Journal of Development Economics 158 (2022) 102893

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Regular article

When student incentives do not work: Evidence from a field experiment in Malawi☆

James Berry [a], Hyuncheol Bryant Kim [b,*], Hyuk Harry Son [c,d]

[a] University of Delaware, United States of America
[b] Hong Kong University of Science and Technology, Hong Kong
[c] Utrecht University, The Netherlands
[d] Cornell University, United States of America

ARTICLE INFO

JEL classification:
I21
O15

Keywords:
student incentives
Education policy
Field experiments

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 *Population-based* scholarship that rewarded overall top performers on an exam and a *Bin-based* scholarship that rewarded the top performers within smaller groups of students with similar baseline scores. We find that the Population-based 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 Bin-based scholarship program.

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 (Fryer, 2017). Standard economic theory predicts that financial incentives can induce student effort and thereby increase academic outcomes. However, there are several potential drawbacks. For example, incentives for test score performance may shift focus towards the test and away from learning the underlying concepts. Incentives for some subjects may also lead students to reallocate effort and learn less other subjects. Finally, incentives may crowd out intrinsic motivation to learn, ultimately decreasing performance (Bénabou and Tirole, 2006; Gneezy et al., 2011). Because impacts of incentive programs vary widely across studies (Fryer, 2017), understanding why incentives do and do not work remains an important area for research.

A common incentive scheme in education is an individual tournament in which the top performing students on an exam are provided with a reward. This scheme may be appealing because they allow policy makers to set a fixed budget for the incentives. However, tournament schemes in which relatively few students receive the reward may induce effort only from top students. 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.

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.[1] The first, which we call the *Population-based* scholarship scheme, provided a scholarship to students in the sample who scored in the top 15 percent on the final end-of-year exam in the sub-district. This scholarship

☆ The authors 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. The authors 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. This project has been implemented by the Malawi office of Africa Future Foundation (AFF) with financial support from the Korea International Cooperation Agency. This research is approved by 1) National Commission for Science and Technology in Malawi (Protocol Number: P.12/14/25) and 2) Institutional Review Board, Cornell University (IRB approval number: 1505005578). The data and replication materials are posted to the Harvard Dataverse (Berry et al., 2021).

* Corresponding author.
E-mail addresses: jimberry@udel.edu (J. Berry), hbkim@ust.hk (H.B. Kim), h.son@uu.nl (H.H. Son).

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

https://doi.org/10.1016/j.jdeveco.2022.102893
Received 5 August 2021; Received in revised form 4 May 2022; Accepted 11 May 2022
Available online 24 May 2022

scheme is similar to that of Kremer et al. (2009), in which scholarships were given to the top 15 percent of 6th grade female students in primary schools in Kenya.

In the second scholarship scheme, the *Bin-based* scholarship, students were grouped into bins by baseline test score, and the top 15 percent of students within each bin received the incentive. Because students compete only with others that have similar baseline test scores, initially low-performing students are more likely to receive the rewards compared with a standard tournament. We hypothesized that this scheme would increase effort and reduce discouragement that may accompany the Population-based scholarship. In addition, like a standard tournament incentive, the Bin-based 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) which proposes a similar scheme for teachers, which they call "pay for percentile".[2]

We implemented a randomized trial where 5th to 8th grade classrooms were assigned to Population-based and Bin-based 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 nine 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.

We find that the Population-based scholarship scheme reduced final exam scores by 0.27 standard deviations (SDs) across the full sample, with the largest negative impacts on students with low initial test scores. The Population-based scholarship scheme also reduced survey-measured motivation of the students, again with the results concentrated among the initially lowest-performing students. By contrast, the Bin-based scholarship scheme did not have statistically significant impacts on test score performance or motivation, with small and negative point estimates. Although our study lacks power to detect statistically significant differences between the impacts of the Population-based and Bin-based scholarships on average, point estimates suggest that the students in the Bin-based group performed better than those in the Population-based 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 Population-based 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 signaling that high performance is valuable. This is related to research on stereotype threat, in which revealing one's social identity can lead individuals to conform to negative stereotypes. For example, Hoff and Pandey (2006) finds that in mixed-caste classrooms in India, caste revelation lowers the performance of low-caste students.

We contribute to two literatures. First, we contribute to the growing literature on incentives to learn in education. Table A.1 presents a summary of evidence on incentives to learn from studies using randomized controlled trials. As shown in the table, evidence on the overall effectiveness of incentives in increasing learning is mixed, across both developed and developing countries and ages of students targeted. While a number of studies find no evidence of effects of incentives on learning, no study that we know of finds significant negative effects, as we find for the Population-based scholarship.

Table A.1 also contains previous studies' findings on heterogeneity by baseline test scores or other measures of ability. Again, the evidence is mixed, with some studies finding larger effects for higher ability students, some finding larger effects for lower ability students, and others finding no evidence of heterogeneity. Of particular note is Leuven et al. (2010) which examines financial rewards given to Dutch University students for passing first-year requirements. Similar to our findings, the authors find positive impacts for high-ability students and negative impacts on low-ability students.

The Population-based scholarship scheme in this study is similar to Kremer et al. (2009), which evaluates 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 an endline exam. The authors find that the program increased test scores both for the targeted girls and for boys who were not eligible for the program. Impacts on girls persisted even after the incentives were removed. Our Population-based 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 and percentile rank. This has important implications on sustainability of merit-based scholarship programs because, even though students may be unaware of their relative score initially, they would learn it if the scheme were repeated in a future period.

Comparing incentive schemes across contexts and domains can prove problematic. Therefore, a smaller but growing literature evaluates the structure of incentives by comparing multiple schemes within the same experiment. For example, studies have compared group and individual incentives (Li et al., 2014; Blimpo, 2014), incentives for effort versus for achievement (Hirshleifer, 2021), incentives targeted to parents versus to children (Berry, 2015), and incentives for students versus for teachers (Behrman et al., 2015). We similarly test different designs within the same context, specifically incentives to top performers for an entire class versus top performers within strata determined by baseline 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 et al., 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. Moreover, List et al. (2018) provides evidence that while incentives for exam performance decreased intrinsic motivation in the short term, these effects did not persist in the longer term.

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).

[2] 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., 2019; Mbiti et al., 2019; Gilligan et al., 2022). The structure is closely related to schemes that provide incentives based on improvement relative to baseline (Behrman et al., 2015; Berry, 2015).

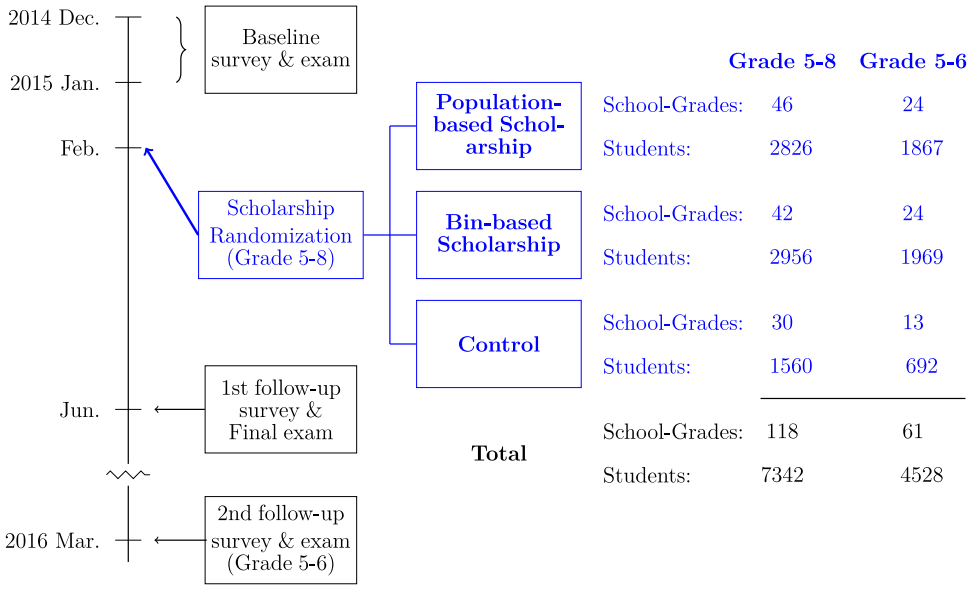


Fig. 1. Experimental timeline and sample composition by treatment category.

Schools are characterized by high pupil–teacher ratios and low levels of infrastructure.[3]

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 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.[4] 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.

2.2.1. Study design

The project chronology is summarized in Fig. 1. The baseline survey and baseline exams were implemented during the first term of the 2014–2015 academic year (December 2014 to January 2015).[5] 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.[6]

Fig. 1 also 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 Population-based scholarship, the Bin-based scholarship, or the control group.[7] 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.

Fig. 2 provides examples of notes for each treatment group as well as the control group. For the Population-based scholarship group, information on the student's overall sub-district rank (hereafter overall rank) as well as the scholarship eligibility condition (top 15 percent) was provided (Panel A). For the Bin-based 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 (Panel B). For the control group, only information on the student's overall rank was provided (Panel C).

The first follow-up survey and final exams took place at the end of the third term (June 2015).[8] The final exam determined eligibility for the scholarships. Awards were distributed 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 Population-based scholarship scheme, within each grade, students scoring in the top 15 percent in the sub-district on the final

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

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

[5] 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.

[6] After the March 2016 exam, we conducted a second-year trial in which we randomly assigned students to the Bin-based 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.

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

[8] As we describe in the next section, we used the PSLCE exam for eighth graders, which took place in May 2015.

(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:			

Fig. 2. Scholarship randomization result announcement note.
Note: Panels (a), (b), and (c) show the scholarship program announcement notes that were given to students assigned to the *Population-based* scholarship group, the *Bin-based* scholarship group, and the control group, respectively.

exam were eligible to receive the award. Under the Bin-based 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.[9]

The awards for Population-based and Bin-based scholarships were identical. The award was a choice among a cash award of USD 9.70 (MWK 4,500) or an in-kind award including a pair of shoes, a school bag, or a school uniform of similar value.[10,11] This represents a significant amount considering that Malawi GDP per capita was only around USD 362.7 in 2014 (World Bank, 2015).

To ensure that students fully understood the scholarship programs (particularly the Bin-based 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 whether hypothetical students would receive the scholarship given their treatment groups and rankings in the final exam. To measure expectations of winning a scholarship, we asked students their perceived likelihood of receiving the scholarship after providing them with the individualized announcements.

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.[12] 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 Online Appendix B. Online Appendix Table B5 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.[13] In addition to the exams, we measured students' school attendance through unannounced checks. These checks were conducted every month between April 2014 and June 2015, four times before the scholarship announcement and four times after.

We also conducted surveys of students at the time of the baseline exams and right before the follow-up exams. A primary objective of the surveys was to measure non-cognitive skills – 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).[14] 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.[15] To

[9] The bottom bin contained 86 students because of integer constraints.

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

[11] The value of the award is comparable to that of Kremer et al. (2009) and Blimpo (2014) where the awards were valued at USD 6.4 and 10, respectively.

[12] 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.

[13] For 8th graders who took the PSLCE instead of the regular final exam, we were able to obtain letter grades for each subject, not a raw test score. The score and overall rank for the reward were calculated by treating A, B, C, D, and F as 6, 5, 4, 3, and 1.

[14] Survey questions used to measure self-esteem, grit, and conscientiousness are shown in Online 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.

[15] 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.

measure impacts on overall non-cognitive skills, we aggregate all four measures into an index, following the method of Kling et al. (2007).[16]

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 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 7,342 students (76.6 percent) who participated in both the baseline survey and baseline exam.

Table 1 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 48.7 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 1 show tests of differences in means between the scholarship groups and the control group. Of the 30 differences examined, only one is statistically significantly at the 10% level. These differences are also not jointly statistically significant (p-values of 0.44 and 0.30 for control vs. Population-based and Bin-based, respectively.

Online Appendix Table A1 displays sample attrition across treatment groups. Out of the control group baseline sample, 83 and 88 percent of the study sample participated in the follow-up survey and final exam, respectively. For the longer-term follow-up survey and exam, 60 and 53 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 Bin-based scholarship group are 3.0 percentage points more likely to take the final exam (statistically significant at the 5 percent level). In Online 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.

Table 1
Balance of baseline variables across treatment groups.

	Control Mean	Pop-based vs. control	Bin-based vs. control	N
	(1)	(2)	(3)	(4)
Age	14.4	0.053	0.158	7342
	[3.60]	(0.179)	(0.188)	
Male	0.487	0.004	−0.018	7342
	[0.500]	(0.020)	(0.018)	
Ethnic group: Chewa	0.914	−0.038	−0.041	7315
	[0.281]	(0.034)	(0.033)	
Household size	7.81	0.301	0.269	7342
	[1.66]	(0.354)	(0.345)	
Asset index	−0.003	0.008	0.046	7063
	[1.89]	(0.177)	(0.176)	
Baseline rank (%)	51.5	−0.625	1.47	7342
	[27.3]	(3.28)	(3.99)	
Baseline Score	0.000	−0.021	0.066	7342
	[0.999]	(0.127)	(0.159)	
Attendance	0.863	0.003	−0.002	7342
	[0.196]	(0.016)	(0.016)	
Study hours per week	16.8	−0.511	−0.212	7265
	[16.4]	(0.790)	(0.790)	
Motivation to study	4.53	−0.014	0.054	7331
	[0.788]	(0.058)	(0.051)	
Self-esteem	2.67	−0.012	−0.005	7325
	[0.339]	(0.022)	(0.022)	
Conscientiousness	3.58	0.028	0.097	7327
	[0.600]	(0.060)	(0.060)	
Grit	3.21	−0.026	−0.009	7325
	[0.450]	(0.020)	(0.024)	
Teacher Effort Index	0.001	0.150	0.252*	7321
	[1.00]	(0.140)	(0.134)	
Parental Effort Index	0.000	−0.025	−0.008	7238
	[1.000]	(0.071)	(0.064)	
P-value of Joint F-test:		0.44	0.30	

Notes: Column 1 reports means for students assigned to the control group. Columns 2 and 3 report mean differences between the scholarship treatment groups and the control group, controlling for grade fixed effects. Standard deviations are in brackets, and standard errors, clustered at the school-grade level, are in parentheses. Baseline score is normalized using the control group mean and standard deviation. The asset index is constructed as the first principal component of variables indicating the ownership of 26 assets. 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). Teacher and parental effort indices are standardized averages of seven and four standardized measures, respectively. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

3. Empirical strategy

To estimate the average impacts of the Population-based and Bin-based scholarship programs, we use the following equation:

$$Y_{igsz1} = \beta_0 + \beta_1 \text{Population-based}_{gsz} + \beta_2 \text{Bin-based}_{gsz} + \delta Y_{igsz0} + \phi X_{igsz} + \eta_g + \gamma_z + \epsilon_{igsz} \quad (1)$$

where Y_{igsz1} is the outcome of interest for student i in grade g in school s in school zone z. *Population-based* and *Bin-based* are indicators for being Population-based and Bin-based 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 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 Bin-based 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:

$$\begin{aligned} Y_{igsz1} = {} & \beta_0 + \beta_1 \text{Population-based}_{gsz} + \beta_2 \text{Bin-based}_{gsz} + \beta_3 Top15_{igsz0} \\ & + \beta_4 \text{Population-based}_{gsz} \cdot Top15_{igsz0} \\ & + \beta_5 \text{Bin-based}_{gsz} \cdot Top15_{igsz0} + \delta Y_{igsz0} \\ & + \eta_g + \gamma_z + \phi X_{igsz} + \epsilon_{igsz} \end{aligned} \quad (2)$$

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 Population-based and Bin-based scholarships on the bottom 85 percent of students, and β_4 and β_5 capture the differences in the impacts of the Population-based and Bin-based scholarship group between the top 15 and bottom 85 percent of students. In addition

[16] 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.

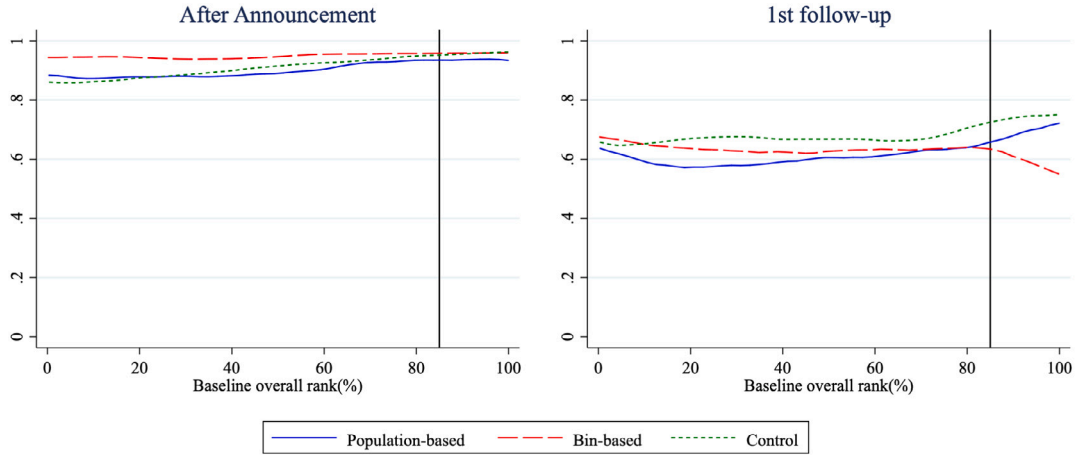


Fig. 3. Understanding of the program.
Notes: This figure presents students' levels of understanding measured by the percent of questions answered correctly on quizzes by baseline rank for each study group, immediately after the scholarship announcements and at the time of the first follow-up surveys.

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 Bin-based 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:

$$\begin{aligned} Y_{igsz1} = & \sum_{k=1,2,3,4,5} \theta_k \cdot k^{th}\text{Quintile}_{igsz0} \\ & + \sum_{k=1,2,3,4,5} \theta_k^p \cdot \text{Population-based}_{gsz} \cdot k^{th}\text{Quintile}_{igsz0} \\ & + \sum_{k=1,2,3,4,5} \theta_k^b \cdot \text{Bin-based}_{gsz} \cdot k^{th}\text{Quintile}_{igsz0} \\ & + \delta Y_{igsz0} + \eta_g + \gamma_z + \phi X_{igsz} + \epsilon_{igsz} \end{aligned} \tag{3}$$

where kth 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 Population-based and Bin-based scholarship treatments, so that the coefficients θ_k^p and θ_k^b capture the impacts of the Population-based and Bin-based 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 2.2.2, AFF provided one-hour introduction sessions to all students to ensure students fully understood the scholarship schemes. We measured 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.

Columns (1) and (2) of Table 2 present regressions corresponding to Eqs. (1) and (2), using the percent of questions about the scholarship schemes answered correctly at program announcement and at first follow-up as outcome variables. Fig. 3 examines heterogeneity by baseline rank non-parametrically by graphing the percent of questions answered correctly (y-axis) by overall baseline rank (x-axis) for each scholarship treatment group. 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 2 shows that there are no statistically significant differences in students' understanding between the scholarship and control groups either right after the program announcement or right before the endline exam.

To examine expectations of winning the scholarship, we present regressions of expectations on treatment status in Columns (3) and (4) of Table 2, with graphs of expectations by baseline rank in Fig. 4.[17] As with understanding of the scholarship, we have measures of expectations at two points: immediately after the scholarship announcement and at the time of the first follow-up survey. Overall, expectations are much higher in both scholarship groups, compared with the control group: as shown in Columns (3) and (4) of Panel A in Table 2, students in the scholarship groups were 30 to 44 percentage points more likely to expect the scholarship. It is worth noting, however, that control group students' expectations of winning the scholarship was non-negligible: 13 percent of control group students expected the scholarship at the time of program announcement, increasing to 24 percent by the first follow-up survey. To the extent that control-group students changed their effort as a result, this could mute the overall effects of the scholarship treatments. Students were also generally optimistic about their chances of winning. Over 40 percent of students in the both scholarship groups expected to win after the program announcement, rising to over 60 percent at the first follow-up survey.

Importantly, we find heterogeneity in expectations by baseline scores that generally align with the structure of the schemes. In the Population-based scholarship group, expectations increase with baseline rank, with a sharp increase for those with baseline ranks above the 85-percent cutoff. Those in the top 15 percent in the Population-based scholarship group were statistically significantly more likely to expect the scholarship than students with the same ranking in the control group, with point estimates of 49 and 21 percentage points, at program announcement and follow-up, respectively. We observe a similar increase in expectations by bin rank in the Bin-based scholarship group at the time of program announcement, although this pattern is not present at the first follow-up survey. We also observe no such patterns in the control group: as shown in Fig. 4, control group expectations do not increase with baseline rank, implying that the non-trivial control-group expectations discussed above are unlikely to affect our findings of heterogeneity by baseline score.

[17] 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?"

Table 2
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				
Population-based	−0.009	−0.021	0.301***	0.442***
	(0.023)	(0.023)	(0.057)	(0.043)
Bin-based	0.036	−0.027	0.358***	0.407***
	(0.022)	(0.024)	(0.066)	(0.044)
R-Squared	0.038	0.092	0.097	0.135
P-value: *Std = Rel*	0.007	0.821	0.330	0.136
Panel B: Heterogeneous treatment effects by overall rank				
Population-based	−0.008	−0.019	0.230***	0.406***
	(0.026)	(0.023)	(0.059)	(0.046)
Bin-based	0.041*	−0.010	0.386***	0.412***
	(0.024)	(0.024)	(0.066)	(0.046)
Pop. × Top 15%	−0.015	−0.018	0.485***	0.211***
	(0.025)	(0.035)	(0.084)	(0.045)
Bin × Top 15%	−0.040*	−0.107***	−0.135	−0.031
	(0.022)	(0.029)	(0.083)	(0.053)
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	0.146
Panel C: Heterogeneous treatment effects by bin rank				
Population-based	−0.011	−0.026	0.290***	0.443***
	(0.023)	(0.023)	(0.057)	(0.044)
Bin-based	0.033	−0.030	0.294***	0.396***
	(0.021)	(0.024)	(0.066)	(0.045)
Pop. × Subg. Top 15%	0.009	0.025	0.080*	−0.004
	(0.017)	(0.026)	(0.044)	(0.041)
Bin × Subg. Top 15%	0.015	0.017	0.395***	0.064
	(0.016)	(0.025)	(0.063)	(0.042)
Controls	Yes	Yes	Yes	Yes
N	5592	5822	5588	5721
R-Squared	0.038	0.092	0.136	0.137
Control Mean	0.915	0.684	0.126	0.243

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, zone, and age fixed effects, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

4.2. Test scores

We now turn to the impacts of the scholarship programs on test scores. Panel A of Table 3 presents the results of estimating Eq. (1) on overall rank (Columns (1) and (2)) and normalized test scores (Columns (3) and (4)).[18] The Population-based scholarship had substantial negative impacts on student performance: students performed 0.27 SDs worse than those in the control group (statistically significant at the 10 percent level). The effects of the Bin-based scholarship were not statistically significant, with negative point estimates ranging from −0.05 to 0.13 SDs. Although the point estimates suggest a substantially larger negative reaction to the Population-based scholarship compared to the Bin-based 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 Fig. 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 Population-based 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 Population-based scholarship, the impacts of the Bin-based 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.[19]

Panel B of Table 3 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 Eq. (2). These results confirm that the decrease in academic achievement in the Population-based treatment is driven by students with initial test scores in the bottom 85 percent: the coefficient on Population-based scholarship is negative and statistically significant, and that on the interaction between Population-based scholarship and being in the top 15 percent at baseline is of opposite sign and larger than the main coefficient on the Population-based scholarship, although it is not

[18] 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).

[19] The negative impacts of the Population-based scholarship were statistically significantly larger among girls than boys (See Online Appendix Table A3).

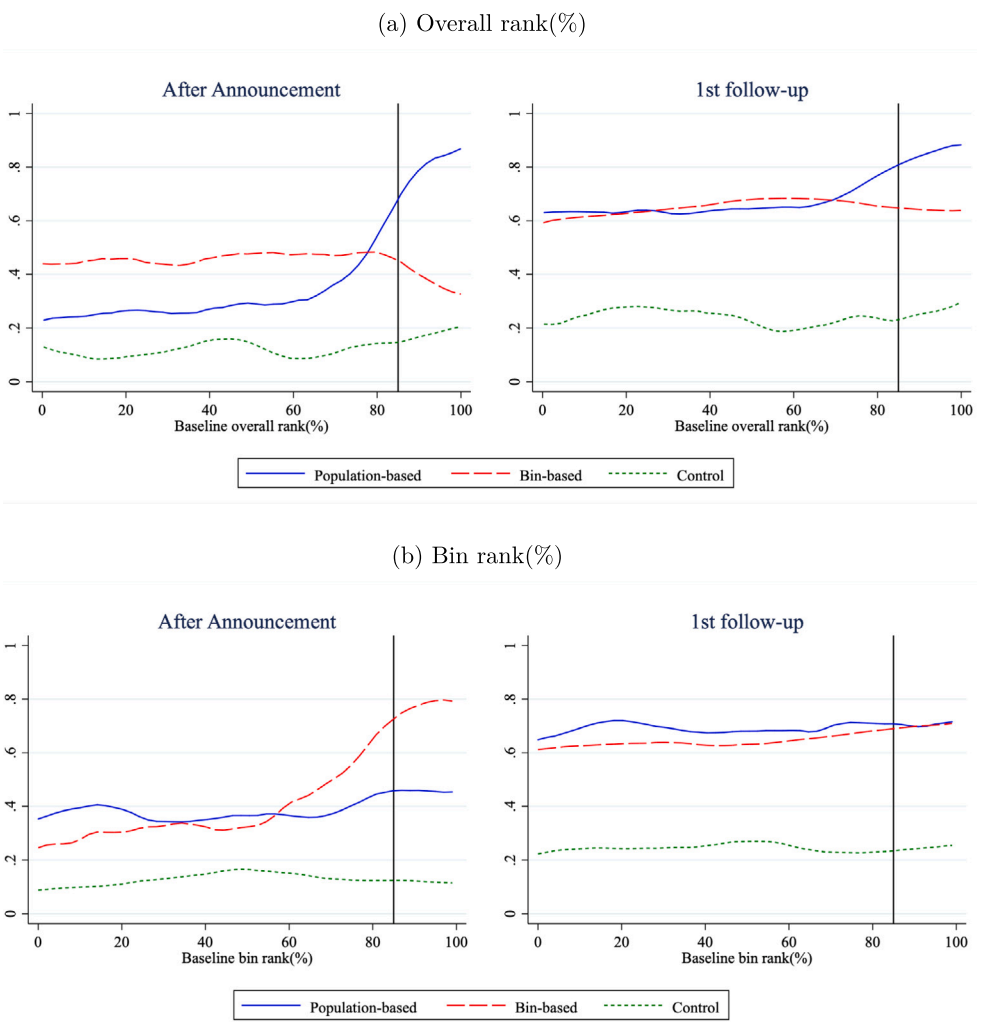


Fig. 4. Expectation of the scholarship.
Notes: This figure presents students' expectations of winning the scholarship by baseline rank for each study group, immediately after the scholarship announcements and at the time of the first follow-up survey.

statistically significant. By contrast, the coefficient on the interaction of the Bin-based 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. In our specification with controls, we cannot reject that the impacts of the Population-based and Bin-based scholarships are equal in each initial performance level, with p-values of the test score for equality of 0.12 and 0.13 for the bottom 85 percent and the top 15 percent, respectively.

Table 4 presents an analysis of heterogeneity by students' initial ranks in more detail, using the series of indicators for being in each quintile as of the baseline test instead of in the top 15 percent, following Equation (3). The results confirm that the negative impacts of the Population-based scholarship program are concentrated in the lower quintiles: coefficients on the interaction of the Population-based treatment (θ_k^p) 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 Bin-based 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 ($\theta_k^p - \theta_k^b$), which is the largest in the lowest quintile (0.52 SDs, statistically significant at the 10 percent level).

Finally, we examine whether the impacts vary by baseline bin rank – that is, the ranking within the 100-student subgroups used to award the Bin-based scholarship. In Panel B of Fig. 5, we plot performance for the two scholarship groups and control groups across the distribution of baseline bin rank. We do not observe differential impacts for those with higher initial ranks within these bins, even for the Bin-based scholarship scheme. These results are confirmed in Panel C of Table 3, 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.[20]

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 5, with average impacts in Panel A and heterogeneity by overall baseline rank in Panel B.

[20] Online Appendix Table A2 presents the analysis of Table 3 by subject. Results are largely similar across the subjects.

Table 3
Test score impacts.

	Sample: Grade 5–8			
	First follow-up			
	Exam rank		Exam score	
	(1)	(2)	(3)	(4)
Panel A: Average treatment effects				
Population-based	−7.402**	−7.368*	−0.265*	−0.266*
	(3.620)	(3.868)	(0.135)	(0.146)
Bin-based	−2.516	−4.730	−0.045	−0.126
	(4.668)	(4.403)	(0.186)	(0.174)
R-Squared	0.234	0.305	0.252	0.324
P-value: Pop = Bin	0.250	0.447	0.207	0.337
Panel B: Heterogeneous treatment effects by overall rank				
Population-based	−8.961**	−8.682**	−0.313**	−0.305**
	(3.833)	(4.138)	(0.139)	(0.153)
Bin-based	−1.543	−4.016	0.018	−0.073
	(4.987)	(4.769)	(0.193)	(0.184)
Pop. × Top 15%	9.697*	7.507	0.301	0.224
	(5.540)	(5.316)	(0.241)	(0.230)
Bin × 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)	(4.729)	(0.223)	(0.209)
R-Squared	0.244	0.312	0.262	0.330
P-value: Pop = Bin at Bot. 85%	0.095	0.211	0.063	0.124
P-value: Pop = Bin at Top 15%	0.169	0.086	0.174	0.125
Panel C: Heterogeneous treatment effects by bin rank				
Population-based	−7.404**	−7.360*	−0.267*	−0.266*
	(3.727)	(3.982)	(0.140)	(0.151)
Bin-based	−2.234	−4.423	−0.029	−0.109
	(4.761)	(4.527)	(0.190)	(0.180)
Pop. × Subg. Top 15%	0.069	0.038	0.010	0.003
	(2.201)	(2.270)	(0.088)	(0.090)
Bin × Subg. Top 15%	−1.731	−1.876	−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
Control Mean	54.913	54.981	0.002	0.004
P-value: Pop = Bin at Bot. 85%	0.231	0.406	0.181	0.289
P-value: Pop = Bin 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 and age fixed effects, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

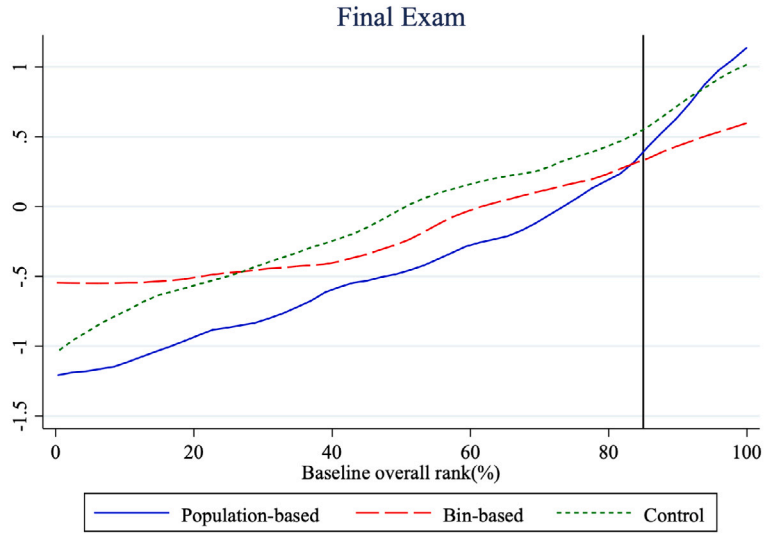
We find few impacts on observed and self-reported student effort. As shown in Column (1) of Table 5, there is a small marginally significant increase in the attendance rate among the Population-based scholarship group (Panel A), but we find no evidence for heterogeneity by baseline test score (Panel 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 5). As shown in Panel A, the point estimates for the Population-based 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 Population-based scholarship is −0.14 SDs, statistically significant at the 1 percent level. The Bin-based 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 SDs and is statistically significant at the 10 percent level.

In terms of heterogeneity by baseline score, Panel B of Table 5 shows that the negative impacts of the Population-based 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 −0.17 SDs and is statistically significant at the 1 percent level. The impact on the top 15 percent is 0.23 SDs higher than the bottom 85 percent (statistically significant at the 10 percent level). By contrast, we do not find similar evidence of heterogeneity for the Bin-based scholarship group. These findings suggest that, by signaling that high performance was valuable, the scholarships may have de-motivated students, particularly those in the Population-based scholarship program that were the least likely to receive the scholarship.

Columns (8) to (10) of Table 5 present impacts on students' perceptions of teacher and parental effort. We do not find evidence for changes in teacher effort as a result of either scholarship program. We do find that parents mentioned the scholarship program more often in the Population-based scholarship group, with effects concentrated among children with the highest baseline test scores. However, even

(a) Overall rank (%)



(b) Bin rank(%)

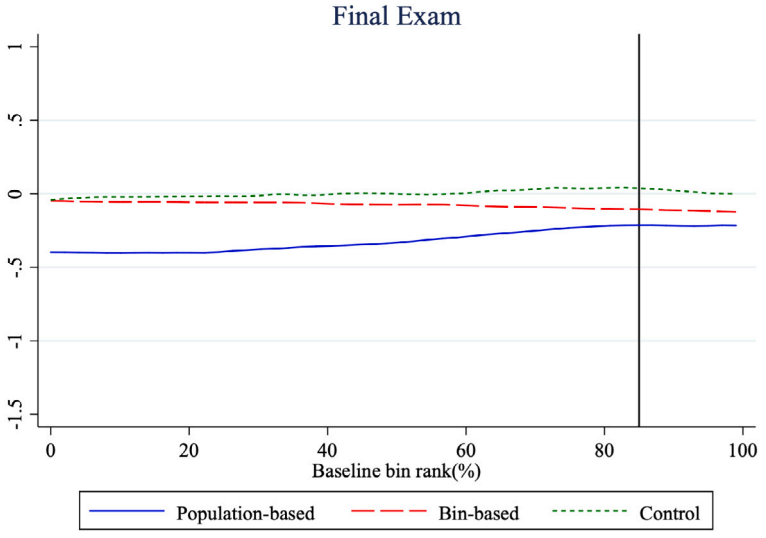


Fig. 5. Final exam scores by baseline rank.
Notes: This figure presents average final exam scores by baseline rank for each treatment group.

though parents of the Population-based 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.[21] A lack of capacity and resources may explain the null impacts of parental effort. However, the results in Column (10) suggest that parents were aware of the program and discussed it with their children. The small attendance impacts of the Population-based scholarship may therefore have been partially a result of parental encouragement to attend school.

4.4. Longer-term impacts

As discussed previously, the Population-based 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 2.2.1, 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 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.

Columns (3) and (4) of Table 6 present the longer-term results of the scholarship programs on test scores. As shown in Panel A, the negative effects of the Population-based scholarship program have faded substantially: the average longer-term impacts are much smaller in absolute value than the short-term impacts and are no longer statistically significant. We note, however, that these estimates are imprecise, with

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

Table 4
Heterogeneity by quintile of baseline test score.

	Sample grade: 5–8		
	Exam score		
	Population-control θ_k^p	Bin-control θ_k^b	Population-bin $\theta_k^p - \theta_k^b$
	(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, zone, and age fixed effects, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

confidence intervals admitting fairly large negative impacts. We also find smaller – and still not statistically significant – negative effects of the Bin-based scholarship program in the longer-term, although again the estimates lack precision to draw stronger conclusions.

Online Appendix 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 Population-based 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 Population-based 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 Population-based scholarship resulted in a statistically significant decrease in test scores, especially for those with lower scores at baseline. We also find negative impacts of the Population-based 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 Population-based 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.[22]

In contrast with the Population-based scholarship, we did not find negative impacts in the Bin-based scholarship group. This suggests that by providing a greater chance to achieve the incentive, the negative effects of the Population-based 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 SDs. This suggests that some caution is warranted in interpreting the results of the Bin-based 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 Population-based 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 Online Appendix Table A5, 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.

Third, the incentives could also have negatively affected the classroom environment by encouraging competition and discouraging collaboration. In our longer-term follow-up survey, we asked children questions about whether they recently helped or received help from friends. As shown in Online Appendix Table A6, we find no evidence that either scholarship group affected these measures.

Instead, key factors for negative impacts of the Population-based scholarship and the lack of impacts of the Bin-based scholarship may have been the pre-existing information of the students and the learning environment in the schools in our sample. These contextual differences may explain the contrasting results in Kremer et al. (2009), which also studies the impact of scholarship in the setting of rural schools in sub-Saharan Africa and upon which the design the Population-based scholarship was based upon. First, because students knew their ranking at baseline, the difficulty in achieving the Population-based scholarship may have been particularly salient, especially for students with the lowest baseline scores in our setting. This contrasts with Kremer et al. (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 et al. (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. 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 Population-based scholarship group, but not in the Bin-based scholarship group.

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

[22] To further examine this hypothesis, we conducted an analysis using causal forests (Wager and Athey, 2018) to predict heterogeneity in the effect of the Population-based scholarship on the non-cognitive skill index. We could then test if prediction corresponded to heterogeneity in outcomes. However, the causal forests did not reveal statistically significant heterogeneity in effects on non-cognitive skills (results not shown).

Table 5
Intermediate Outcomes.

	Sample: Grade 5–8									
	Student input		Non-cognitive skills					Teacher and parental response		
	Attendance	Study Hours	Motivation to study hard	Self esteem	Grit	Conscientiousness	Non-cognitive skill index	Teacher effort index	Parental effort index	Parents mentioned scholarship
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Average treatment effects										
Population-based	0.024*	−0.970	−0.071**	−0.030*	−0.034	−0.045	−0.136***	−0.044	−0.036	0.126*
	(0.013)	(1.033)	(0.035)	(0.017)	(0.023)	(0.032)	(0.051)	(0.102)	(0.084)	(0.063)
Bin-based	0.007	−1.536	−0.035	−0.028	−0.027	−0.026	−0.100*	−0.061	0.022	0.089
	(0.015)	(1.154)	(0.038)	(0.017)	(0.023)	(0.034)	(0.054)	(0.088)	(0.082)	(0.071)
R-Squared	0.192	0.076	0.022	0.050	0.049	0.080	0.117	0.085	0.044	0.039
P-value: Pop = Bin	0.207	0.541	0.238	0.920	0.698	0.526	0.491	0.850	0.249	0.574
Panel B: Heterogeneous treatment effects by overall rank										
Population-based	0.024*	−0.959	−0.089**	−0.035*	−0.040*	−0.059*	−0.174***	−0.047	−0.042	0.081
	(0.013)	(1.118)	(0.038)	(0.018)	(0.023)	(0.031)	(0.054)	(0.106)	(0.090)	(0.067)
Bin-based	0.009	−1.396	−0.049	−0.026	−0.011	−0.021	−0.089	−0.060	0.047	0.111
	(0.015)	(1.233)	(0.042)	(0.019)	(0.024)	(0.031)	(0.056)	(0.093)	(0.087)	(0.069)
Pop. × Top 15%	−0.008	0.080	0.115*	0.032	0.030	0.084	0.227*	0.018	0.030	0.277**
	(0.023)	(1.716)	(0.063)	(0.039)	(0.051)	(0.094)	(0.134)	(0.128)	(0.110)	(0.108)
Bin × Top 15%	−0.019	−1.024	0.066	−0.016	−0.098**	−0.034	−0.080	−0.020	−0.158	−0.058
	(0.027)	(2.049)	(0.066)	(0.034)	(0.040)	(0.092)	(0.121)	(0.124)	(0.111)	(0.121)
Top 15%	0.043**	1.521	−0.004	0.024	0.090***	0.026	0.086	0.065	0.132	−0.230***
	(0.016)	(1.521)	(0.050)	(0.030)	(0.029)	(0.083)	(0.101)	(0.098)	(0.092)	(0.086)
N	7046	5213	5726	5813	5813	5815	5821	5809	5749	5819
R-Squared	0.194	0.077	0.023	0.052	0.054	0.083	0.122	0.086	0.046	0.043
Control Mean	0.766	16.615	4.361	2.748	3.303	3.719	0.012	0.005	−0.013	3.231

Notes: Standard errors, clustered at the school-grade level, are in parentheses. Controls include baseline test score, grade, zone, and age fixed effects, ethnic group, household size, and a household asset index. Each index is constructed by taking the standardized average of the standardized components. Non-cognitive skill index is the aggregate of the motivation, self esteem, grit, conscientiousness measures. Teacher and parental effort indices are aggregates of seven and four measures, respectively. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

Table 6
Longer term test score impacts.

	Sample: Grade 5–6			
	1st Follow-up		2nd Follow-up	
	(1)	(2)	(3)	(4)
Panel A: Average treatment effects				
Population-based	−0.442**	−0.518**	−0.266	−0.116
	(0.208)	(0.248)	(0.191)	(0.235)
Bin-based	−0.210	−0.374	−0.188	−0.078
	(0.273)	(0.277)	(0.162)	(0.201)
R-Squared	0.232	0.318	0.103	0.218
P-value: Pop = Bin	0.341	0.465	0.711	0.852
Panel B: Heterogeneous treatment effects by overall rank				
Population-based	−0.474**	−0.547**	−0.328*	−0.183
	(0.226)	(0.267)	(0.188)	(0.237)
Bin-based	−0.131	−0.322	−0.145	−0.018
	(0.293)	(0.298)	(0.144)	(0.195)
Pop. × Top 15%	0.212	0.185	0.377	0.361
	(0.275)	(0.296)	(0.289)	(0.260)
Bin × Top 15%	−0.442	−0.277	−0.220	−0.294
	(0.353)	(0.326)	(0.336)	(0.286)
Top 15%	0.118	0.123	0.006	0.047
	(0.249)	(0.263)	(0.208)	(0.203)
Additional Controls	No	Yes	No	Yes
N	4040	3860	2476	2371
R-Squared	0.241	0.323	0.112	0.228
Control Mean	0.000	0.006	0.003	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 and age fixed effects, ethnic group, household size, and a household asset index. * denotes significance at 0.10; ** at 0.05; and *** at 0.01.

resources. We evaluate two incentive schemes: a Population-based scholarship program that provided scholarships for students whose test scores were within the top 15 percent and a novel Bin-based scholarship scheme that provided scholarships for the top students within smaller groups with similar baseline scores.

We find that the Population-based scholarship statistically 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 Bin-based 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 Population-based 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 et al., 2011; Hoff and Pandey, 2006).

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 et al. (2009), but future work is needed to more rigorously estimate the factors that contribute to the success (or failure) of such schemes.

Although the Bin-based scholarship scheme did not have the adverse distributional effects of the Population-based scholarship, its failure to produce positive impacts is also notable. Even if the design could be modified to work, this type of scheme faces a further challenge of strategic behavior on the part of students. If – unlike our Bin-based-scholarship – the incentive scheme was known *ex ante* or was to be repeated, students may attempt to game the incentive scheme by decreasing performance on the initial exam to achieve the incentive with less effort. Understanding whether students respond strategically to repeated incentives of this type would therefore be a useful area of future research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Will be provided upon Request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jdeveco.2022.102893.

References

Angrist, Joshua, Lang, Daniel, Oreopoulos, Philip, 2009. Incentives and services for college achievement: evidence from a randomized trial. Am. Econ. J.: Appl. Econ. 1 (1), 136–163.

Angrist, Joshua, Oreopoulos, Philip, Williams, Tyler, 2014. When opportunity knocks, who answers? New evidence on college achievement awards. J. Hum. Resour. 49 (3), 572–610.

Barlevy, Gadi, Neal, Derek, 2012. Pay for percentile. Am. Econ. Rev. 102 (5), 1805–1831.

Barrow, Lisa, Richburg-Hayes, Lashawn, Rouse, Cecilia Elena, Brock, Thomas, 2014. Paying for performance: the education impacts of a community college scholarship program for low-income adults. J. Labor Econ. 32 (3), 563–599.

Behrman, Jere R., Parker, Susan W., Todd, Petra E., Wolpin, Kenneth I., 2015. Aligning learning incentives of students and teachers: Results from a social experiment in Mexican High Schools. J. Political Econ. 123 (2), 325–364.

Bellés-Obrero, Cristina, 2020. Who is Learning? A Field Experiment Comparing Three Different Incentive Schemes in the Same Educational Setting. Working paper, University of Mannheim.

Bénabou, Roland, Tirole, Jean, 2006. Incentives and prosocial behavior. Am. Econ. Rev. 96 (5), 1652–1678.

Berry, James, 2015. Child control in education decisions an evaluation of targeted incentives to learn in India. J. Hum. Resour. 50 (4), 1051–1080.

Berry, James, Kim, Hyuncheol Bryant, Son, Hyuk Harry, 2021. Replication data for: When student incentives don't work: Evidence from a field experiment in Malawi. Harvard Dataverse UNF:6:Jdmc919H5J9CPJ3R3yql2Q==V1.

Bettinger, Eric P., 2011. Paying to learn: The effect of financial incentives on elementary school test scores. Rev. Econ. Stat. 94 (3), 686–698.

Blimpo, Moussa P., 2014. Team incentives for education in developing countries: A randomized field experiment in Benin. Am. Econ. J.: Appl. Econ. 6 (4), 90–109.

Cameron, Judy, Pierce, W. David, 1994. Reinforcement, reward, and intrinsic motivation: A meta-analysis. Rev. Educ. Res. 64 (3), 363–423.

De Paola, Maria, Scoppa, Vincenzo, Nisticó, Rosanna, 2012. Monetary incentives and student achievement in a depressed labor market: results from a randomized experiment. J. Hum. Cap. 6 (1), 56–85.

Deci, Edward L., Koestner, Richard, Ryan, Richard M., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol. Bull. 125 (6), 627–668.

Duckworth, Angela Lee, Quinn, Patrick D., 2009. Development and validation of the short grit scale (grit-s). J. Pers. Assess 91 (2), 166–174.

Fryer, Roland G., 2011. Financial incentives and student achievement: evidence from randomized trials. Q. J. Econ. 126 (4), 1755–1798.

Fryer, Roland G., 2017. The production of human capital in developed countries: Evidence from 196 randomized field experimentsa. In: Banerjee, Abhijit Vinayak, Duflo, Esther (Eds.), Handbook of Economic Field Experiments. In: Handbook of Economic Field Experiments, vol. 2, North-Holland, pp. 95–322.

Fryer, Roland G., Devi, Tanaya, Holden, Richard T., 2016. Vertical versus horizontal incentives in education: evidence from randomized trials. Working Paper, Harvard University.

Gilligan, Daniel O., Karachiwalla, Naureen, Kasirye, Ibrahim, Lucas, Adrienne M, Neal, Derek, 2022. Educator incentives and educational triage in rural primary schools. J. Hum. Resour. 57, 79–111.

Gneezy, Uri, Meier, Stephan, Rey-Biel, Pedro, 2011. When and why incentives (Don't) work to modify behavior. J. Econ. Perspect. 25 (4), 191–210.

Hirshleifer, Sarojini R., 2021. Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance. Working Paper, University of California Riverside.

Hoff, Karla, Pandey, Priyanka, 2006. Discrimination, social identity, and durable inequalities. Am. Econ. Rev. 96 (2), 206–211.

John, Oliver P., Srivastava, Sanjay, 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In: Handbook of personality: Theory and research, Vol. 2, no. 1999. pp. 102–138.

Kling, Jeffrey R., Liebman, Jeffrey B., Katz, Lawrence F., 2007. Experimental analysis of neighborhood effects. Econometrica 75 (1), 83–119.

Kremer, Michael, Miguel, Edward, Thornton, Rebecca, 2009. Incentives to learn. Rev. Econ. Stat. 91 (3), 437–456.

Lee, David S., 2009. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Rev. Econ. Stud. 76 (3), 1071–1102.

Leuven, Edwin, Oosterbeek, Hessel, van der Klaauw, Bas, 2010. The effect of financial rewards on student achievement: Evidence from a randomized experiment. J. Eur. Econ. Assoc. 8 (6), 1243–1265.

Levitt, Steven D., List, John A., Sadoff, Sally, 2016. The effect of performance-based incentives on educational achievement: evidence from a randomized experiment. NBER Working Paper 22107.

Li, Tao, Han, Li, Zhang, Linxiu, Rozelle, Scott, 2014. Encouraging classroom peer interactions: Evidence from Chinese migrant schools. J. Public Econ. 111, 29–45.

List, John A., Livingston, Jeffrey A., Neckermann, Susanne, 2018. Do financial incentives crowd out intrinsic motivation to perform on standardized tests? Econ. Educ. Rev. 66, 125–136.

Loyalka, Prashant Kumar, Sylvia, Sean, Liu, Chengfang, Chu, James, Shi, Yaojiang, 2019. Pay by design: Teacher performance pay design and the distribution of student achievement. J. Labor Econ. 37 (3), 621–662.

Mbiti, Isaac, Romero, Mauricio, Schipper, Youdi, 2019. Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania. NBER Working Paper 25903.

Rosenberg, Morris, 1965. Society and the adolescent self-image. Science 148 (3671), 804.

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

Visaria, Sujata, Dehejia, Rajeev, Chao, Melody M., Mukhopadhyay, Anirban, 2016. Unintended consequences of rewards for student attendance: Results from a field experiment in Indian classrooms. Econ. Educ. Rev. 54, 173–184.

Wager, Stefan, Athey, Susan, 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. 113 (523), 1228–1242.

World Bank, 2015. World development indicators 2015.

Table A.1
RCTs of Pay for Performance.

Reference	Structure of incentive	Population	Effects	Heterogeneity by baseline ability
Angrist et al. (2009)	Passing college matriculation exams.	Low-achieving high school students in Israel.	Increase in pass rates of 5-7pp, stat. sig. in some specifications.	Pooled boys and girls not reported. Largest effects with highest predicted performance.
Angrist et al. (2014)	Linear incentives for grades above 70.	Second-year students at a Canadian commuter college.	Increase in courses graded but no statistically significant increase in GPA.	Not reported.
Barrow et al. (2014)	GPA threshold.	Community college students in New Orleans.	Statistically significant impacts on GPA in first semester but not second.	Not reported.
Behrman et al. (2015)	Incentives for mathematics performance that varied by baseline level and level reached.	High school sudents in Mexico.	Individual incentives had effects of 0.2-0.3 sd in math.	Largest effects of individual incentives on those with highest baseline scores.
Bettinger (2011)	Thresholds for passing or advanced passing on five different achievement tests.	Students in grades 3-6 in an Ohio public school district.	Increase in math test scores of 0.15 sd but not other subjects.	Strongest effects at top and bottom quartiles of math test score distribution.
Blimpo (2014)	Two performance thresholds on comprehensive end-of-year exams.	10th grade students in Benin.	0.29 sd overall (individual target arm).	No evidence of heterogeneity by baseline score.
Fryer et al. (2016)	Houston: piece rate for mastering math objectives. Washington, DC: students could earn for multiple achievement and behavior metrics	6-8th graders in Washington, DC and 5th graders in Houston	In Houston, students mastered 1.1 SD more math objectives. In Washington, DC, 0.14 SD higher in math, 0.15 higher in reading	Houston: largest effects on high-ability students. Washington, DC: not reported.
Fryer (2011)	Dallas: piece rate per book read; New York City: linear incentives on interim assessments; Chicago: incentives for each letter grade in five different courses.	Second graders in Dallas, 4th and 7th graders in New York City, and 9th graders in Chicago	No statistically significant impacts on test scores	No evidence for heterogeneity by previous year's test scores.
Kremer et al. (2009)	School fees and a grant for girls in top 15 percent on district-wide exams in five subjects.	Girls in grade 6 in Kenyan primary schools.	Test scores increased by 0.19 sd.	No evidence for heterogeneity by previous year's test scores.
Leuven et al. (2010)	Financial incentives for passing all first-year requirements before the start of the next academic year.	First-year undergraduate students in economics and business at the University of Amsterdam.	No statistically significant impacts on achievement overall.	Positive stat. sig. effects for high ability students and negative effects for low ability students, categorized by high school math grades.
Levitt et al. (2016)	Monthly achievement standard of combined attendance, behavior, grades, and test scores.	High school freshmen in Chicago Heights, IL.	Modest statistically significant increases in meeting achievement standards.	Largest impacts on below-median students and students whose baseline achievement was near threshold.
List et al. (2018)	Combination of improved test scores, maintained grades, and avoiding unexcused absences.	Students in grades 3-8 in Chicago Heights, IL.	Improvement in incentivized tests by 0.3 to 0.37 standard deviations.	Largest improvements for lowest grade tercile.
Sharma (2010)	Linear incentives for semester and end of year exams across nine subjects.	Grade 8 students in Nepali Public Schools.	Increase in test scores of 0.03-0.09 sd, stat. sig. in some specifications.	Not reported; larger effects for students in the bottom quartiles of the distribution.
Hirshleifer (2021)	Piece rate incentives for unit tests in math.	Grade 4-6 students in Mumbai and Puna, India.	Increase in test scores of 0.24-.35 sd, statistically significant in some specifications.	No evidence for heterogeneity by baseline test score.
De Paola et al. (2012)	30 best-performing students in credits earned and grades obtained in each reward group (high or low).	First-year undergraduate students in an Italian university.	Increases in points (sum of GPAs) and credits earned.	Largest effects on those with highest predicted performance.
Bellés-Obrero (2020)	Three treatment arms: grade threshold, top percentile, and improvement.	Students in a microeconomics class at a public distance learning university in Spain.	No statistically significant impacts of any treatment arm on exam grades.	Some evidence of larger impacts for those with lower predicted grades.

Notes: This table lists studies of incentives for academic performance that are evaluated using randomized trials. Effect sizes are statistically significant at at least the 10 percent level unless otherwise noted.