a) Overall rank(%)



(b) Bin rank(%)



Notes: This figure presents average final exam scores by baseline rank for each study group.

Appendices

Contents

A Tables and Figures Referenced in Text

Table A1: Sample Attrition

	Dependent Variable: Participated					
	Sample: Grade 5-8		Sample: Grade 5-6			
	1st Follow-up		1st Follow-up		2nd Follow-up	
	Survey	Exam	Survey	Exam	Survey	Exam
	(1)	(2)	(3)	(4)	(5)	(6)
Standard	-0.019	0.022	-0.012	0.023	0.043	0.036
	(0.017)	(0.015)	(0.019)	(0.014)	(0.032)	(0.039)
Relative	-0.025	0.029**	-0.014	0.027*	0.025	0.043
	(0.017)	(0.014)	(0.021)	(0.015)	(0.034)	(0.033)
N	7385	7385	4562	4562	4393	4393
R-Squared	0.003	0.004	0.000	0.001	0.001	0.001
Mean of Dep. Var.	0.827	0.896	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: Test Score Impacts for Each Subject

	Dep. Var: Test score					
	Chichewa	Math	English	Science	Social studies	Art and Life skills
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Average treatment effects						
Standard	-0.193*	-0.210	-0.291**	-0.113	-0.226	-0.064
	(0.100)	(0.139)	(0.122)	(0.106)	(0.152)	(0.108)
Relative	-0.207*	0.109	-0.169	-0.000	-0.209	0.132
	(0.107)	(0.157)	(0.141)	(0.124)	(0.166)	(0.149)
R-Squared	0.125	0.076	0.141	0.184	0.156	0.143
P-value: *Std = Rel*	0.880	0.008	0.241	0.400	0.897	0.192
Panel B: Heterogeneous treatment effects by overall rank						
Standard	-0.219**	-0.203	-0.323**	-0.173	-0.267*	-0.113
	(0.108)	(0.154)	(0.130)	(0.108)	(0.154)	(0.104)
Relative	-0.189	0.102	-0.132	0.027	-0.189	0.149
	(0.116)	(0.173)	(0.147)	(0.130)	(0.179)	(0.159)
Std. x Top 15%	0.140	-0.078	0.119	0.319*	0.246	0.223
	(0.122)	(0.176)	(0.190)	(0.170)	(0.226)	(0.220)
Rel. x Top 15%	-0.157	-0.168	-0.300	-0.199	-0.196	-0.209
	(0.145)	(0.245)	(0.267)	(0.213)	(0.223)	(0.250)
Top 15%	0.297***	0.554***	0.444***	0.462***	0.650***	0.529***
	(0.094)	(0.148)	(0.161)	(0.144)	(0.169)	(0.195)
R-Squared	0.138	0.104	0.158	0.214	0.203	0.175
Panel C: Hegerogeneous treatment effects by bin rank						
Standard	-0.204**	-0.199	-0.312**	-0.132	-0.234	-0.058
	(0.101)	(0.141)	(0.125)	(0.111)	(0.158)	(0.113)
Relative	-0.216**	0.133	-0.173	0.004	-0.198	0.167
	(0.108)	(0.160)	(0.143)	(0.131)	(0.172)	(0.156)
Std. x Subg. Top 15%	0.069	-0.034	0.141	0.120	0.052	-0.037
	(0.090)	(0.089)	(0.086)	(0.104)	(0.090)	(0.087)
Rel. x Subg. Top 15%	0.056	-0.164*	0.024	-0.028	-0.065	-0.212**
	(0.092)	(0.099)	(0.083)	(0.092)	(0.087)	(0.092)
N	6277	6317	6252	6248	6229	6220
R-Squared	0.126	0.079	0.141	0.185	0.157	0.145
Mean of Dep. Var.	0.054	0.032	0.030	0.028	0.033	0.036

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, 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 A3: Test Score Impacts by Gender

	Exam Rank			Exam Score		
	Girls	Boys	Girls - Boys	Girls	Boys	Girls - Boys
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Average treatment effects						
Standard	-9.795**	-6.327*	-3.468*	-0.384***	-0.245*	-0.139**
	(3.741)	(3.502)	(1.932)	(0.135)	(0.136)	(0.068)
Relative	-3.970	-3.548	-0.422	-0.136	-0.124	-0.011
	(4.744)	(4.445)	(1.879)	(0.188)	(0.185)	(0.067)
R-Squared	0.825	0.825	0.825	0.263	0.263	0.263
Panel B: Heterogeneous treatment effects by overall rank						
Standard	-11.615***	-7.331*	-4.283**	-0.442***	-0.282**	-0.160**
	(3.788)	(3.854)	(2.069)	(0.137)	(0.139)	(0.072)
Relative	-3.667	-1.555	-2.112	-0.107	-0.032	-0.075
	(4.943)	(4.849)	(1.891)	(0.195)	(0.195)	(0.068)
Top 15%	-0.899	5.966	-6.865	-0.046	0.163	-0.209
	(6.338)	(4.369)	(4.488)	(0.234)	(0.189)	(0.166)
Std. x Top 15%	15.156**	4.859	10.296**	0.447*	0.178	0.269
	(6.618)	(5.081)	(4.892)	(0.248)	(0.210)	(0.190)
Rel. x Top 15%	-0.835	-9.560	8.725*	-0.188	-0.436	0.248
	(8.665)	(6.847)	(5.228)	(0.297)	(0.266)	(0.187)
N	6323	6323	6323	6323	6323	6323
R-Squared	0.828	0.828	0.828	0.272	0.272	0.272
Mean of Dep. Var.	51.906	57.674	54.342	-0.120	0.113	-0.022

Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects, the baseline value of the outcome variable, 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 A4: Longer Term Intermediate Outcomes

	Sample: Grade 5-6								
	1st Follow-up					2nd Follow-up			
	Student input		Non-cognitive skills			Student input	Non-cognitive traits		
	Atten-dance	Study Hours	Motiva-tion to study hard	Self esteem	Con-scien tious-ness	Study Hours	Motiva-tion to study hard	Self esteem	Con-scien tious-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: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects, the baseline value of the outcome varlabie, 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 A5: Test Score Impacts (Non-Cognitive Skills Controlled)

	Sample: Grade 5-8			
	Exam Rank		Exam score	
	(1)	(2)	(3)	(4)
Panel A: Average treatment effects				
Standard	-6.739*	-7.010*	-0.236*	-0.245*
	(3.639)	(3.856)	(0.137)	(0.147)
Relative	-2.774	-5.208	-0.048	-0.137
	(4.717)	(4.416)	(0.190)	(0.177)
R-Squared	0.253	0.317	0.270	0.334
P-value: *Std = Rel*	0.349		0.289	
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-8.554**	-8.554**	-0.297**	-0.298*
	(3.881)	(4.133)	(0.142)	(0.154)
Relative	-2.066	-4.677	0.004	-0.093
	(5.075)	(4.790)	(0.199)	(0.186)
Std. x Top 15%	10.724**	8.587*	0.355	0.281
	(5.095)	(5.071)	(0.219)	(0.217)
Rel. x Top 15%	-3.790	-2.573	-0.271	-0.214
	(6.687)	(5.620)	(0.266)	(0.237)
Top 15%	-31.711*	-42.303***	-1.774**	-2.149***
	(17.202)	(14.314)	(0.701)	(0.618)
Additional Controls	No	Yes	No	Yes
N	5829	5596	5829	5596
R-Squared	0.267	0.324	0.286	0.343
Mean of Dep. Var.	52.316	52.437	-0.123	-0.117

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

Table A6: Test Score Impacts by Grade

	Grade 5-7		Grade 8	
	(1)	(2)	(3)	(4)
Panel A: Average treatment effects				
Standard	-0.351**	-0.332*	-0.154	-0.029
	(0.169)	(0.188)	(0.179)	(0.192)
Relative	-0.086	-0.220	0.012	-0.023
	(0.220)	(0.212)	(0.196)	(0.138)
R-Squared	0.234	0.319	0.396	0.475
P-value: *Std = Rel*	0.187	0.469	0.244	0.966
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-0.420**	-0.395**	-0.124	0.058
	(0.173)	(0.197)	(0.192)	(0.194)
Relative	-0.044	-0.189	0.047	-0.003
	(0.229)	(0.224)	(0.198)	(0.134)
Std. x Top 15%	0.476	0.389	-0.222	-0.530*
	(0.308)	(0.295)	(0.261)	(0.302)
Rel. x Top 15%	-0.232	-0.155	-0.251	-0.252*
	(0.358)	(0.313)	(0.194)	(0.146)
Top 15%	-0.019	0.048	0.134	0.143
	(0.287)	(0.277)	(0.141)	(0.120)
R-Squared	0.245	0.326	0.398	0.483
Panel C: Hegerogeneous treatment effects by bin rank				
Standard	-0.348*	-0.324	-0.151	-0.020
	(0.177)	(0.199)	(0.176)	(0.194)
Relative	-0.076	-0.199	0.042	0.022
	(0.228)	(0.224)	(0.199)	(0.147)
Std. x Subg. Top 15%	-0.023	-0.048	-0.018	-0.063
	(0.123)	(0.125)	(0.118)	(0.119)
Rel. x Subg. Top 15%	-0.059	-0.122	-0.188	-0.299**
	(0.125)	(0.123)	(0.136)	(0.144)
Additional controls	No	Yes	No	Yes
N	5955	5159	1351	1164
R-Squared	0.234	0.319	0.398	0.478
Mean of Dep. Var.	-0.234	-0.186	-0.011	0.030

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 A7: 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 competitiveness index
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Average treatment effects						
Standard	0.075	-0.042	0.081	0.075	0.036	0.061
	(0.103)	(0.063)	(0.061)	(0.066)	(0.066)	(0.081)
Relative	-0.208	0.010	-0.049	0.003	-0.053	-0.085
	(0.135)	(0.061)	(0.064)	(0.067)	(0.065)	(0.080)
R-Squared	0.083	0.018	0.021	0.008	0.008	0.038
P-value: *Std = Rel*	0.015	0.171	0.006	0.178	0.121	0.002
Panel B: Heterogeneous treatment effects by overall rank						
Standard	0.121	-0.051	0.081	0.093	0.034	0.075
	(0.109)	(0.073)	(0.052)	(0.066)	(0.076)	(0.084)
Relative	-0.236	-0.013	-0.043	0.039	-0.064	-0.092
	(0.147)	(0.070)	(0.061)	(0.066)	(0.066)	(0.086)
Std. x Top 15%	-0.304*	0.062	0.001	-0.122	0.025	-0.089
	(0.161)	(0.101)	(0.143)	(0.157)	(0.165)	(0.127)
Rel. x Top 15%	0.062	0.119	-0.027	-0.209	0.088	0.017
	(0.175)	(0.110)	(0.150)	(0.191)	(0.195)	(0.151)
Top 15%	0.221	-0.052	-0.010	0.138	-0.139	0.038
	(0.143)	(0.091)	(0.130)	(0.142)	(0.140)	(0.119)
N	2698	2697	2690	2692	2698	2700
R-Squared	0.088	0.019	0.021	0.009	0.010	0.038
Mean of Dep. Var.	3.754	4.072	3.889	3.828	4.096	-0.011

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. The classroom competitiveness index is generated by taking the average of the standardized measures of the other outcomes in the table, where the mean and standard deviation in the control group is used in the standardization. The resulting index is then standardized relative to the control group . * denotes significance at 0.10; ** at 0.05; *** 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	Dis-agree	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 Im 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 dont 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...	Dis-agree strongly	Dis-agree 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

B Feedback Intervention

B.1 Description

The feedback intervention provided rank information on the midterm exam, administered at the end of the second term (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.[26]

This intervention was designed to test whether additional information on student performance could influence the effectiveness of the scholarship schemes. In particular, if students lack precise information on their likelihood of obtaining the scholarship, providing them with feedback on performance may enhance the distributional impacts of the scholarship incentives, encouraging those at the top or discouraging those at the bottom.[27]

At the beginning of the third term (March of 2015), each 5th, 6th, and 7th grade student received a note providing their ranking as of the midterm exam. Figure B1 presents examples of these notes. The feedback treatment group received information on their numeric and percentile rank at the baseline and midterm exams (Panels B1a, B1c, and B1e), while the no-feedback group received information only on the baseline exam (Panels B1b, B1d, and B1f). Feedback differed depending on the scholarship treatment group. In the *Standard* scholarship group and the scholarship control group, students in the feedback treatment received their overall rankings in the midterm exam relative to all students in the program (Panels B1a and B1e). Students in the *Relative* scholarship group received information on their bin rankings in the midterm exam (Panel B1c).

[26]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 term.

[27]While no research that we are aware of examines the interaction of feedback on student ranking and incentives, several of previous papers examine the overall effects of feedback on exam performance (Tran and Zeckhauser, 2012; Azmat and Iriberri, 2010; Ashraf, Bandiera, and Lee, 2014).

B.2 Implementation Issues

As described in the main text, there were two mistakes made in computing ranking information for the feedback intervention that complicates interpretation. First, the bin ranks in the *Relative* Scholarship intervention were calculated incorrectly, and the resulting bin rank provided was not informative of the actual rank. Second, the ranks in the *Standard* Scholarship and control groups were based on the total number of students on the school rosters, rather than the number of students who were in the study sample, which only included those who took the baseline exam and completed the baseline survey. This results in an overstatement of performance, particularly for the lowest-performing children. Based on the procedure used, the minimum rank provided was 22.7 percent. For completeness we present the results of the feedback information in the following section, but these results should be taken with some caution as a result of these issues.

B.3 Analysis

Table B1 presents the difference in means between the feedback and no-feedback treatment groups. Of the 16 variables examined, two variables are significantly different at the 5 percent level and two variable at the 10 percent level.Table B2 displays sample attrition across feedback treatment groups. We do not observe any statistically significant difference between the feedback and no-feedback groups.

Panel A, Column (1) of Table B3 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 B2 plots final exam score by overall midterm exam 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 (2) of Table B3, Panel B, the impacts in the top 15 percent are not significant for either scholarship group or for the control group.

We present additional analyses to explore whether feedback may have been more valuable when it carried a stronger signal about student progress. In our study, students were told their rankings as of the baseline test, and those with a larger difference between midterm and baseline test scores may have responded more strongly to the feedback. Therefore, we examine how feedback influenced students' perceptions of their performance, expectations of winning the scholarship, and the final test scores.[28] Table B4 presents the results of regressions of these outcomes on baseline scores and a dummy for the feedback treatment interacted with the difference between midterm and baseline test scores, for the sample of students in the two scholarship treatments. We find that both baseline test score and the improvement between the baseline and midterm exam are indeed correlated with perceptions of performance, expectations of winning the scholarship, and final test scores. However, we do not find evidence that a larger difference between baseline and midterm scores was associated with a larger feedback effects. This implies that feedback treatment may have conveyed little additional information beyond what students already knew; they may have received and understood their exam performance without the feedback provided within the experiment.

Although we caution against drawing strong conclusions due to implementation issues, 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.

[28]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%)".

B.4 Robustness of Scholarship Impacts to Exclusion of Feedback Group

Because the provided feedback was incorrect, particularly in the *Relative* Scholarship group, it is important to examine whether it may have influenced impact estimates of the scholarship programs. In theory, this could occur if the incorrect information influenced the scholarship treatment and control groups in different ways. However, as we describe, there is no evidence suggesting that this occurred.

We present two pieces of evidence to examine this issue. First, as documented in the previous section, we found no evidence for impacts of feedback in any of the treatment groups, or by student performance. These results imply the scholarship impact estimates were unaffected by the feedback intervention. Second, we can more directly check for the influence of the feedback intervention on the scholarship impacts by re-estimating the scholarship impacts of Table 4 on the randomly-assigned no-feedback sub-sample. These results are shown in Table B5. As shown in the table, the scholarship impacts are largely unchanged when the feedback group is excluded.

Appendix B References

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.

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

Table B1: Balance of Baseline Variables Across Feedback Treatment

	No Feedback Mean	Feedback vs. c	N
	(1)	(2)	(3)
Age	13.8	0.206**	6103
	[4.18]	(0.093)	
Male	0.459	0.013	6103
	[0.498]	(0.013)	
Ethnic group: Chewa	0.890	-0.003	6077
	[0.313]	(0.006)	
Household size	7.88	0.038	6103
	[1.53]	(0.031)	
Asset index	0.027	-0.091*	5848
	[1.94]	(0.051)	
Baseline rank(%)	52.6	-0.248	6061
	[27.9]	(0.590)	
Baseline Score	-0.002	-0.012	6061
	[1.07]	(0.021)	
Attendance	0.837	0.005	6103
	[0.196]	(0.005)	
Study hours per week	15.6	0.168	6031
	[16.1]	(0.373)	
Motivation to study	4.49	0.0008	6092
	[0.838]	(0.021)	
Self-esteem	2.63	0.011	6087
	[0.332]	(0.007)	
Conscientiousness	3.56	0.004	6089
	[0.576]	(0.015)	
Grit	3.15	0.021*	6087
	[0.425]	(0.012)	
Teacher Index	0.095	0.002	6083
	[0.968]	(0.023)	
Parental Effort Index	-0.099	0.053**	6024
	[1.12]	(0.024)	

Notes: Column 1 reports means of baseline variables for subjects assigned to the no feedback group. Column 2 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. Refer to the table note of Table 2 for definition of other variables.* denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table B2: Sample Attrition (Feedback Treatment)

	Dependent Variable: Participated					
	Sample: Grade 5-8		Sample: Grade 5-6			
	1st Follow-up		1st Follow-up		2nd Follow-up	
	Exam	Survey	Exam	Exam	Survey	Exam
	(1)	(2)	(3)	(4)	(5)	(6)
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 B3: 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: Average treatment effects			
Feedback	0.029	0.064	0.015
	(0.024)	(0.052)	(0.028)
R-Squared	0.309	0.240	0.222
Panel B: Interaction effects of scholarhip and feedback			
Feedback	0.051	0.086	0.036
	(0.064)	(0.081)	(0.081)
Standard	-0.323	-0.185	-0.276
	(0.202)	(0.242)	(0.174)
Relative	-0.204	-0.109	-0.105
	(0.227)	(0.247)	(0.210)
Std. x FB	-0.019	-0.010	-0.018
	(0.072)	(0.103)	(0.090)
Rel. x FB	-0.032	-0.039	-0.033
	(0.073)	(0.135)	(0.090)
N	5159	1057	4102
R-Squared	0.319	0.245	0.232
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.

Table B4: Self-evaluation, Expectations and Exam Scores by the Intensity of Feedback

	Sample: Grade 5-7		
	Self-evaluation	Expectation	Exam score
	(1)	(2)	(3)
Feedback	0.006	-0.009	0.031
	(0.027)	(0.016)	(0.027)
Mid-Base	0.007***	0.003***	0.023***
	(0.001)	(0.001)	(0.003)
Feedback * (Mid-Base)	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)
Baseline Score	0.210***	0.071***	0.762***
	(0.023)	(0.012)	(0.077)
N	3653	3627	3926
R-Squared	0.095	0.032	0.461
Mean of Dep. Var.	3.237	0.667	-0.186

Notes: The sample includes students in the *Standard* and *Relative* scholarship groups. Self-evaluation is the students perception of performance within the classroom. Responses are scaled of 1 to 5, ranging from [very bad (0-20%)] to [very good (81-100%)]. 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?] Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects, zone fixed effects, baseline value of dependent variables, and demographic controls such as age, ethnic group, 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.

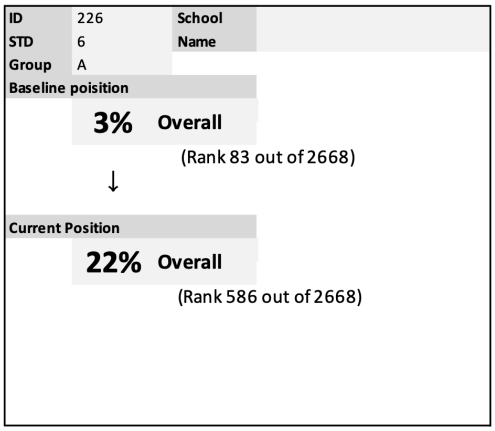
Table B5: Test Score Impacts on No Feedback Group

	Sample: Grade 5-8			
	Exam Rank		Exam Score	
	Full Sample	No Feedback	Full Sample	No Feedback
	(1)	(2)	(3)	(4)
Panel A: Average treatment effects				
Standard	-7.368*	-9.612*	-0.266*	-0.328
	(3.868)	(5.163)	(0.146)	(0.199)
Relative	-4.730	-7.201	-0.126	-0.208
	(4.404)	(5.577)	(0.174)	(0.223)
R-Squared	0.305	0.301	0.324	0.305
P-value: *Std = Rel*	0.447	0.520	0.337	0.449
Panel B: Heterogeneous treatment effects by overall rank				
Standard	-8.682**	-12.029**	-0.305**	-0.418**
	(4.138)	(5.511)	(0.153)	(0.210)
Relative	-4.016	-7.352	-0.073	-0.178
	(4.769)	(6.055)	(0.184)	(0.236)
Std. x Top 15%	7.507	15.209*	0.224	0.570
	(5.316)	(8.270)	(0.230)	(0.357)
Rel. x Top 15%	-4.348	1.295	-0.299	-0.122
	(6.057)	(8.289)	(0.253)	(0.352)
Top 15%	3.847	-0.776	0.118	-0.014
	(4.730)	(7.367)	(0.209)	(0.326)
R-Squared	0.312	0.312	0.330	0.317
P-value: *Std = Rel* at Bot. 85%	0.211	0.249	0.124	0.145
P-value: *Std = Rel* at Top 15%	0.086	0.037	0.125	0.028
Panel C: Hegerogeneous treatment effects by bin rank				
Standard	-7.360*	-9.415*	-0.266*	-0.331
	(3.982)	(5.477)	(0.151)	(0.212)
Relative	-4.423	-6.863	-0.109	-0.200
	(4.527)	(5.869)	(0.180)	(0.236)
Std. x Subg. Top 15%	0.038	-1.386	0.003	0.021
	(2.270)	(4.602)	(0.090)	(0.175)
Rel. x Subg. Top 15%	-1.877	-2.014	-0.106	-0.033
	(2.227)	(4.135)	(0.088)	(0.156)
Controls	Yes	Yes	Yes	Yes
N	6323	2568	6323	2568
R-Squared	0.305	0.302	0.324	0.305
Mean of Dep. Var.	51.489	51.044	-0.146	-0.198
P-value: *Std = Rel* at Bot. 85%	0.406	0.514	0.289	0.424
P-value: *Std = Rel* at Top 15%	0.770	0.604	0.737	0.626

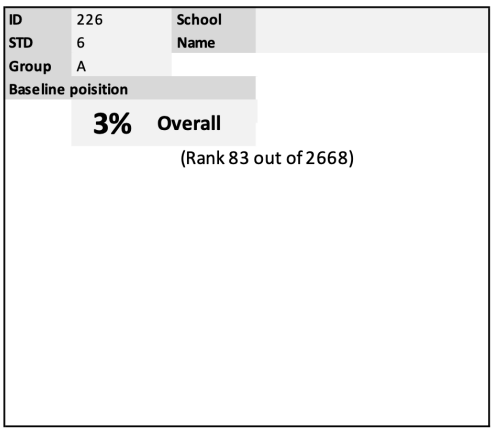
Notes: Standard errors, clustered at the school-grade level, are in parentheses. All specifications include grade fixed effects, the baseline value of the outcome variable, zone fixed effects, age, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05;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

Figure B1: Feedback Note

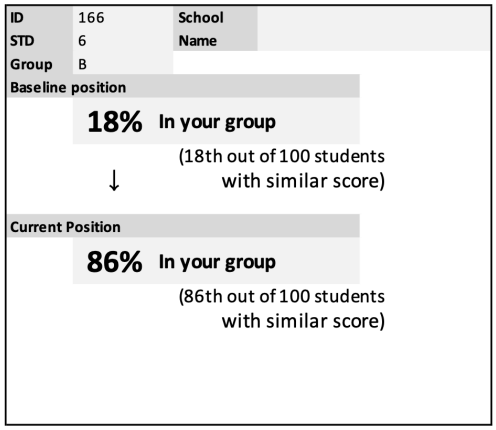
(a) Feedback and *Standard*



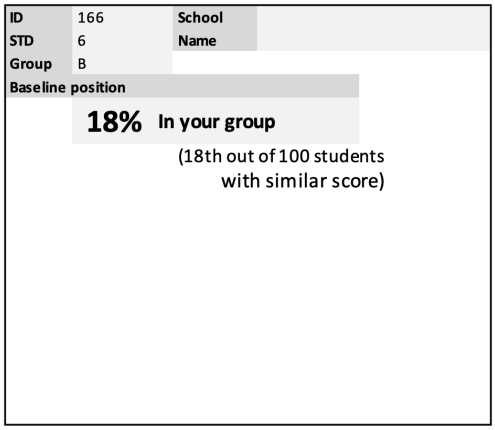
(b) No Feedback and *Standard*



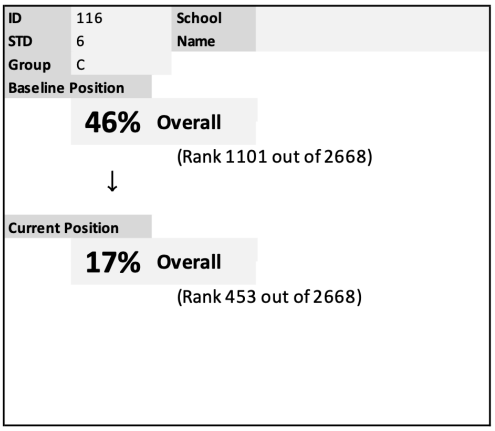
(c) Feedback and *Relative*



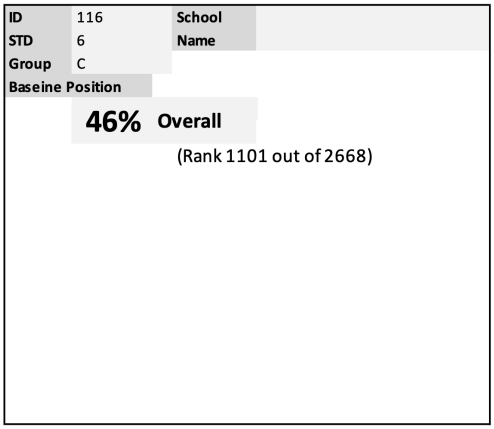
(d) No Feedback and *Relative*



(e) Feedback and Control



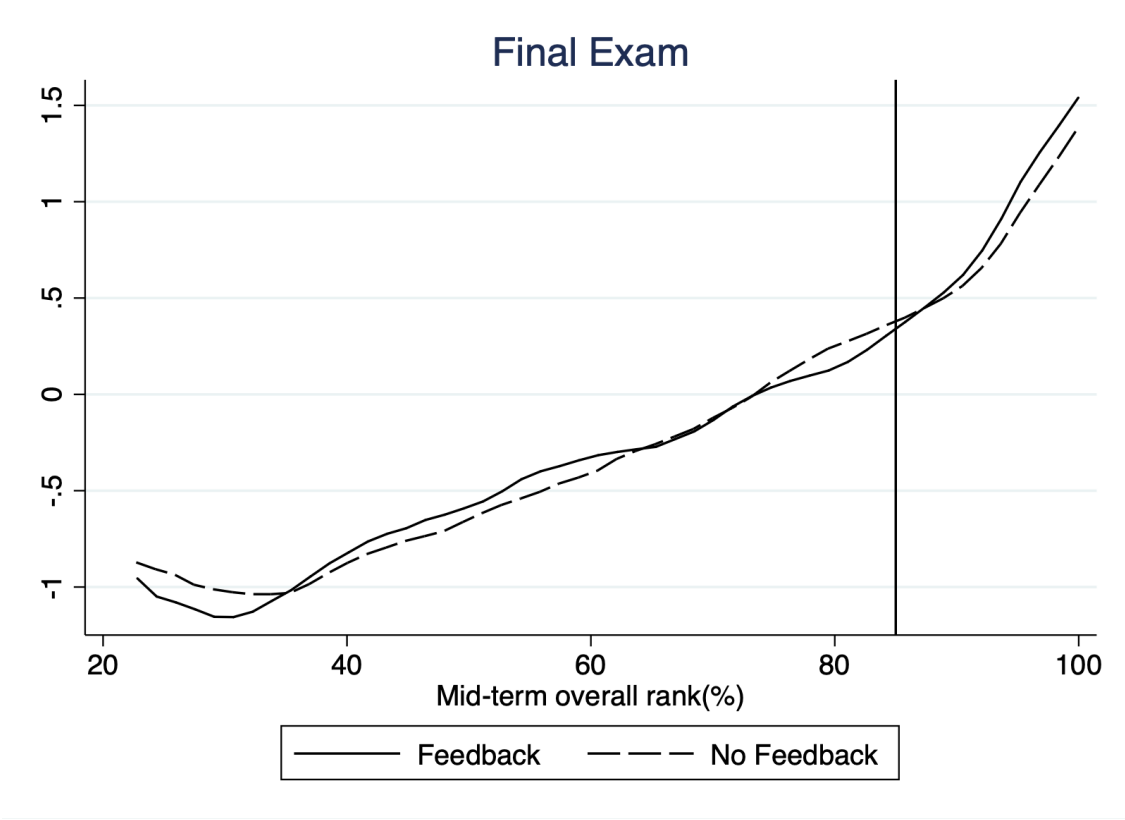
(f) No Feedback and Control



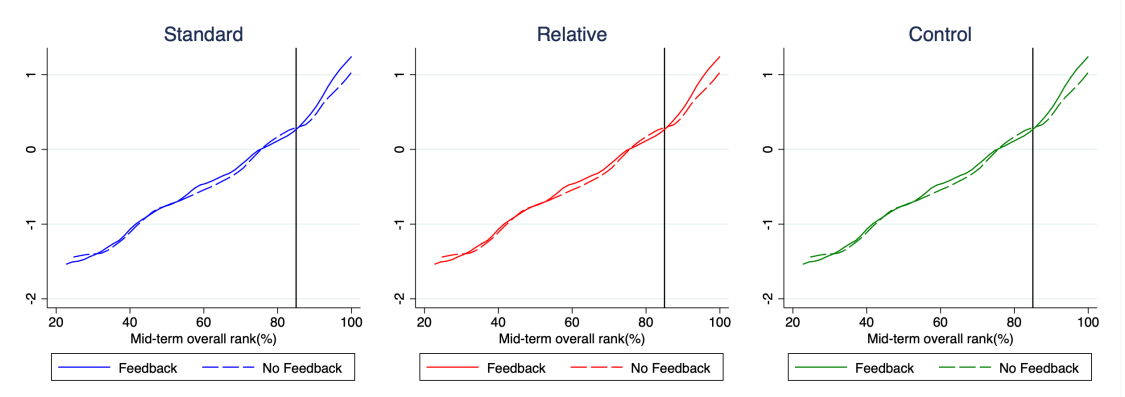
Notes: This figure displays the feedback notes that students received in the second term. The left column presents feedback notes given to the feedback treatment group and the right column presents feedback notes given to the control group. The feedback treatment group received information on their rank 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 B2: Feedback Effect on Final Exam Score by Midterm Rank

(a) Whole sample



(b) By treatment group



Notes: This figure presents average final exam scores by midterm overall rank. Panel A presents the results for all students, and Panel B presents the results by scholarship treatment status.

C Attrition

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 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 C1, 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 C2. 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.

Appendix C References

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

Table C1: 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)
Standard	-7.402**	-8.321**	-6.707*	-0.265*	-0.304**	-0.232*
	(3.671)	(3.671)	(3.671)	(0.138)	(0.138)	(0.138)
Relative	-2.516	-3.724	-1.374	-0.045	-0.122	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 4 of Table A1. 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 C2: Attrition on First Follow-up Exam by Scholarship Treatment and Baseline Test Score

		Baseline Variable			
		Baseline Score		Top 15 Percent	
	(1)	(2)	(3)	(4)	(5)
Standard	0.022		0.024		0.023
	(0.015)		(0.015)		(0.015)
Relative	0.029**		0.028**		0.026*
	(0.014)		(0.014)		(0.014)
Baseline		0.032***	0.028**	0.036***	0.034
		(0.005)	(0.012)	(0.009)	(0.022)
Baseline × *Standard*			0.001		-0.008
			(0.014)		(0.026)
Baseline × *Relative*			0.007		0.010
			(0.015)		(0.027)
N	7385	7342	7342	7385	7385

Notes: Each column regresses attrition on the first follow-up exam on the variables indicated. Standard errors, in parentheses, are clustered at the school-grade level. All specifications include grade fixed effects. Columns 2 and 3 use the continuous measure of the baseline test score, while Columns 4 and 5 use a dummy indicating whether the student was in the top 15 percent at baseline. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.