Economics of Education Review 85 (2021) 102195

Contents lists available at ScienceDirect

Economics of Education Review

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

Understanding the response to financial and non-financial incentives in education: Field experimental evidence using high-stakes assessments

Simon Burgess [a], Robert Metcalfe [b], Sally Sadoff [c,*]

[a] University of Bristol, 2C1, The Priory Road Complex, Priory Road, Clifton, BS8 1TU, UK
[b] University of Southern California, 3620 South Vermont Ave. Kaprielian (KAP) Hall, 300, Los Angeles, CA 90089-0253, USA
[c] UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

ARTICLE INFO

JEL codes:
I20
I24
C93

Keywords:
Education
Effort
Incentives
High-stakes Assessments
Field experiment

ABSTRACT

We analyze the impact on high-stakes assessments of incentivizing students' effort in a field experiment with over 10,000 high school students. We contribute to the literature by using our rich data and machine learning techniques to explore treatment heterogeneity; by comparing financial and non-financial rewards in rewarding effort rather than grades; and by using high-stakes outcomes. We find little average impact of incentives in the overall population, but we identify a "right tail" of highly responsive students: in the upper half of the responsiveness distribution, test scores improve by 0.1-0.2 SD, about half the attainment gap between poor and non-poor students.

1. Introduction

Many countries struggle with persistently underperforming students and schools. Low educational achievement has a lasting impact on individual lives and represents lost output for the economy as a whole (Hanushek, 2009; Michelmore & Dynarski, 2016). Many governments are turning to new ideas in an attempt to deal with this problem, including the potential for incentives to increase student motivation and change behaviors in schools. As we detail below, the results of many of these studies have been disappointing for policy-makers, showing null or small effects on average. Increasingly, however, researchers are using richer data and stronger techniques to move beyond estimating the average effects of an intervention. Establishing the nature of treatment heterogeneity helps to illuminate the mechanisms behind the effectiveness (for some) of the intervention, and also allows policy to be targeted more effectively and cheaply.

In this paper, we report the results of a randomized experiment that includes over 10,000 students in 63 high schools spread across England, as shown in Fig. 1. We randomized schools to one of three treatment groups: Financial Incentives, Non-financial Incentives or Control. Both of the incentive treatment groups rewarded students on multiple dimensions of effort (attendance, behavior, classwork, and homework) measured over a five-week period (a half-term), with four such periods between September and April. The financial incentives offered students up to £80 per half-term (for a total of £320 over the year).[1] The non-financial incentives offered students the chance to qualify for a high-value event determined jointly by the school and the students, such as a trip to the national soccer stadium or an amusement park, costing approximately £40 per student per term (for a total of £80 over the year).

Our field experiment was implemented in high schools in England, in the final year of compulsory schooling when the pupils are 16 years old, leading up to the high stakes assessments called the General Certificate of Secondary Education (GCSE). GCSEs serve as the primary gatekeeper for students to stay in school and progress to college.[2] They are also key

* Corresponding author.
E-mail addresses: Simon.Burgess@bristol.ac.uk (S. Burgess), ssadoff@ucsd.edu (S. Sadoff).

[1] This is equivalent to about 520 USD using July 2014 exchange rates and equates to just over 4 h per week at the youth sub-minimum wage that these pupils could earn at the end of this school year. See https://www.gov.uk/government/news/government-approves-new-national-minimum-wage-rate-of-6-31, accessed 21/7/2014. In related work, Dearden et al. (2009) find that incentives of £1,300-£1,700 per year to continue past compulsory schooling increase enrolment rates by 4-7 percentage points (authors' calculations based on Dearden et al., 2009, Table 1). The incentives in that study are offered only after students complete year 11 and take the GCSEs.

[2] Appendix A provides details on the structure of the education system in England.

https://doi.org/10.1016/j.econedurev.2021.102195
Received 25 August 2020; Received in revised form 27 July 2021; Accepted 12 October 2021
Available online 9 November 2021

qualifications in the labor market for those not continuing in education. Our primary outcome of interest is the GCSE assessment scores on the subjects for which behavior was incentivized: English, math, and science.

We find little average impact in the full sample of either the financial or non-financial incentives on GCSE performance. This is perhaps not surprising given that we would expect our intervention incentives primarily to have an impact among those students who are not already motivated by the much larger inherent returns to education. In contrast, students who are motivated and putting forth effort at baseline will likely be less responsive to our incentives – because they have less room for improvement (Gneezy et al., 2019). Our analysis then focuses on establishing that the overall null effect is not common across all students, and on identifying a "right tail" of students whose performance can be significantly and substantially moved by the intervention incentives.

To do this, we take advantage of our large and diverse sample of students. As discussed in more detail in Section 3, we targeted schools serving students living in neighborhoods in the highest decile of eligibility for free school meals (FSM) is about 40%. This differs from experiments in the U.S. targeting low-income neighborhoods or low-performing schools, which tend to be more highly segregated.[3]

Adopting a simple subsample analysis first, we show that financial incentives have a significant impact on the performance of native students, but little impact among the children of immigrants, as measured by English as an Additional Language (EAL) status. We estimate that financial incentives increase the math and science GCSE scores of non-EAL students by around 0.13–0.14 standard deviations (SDs), with little impact on scores in English.

The danger of this subsample approach is researcher discretion and spurious findings: simple searches for responsive subgroups may pick up noise rather than a true treatment impact. And while corrections for multiple hypothesis testing are useful, they can be conservative in settings with many hypotheses. We therefore turn to newly developed applications of machine learning to examining heterogeneity of causal impacts. Part of the appeal of using machine learning techniques is precisely that they remove discretion and "help identify treatment heterogeneity in a principled way" (Davis & Heller, 2017, p. 546). We use

Fig. 1. Location of experimental schools across England.

neighborhood poverty. Within these schools, there is substantial diversity at the student level. As shown in Table 1, about half the students are white, a quarter have Asian ethnicity and a quarter are black; about half are categorized as having English as an additional language (EAL), a proxy for immigrant status; and the average poverty rate as measured by

[3] In Fryer (2011) for example, 88% of students are minorities (black or Hispanic) and 86% are eligible for free lunch, a proxy for poverty status (authors' calculations based on Fryer, 2011, Table 1).

Table 1
Baseline characteristics by treatment Group.

	Control	Financial Incentive	Non-Financial Incentive
Female	0.561	0.493	0.469
	(0.496)	(0.500)	(0.499)
Asian ethnicity	0.250	0.278	0.220
	(0.433)	(0.448)	(0.414)
Black ethnicity	0.248	0.256	0.259
	(0.432)	(0.437)	(0.438)
White ethnicity	0.422	0.397	0.430
	(0.494)	(0.489)	(0.495)
Free School Meals (FSM)	0.387	0.374	0.441
	(0.487)	(0.484)	(0.497)
English as an additional language (EAL)	0.478	0.535	0.491
	(0.500)	(0.499)	(0.500)
Special Education Needs (SEN)	0.021	0.017	0.021
	(0.142)	(0.128)	(0.143)
Baseline (KS2) English score	3.545	3.452	3.460
	(1.365)	(1.423)	(1.461)
Baseline (KS2) Math score	3.510	3.484	3.465
	(1.367)	(1.445)	(1.477)
Baseline (KS2) Science score	3.721	3.679	3.670
	(1.413)	(1.482)	(1.516)
School is in London	0.507	0.502	0.685
	(0.500)	(0.500)	(0.465)
School is an Academy	0.372	0.416	0.308
	(0.483)	(0.493)	(0.462)
Schools	33	15	15
Students	5553	2407	2689

Note: The table reports means and standard deviations for each variable. There are no differences between the control group and either treatment group significant at the 10/5/1 percent levels, using robust standard errors clustered on the school level.

both a well-established linear leave-one-out approach and the recently developed Causal Forest methods (Athey & Imbens, 2016b; Davis & Heller, 2017; Wager & Athey, 2018) to flexibly estimate treatment heterogeneity and test our findings.

Across both methods, we find evidence that there is a "right tail" of students—in this case, the upper half—who experience significant impacts on assessment scores. Using our preferred approach, we find that among students predicted to have above median treatment effects, the financial incentives improve math and science GCSE scores by 0.14–0.16 SDs. These effect sizes are similar to the more simply estimated effects for non-EAL students, which emerges from the machine learning analyses as a key predictor for high responsiveness. Critical for policy, the students predicted to be most responsive to both financial and non-financial incentives in math are those with lower performance at baseline. Our results suggest that incentives could help close achievement gaps among these students by about half.[4]

While it is politically and logistically difficult to target individual students within a school, we provide evidence to show that targeting the intervention on particular schools is feasible and effective. We are able to identify schools with relatively large shares of pupils predicted to have high responsiveness to financial incentives. If the intervention were simply run on these schools, we would expect school-average effect sizes of around 0.10 – 0.15 SDs (see Section 4.4).

The design of our study brings together several strands of the existing literature on incentives in education (see e.g., Fryer 2017, Sadoff, 2014 for review). The intervention we use incentivizes pupil inputs not outputs, that is, behaviors not grades. Incentivizing inputs may be more effective than incentivizing outputs if students lack information on how to improve their performance. However, if the incentivized inputs are not sufficiently related to the relevant output, then behavioral improvements may have little impact on performance. Our incentives are closest in design to prior work offering repeated short-term financial incentives based on measures of effort and engagement, such as attendance, behavior, grades, homework, classwork, low stakes test performance, and completing math or reading tasks (Fryer, 2017; Fryer & Holden, 2013; Hirshleifer, 2017; Levitt et al., 2016b). Our results showing little impact in the overall population are consistent with prior findings. Interestingly, our finding that financial incentives have the largest impact among non-EAL students aligns with Fryer (2011) in a very different context. He tests the impact of incentives for reading books among U.S. primary students and finds that while there is little overall impact in the population, there are large effects on test scores among English speaking students, which are netted out by negative impacts among English language learners.

Our non-financial incentives build on a growing body of research in behavioral economics demonstrating the power of non-financial rewards (Levitt et al., 2016b provide a discussion, and Damgaard & Nielsen, 2018, a survey). Recent work has examined the impact of non-financial incentives on low-stakes test taking effort and performance including both non-material rewards (e.g., certificates) and material rewards, such as trophies, cellular phone credits, or school supplies (Fryer, 2013; Jalava et al., 2015; Levitt et al., 2016b; Wagner & Reiner, 2015). These rewards are particularly attractive in educational contexts because they are low cost and more familiar to schools than cash rewards; educators may also believe that they do less to crowd out intrinsic motivation. In our context, non-financial incentives may be more cost-effective than financial incentives if students value the joint event above its face value. This could occur if, for example, it is difficult for students to coordinate such an event on their own, they value the public recognition, or they do not want to miss out on joining their peers.[5] To our knowledge, no previous study has compared financial and non-financial incentives of this type within a single experiment.[6]

Our study is also among the first to apply cross-validation methods to examining heterogeneous treatment effects. These methods have been used extensively to predict behavior – in education, for example, to predict teacher value added (Chetty et al., 2014a). In the context of randomized experiments, cross-validation methods have been used to predict the outcome of interest in order to examine treatment heterogeneity along that single dimension (Abadie et al., 2018; Angrist et al., 2016). However, cross-validation methods have only recently gained attention as a tool for predicting responsiveness to randomized interventions along multiple dimensions (Athey & Imbens, 2016a; Imai & Ratkovic, 2013; Wager & Athey, 2018; Handel & Kolstad, 2017; Davis & Heller, 2017).

Such tools can help researchers better understand the distributional effects of interventions, which can inform how best to target policies. Targeting may be particularly important in contexts like ours because there is often concern that extrinsic incentives can have negative effects on students who are already highly motivated, for example through crowd out of intrinsic motivation (e.g., Kohn, 1999) or 'choking' under increased pressure (e.g., Beilock, 2010). There is less concern about crowd out among those students who have little motivation at baseline – i.e., those who the incentives are designed to target. A better understanding of how to target educational interventions can also improve the

[4] Comparing students predicted to have High (above-median) vs. Low (below median) treatment effects, the predicted attainment gap in the math GCSE is 0.38-0.4 SD and the effect size of the intervention in math is 0.16-0.2 SD.

[5] There is also the potential that public rewards at the group level can generate negative peer effects (Austen-Smith and Fryer, 2005; Burstzyn and Jensen, 2015)

[6] Levitt et al. (2016b) compare financial and non-financial incentives in the context of one-time incentives for effort on a low-stakes diagnostic test. The non-financial incentives cost about 15–30% of the financial incentives, compared to our study in which the non-financial incentives cost about 25–33% of the financial incentive (including administrative costs).

efficiency of spending on social programs and help craft policies that meet the needs of individual students.

The remainder of the paper is organized as follows: the next section sets out the intervention design and details of implementation. Section 3 describes the data, randomization and estimation issues. Section 4 presents the results, and Section 5 offers wider conclusions.

2. Program details

2.1. Incentive design

The experiment took place during the 2012–2013 school year and included 10,649 year 11 students in 63 schools, which we randomized at the school level to one of the following treatment groups: Financial Incentives, Non-Financial Incentives, or Control. Students in the incentive treatments earned rewards every half-term (with four 5-week half-terms in the year) based on the following measures of behavior: attendance, conduct, homework, and classwork. The attendance standard required that a student have no unauthorized absences in the half-term. The conduct standard required that a student have no more than one instance of poor conduct resulting in a sanction. The homework and classwork standards required that a student complete the work on time and at a level consistent with the individual student's target grade in each of the three compulsory GCSE courses: math, English and science. A student's target grades in each subject, which were set before the experiment began, are determined by teachers and are a regular part of schooling in England.[7] Using target grades allows the performance measures to depend primarily on student effort relative to baseline ability – rather than setting a single threshold across all students.

Students earned rewards each term based on their performance on each standard. In the Financial Incentive treatment, students could earn up to £80 per period: £30 for meeting the class-work threshold, £30 for meeting the homework threshold, £10 for attendance and £10 for conduct. The structure for the Non-Financial treatment mirrored that of the Financial treatment, but with rewards in tickets rather than cash, with up to 8 tickets per half-term in the same 3/3/1/1 ratio described above. The structure is summarized below:

Behavior	Criterion	Financial Incentives	Non-Financial Incentives
Attendance	No absences	£10	1 ticket
Conduct	No more than one sanction	£10	1 ticket
Homework	Complete all work on time at or above individual target level	£30	3 tickets
Classwork	Complete all work on time at or above individual target level	£30	3 tickets

For the Non-Financial Incentive group, each student was able to participate in two events per year, in December after the second half-term and again in April after the fourth half-term.[8] In order to participate in the trip, a student needed to earn at least three-fourths of the tickets over the two half-terms (i.e., 12 out of the maximum of 16).[9] The events were chosen by the students and school administration collectively at the start of term, given a budget constraint.[10]

In both the Financial and Non-Financial Incentives, we used loss framing to increase the power and salience of the rewards. We told students that they began the half-term with the full reward (of £80 or 8 tickets) which they would (partially) lose if they missed the behavioral thresholds.[11] About a week after each half-term ended, we sent students a simple letter informing them whether they hit or missed the thresholds on the four behavior measures and their reward from the program (see Appendix B for an example letter, and for further details on the scheme). Students in the Financial Incentive treatment received payments by either cash or check through their school. Students in the Non-Financial Incentive treatment received notice of their (virtual) tickets through the feedback letters. Students in Control schools received no reward and were not sent feedback letters. Therefore, any treatment effects of the intervention measure the combined impact of incentives and feedback. The mean reward won was £211 (out of £320), with a very similar fraction of event 'tickets' won (20/32).

Overall, payments to schools totaled £729k. This included a compensation payment to each school of £2k, as well as £540k to Financial treatment schools and £119k to Non-Financial treatment schools. The monetary value of the Financial treatment was greater than that of the Non-Financial treatment. This was partly by design and partly because the schools and students made choices well within the budget.[12] In Appendix C and D we describe in detail how we provided information to schools and students, how schools reported outcomes to us, and how we provided feedback to the students.

2.2. Sampling frame and recruitment

The sampling frame was composed of regular state secondary schools in very disadvantaged areas, defined as the highest decile of neighborhood poverty as measured by the Income Deprivation Affecting Children Index (IDACI), yielding a total of 296 schools.[13] Inclusion was not conditional on school performance, so it did include some high-performing schools. Schools were removed from the sampling frame as follows: if they were in Special Measures (intense intervention to turn the school's performance around), if they were scheduled to close, or if they were a combined primary-secondary school.[14] The remaining sample included 279 schools covering 60 Local Authorities (out of 150

[7] Target grades are set by teachers for each pupil in each subject based on their interactions with the pupils in class and on informal class tests. This individual student level target is set to be challenging yet attainable for each student (although the effectiveness of such targets on achievement has not been formally tested).

[8] In year 11, teaching finishes in April to allow students time for personal study for the key GCSE assessments.

[9] Neither we nor the schools had the capacity to organize four substantial events in the year. Tickets from the first and second half-terms were pooled to determine qualifying for the first event. Tickets from the third and fourth half-terms were pooled to determine qualifying for the second event. Because there are 8 tickets available to earn per half-term and students needed earn at least 12 tickets to qualify for the event, students needed to earn at least 4 tickets each half-term to qualify.

[10] For example, in one school pupil representatives in year 11 sought suggestions and views from their cohort; in another, the school's Student Council worked with the project liaison teacher in the school to make the decision. Examples of events chosen include tickets to the School Prom and trips to Wembley (home of the England football team and a major venue of the 2012 London Olympics), the Houses of Parliament, large theme parks, and Winter Wonderland in Hyde Park (with each pupil having the trip, including an ice skating session, paid for as well as receiving £10 cash to spend in the park).

[11] We were unable to persuade any banks to create escrow accounts that would allow us to endow pupils with upfront rewards. We instead used language to frame the incentive as a loss – e.g., "this money is yours to lose or to keep", "your place is booked on the coach – don't miss the trip".

[12] The cost per student is harder to control in the event treatment, partly because of significant fixed costs (for example, hiring a bus for transportation), and partly because the nature of the event was chosen by the pupils and school. Schools worked within the following budget: £1,000 per term for the first 30 students who qualified, and £25 per additional qualifying student, up to a maximum total amount of £6,000 per event.

[13] See http://webarchive.nationalarchives.gov.uk/20120919132719/http://www.communities.gov.uk/communities/research/indicesdeprivation/deprivation10/

[14] We also omitted a single school, Mossbourne. This is the most famous state school in the country and many leading politicians mention it and visit it. The Headteacher who brought it to levels of very high performance has now become the Chief Executive of the Office for Standards in Education. The school attracts a huge amount of media attention, and undoubtedly, research interest.

in England). Recruitment halted for budgetary reasons after 84 schools signed up. After an initial training event explaining the details of the intervention, some schools dropped out, leaving 63 schools in the randomization. All year 11 students in a school were included in the study unless their parents signed a reverse consent form stating that they did *not* want their child to participate. Only 17 parents (<0.2%) opted their child out of the study.

The recruited schools and students are broadly representative of the sampling frame as shown in Appendix Table 1, and located across the country as shown in Fig. 1. The schools in the experiment are more likely to be in London, more likely to have a new principal and have students of slightly lower baseline ability. In Appendix C we describe in detail our school recruitment procedures, how we obtained consent, and how we explained the intervention to schools and students.

3. Data and research design

3.1. Data

Our primary outcome of interest is performance on the high-stakes General Certificate of Secondary Education (GCSE) qualifications, the compulsory set of examinations in England for those who are 16 years old. GCSEs are typically two-year courses taken in the final two years of compulsory schooling (year 10 and year 11) when students are 15 to 16 years old. Students take courses in a number of subject areas with virtually all students required to take GCSEs in English, math and science. Students must generally achieve a good pass (a grade of C or higher) in at least five subjects (including English and math) in order to progress to University. Good GCSE performance is also a common condition for employment.

A student's GCSE score depends primarily on a standardized national exam taken at the end of the year. GCSE exams are nationally set and remotely marked and have very high measurement fidelity. In some subjects, a portion of the score depends on performance during the two years of coursework, referred to as the "controlled assessment". Our intervention only takes place during the second year of the GCSE, so some marks will already have been banked for the final grade in the first year before treatment. Our data provide the overall grade for the course, not the 2013 exam mark separately.[15] Therefore, our treatment effect estimates may be an underestimate of the impact of incentives on GCSE performance. This is particularly relevant in English, in which the controlled assessment makes up 60 percent of the GCSE score, and to a lesser extent in Science, which gives 25 percent weight to the controlled assessment. The controlled assessment does not count towards the math score (Ofqual, 2013).

We focus on GCSE scores in the core subjects: math, English (language), and science.[16] We also examine overall performance: total capped points score (each student's best 8 scores) and whether they achieved the most prominent national benchmark of at least 5 good passes (grade C or higher).[17] In the cohorts in our data, students receive one of the following grades on the GCSE, with A* being the highest: A*, A–G, or U where U (ungraded/unclassified) signifies that a student achieved nothing worthy of credit. As discussed below, we convert the letter grades to numbers as follows: $A^* = 8, A = 7, B = 6, C = 5, D = 4, E = 3, F = 2, G = 1, U = 0$. We then standardized the GCSE scores using the national cohort (by year) to have mean 0 and standard deviation 1.

Our secondary outcomes of interest are the impact of treatment on the behaviors we incentivized: attendance, conduct, classwork and homework. The school administration reported the attendance and conduct measures. Classroom teachers reported students' classwork and homework measures in the relevant subject. The teacher-reported measures are potentially biased if teachers in treated schools report inflated performance so that students can receive the incentives. We therefore use these measures largely to examine potential mechanisms for the heterogeneity within schools of the impact of treatment on GCSE scores.

Our two primary sources of administrative baseline and outcome data are Edubase and the National Pupil Database (NPD).[18] We took the following data from the school level dataset, Edubase: school's location, number of students, expenditure per pupil, date of hiring for the Headteacher, plus indicators for whether it is a single sex school, whether it is an Academy (similar to charter schools in the U.S.), and whether it also teaches pupils beyond GCSE. The NPD provides student-level data on demographics and full exam histories for all students in England. The demographics include: gender (female, male), ethnicity (detailed categories which we aggregate to Asian, Black, or White[19]), English as an Additional Language (EAL) status, whether the student has a statement of special educational needs (SEN), birth month, and eligibility for free school meals (FSM), which is a proxy for low income status. The exam performance data include both the GCSE scores discussed above and scores from the Keystage 2 (KS2) tests taken at the end of primary school (year 6, age 11) in math, English and science, which we use as baseline ability measures. Both the GSCE and the KS2 are nationally set and remotely marked. Finally, we administered a short survey to schools before the randomization took place asking whether they were implementing their own incentive scheme at baseline. We also administered a survey to students, but due to low response rates (below 10%), we do not report the results.

3.2. Randomization

Randomization took place at the school level. School level randomization minimizes spillovers between treatment groups, allows us to measure the impact of treatment inclusive of peer effects, and is particularly important for the Non-Financial treatment, which offers a school-based group event. In addition, schools strongly prefer school level randomization because all their students receive (or do not receive) the same incentive. As we discuss further in Section 4, this is important for understanding how we might target incentives effectively given policy constraints. The drawback of school level randomization (relative to individual level randomization) is the substantial loss of power that this entails. While we work with over 10,000 students, the true variation in treatment is only across 63 schools. Given the typical cohort size of 180 and realised ICC values in our sample of 0.099 – 0.171, this implies a

[15] Our incentives for homework and classwork were not directly tied to GCSE coursework. However, performance in class could affect a small portion of a student's GCSE course grade. As discussed below, we do not find evidence that this is a significant driver of the treatment impacts on GCSE performance. It is also possible (though not common) to take the GCSE math course exams a year early. Our data do not indicate the date of the exam.

[16] Students can take multiple GSCEs to count towards science, including Physics, Chemistry, Biology and a general Core Science exam. We use a composite measure capturing the highest point score achieved (GCSE equivalencies) in these exams (See Appendix D for details).

[17] In our sample, the mean number of full GCSEs taken is 6.6.

[18] For school data, see http://www.education.gov.uk/edubase/home.xhtml accessed 22 July 2014. For pupil data See https://www.gov.uk/national-pupil-database-apply-for-a-data-extract accessed 22 July 2014.

[19] The ethnicity categories are as follows: 'Asian' includes pupils with Bangladeshi, Indian, Pakistani, Chinese, Other Asian, and Mixed White and Asian ethnicities; 'Black' includes Black African, Black Caribbean, Other Black heritage, Mixed White and Black African, and Mixed White and Black Caribbean ethnicities; and 'White' includes White British, White Irish, White Other, and White Irish Traveller ethnicities. The very few pupils who fit into none of these groups ("Refused", "Other ethnicities" or "Other mixed ethnicities") are in the omitted category with 'White' in the regressions.

design effect of 0.19 – 0.32.[20]

Budgetary constraints allowed us to assign 15 schools to the Financial Incentive, 15 schools to the Non-Financial Incentive and the remaining 33 schools to Control. We use a matched-triplets design, which allows us to conduct analyses correcting for non-compliance and attrition (discussed below). We first generated triplets of schools matched on the presence of a pre-existing reward scheme in the school and on which broad ethnic group was the majority group in the school: Asian, black, or white as defined above.[21] In the first 15 randomly chosen triplets, we assigned the first randomly chosen school to Financial Incentives, the second randomly school to Non-Financial Incentives, and the third randomly chosen school to Control. We assigned all schools in the remaining triplets to Control (full details of the randomization procedure are given in Appendix E).

While we have a large sample of students, we randomized across only 63 schools and simulation evidence presented in Bruhn & McKenzie (2009) suggest that for small samples, matched-pairs, re-randomization (the method employed here), and stratification all perform better than a pure random draw. Following the recommendation of Bruhn & McKenzie (2009), we have estimated our treatment effects including all individual student baseline characteristics used to check balance. We used re-randomization to ensure balance across a rich set of characteristics, including school composition, type and location, and recent past measures of school performance including raw output and value-added, levels and trends.[22] We re-randomized until all p-values from binary comparisons of the control and treatment groups were above the chosen significance level of 10%, with standard errors clustered at the school level. Table 1 presents summary statistics by treatment group for pre-treatment characteristics used in the primary analysis: gender, race/ethnicity, eligibility for free school meals (FSM), English as an additional language (EAL) status, special education needs (SEN) status, baseline test scores, whether school is in London and whether the school is an Academy.[23] We also report p-values for binary tests of difference between treatment groups and Control with standard errors clustered at the school level; there are no statistically significant differences. The results below include controls for the full set of covariates as recommended by Bruhn & McKenzie (2009).

3.3. Estimation

We estimate three models in the results section. The primary analysis uses all the randomized schools and focuses on test scores in math, English and science. We use data from all the schools in 2012/13, the year when our treated cohort was in year 11. In addition, through the census administrative data in the National Pupil Database (NPD), we have all the same characteristics – demographics, prior attainment and GCSE test scores – for the prior cohorts of year 11 students in all 63 schools.

The first model estimates the effect of treatment using only our experimental cohort. We estimate the subject *(j)*-specific intent-to-treat (ITT) effects of the Financial Incentive, π_{Fj}, and the Non-Financial Incentive, π_{Nj}, using the following model at the student level:

$$g_{ijs} = \alpha + \pi_{Fj}F_s + \pi_{Nj}N_s + \beta X_{ij} + \gamma Z_s + \varepsilon_{ijs} \quad (1)$$

where g_{ijs} is the score of student i in subject j in school s; F_S is an indicator variable for the Financial incentive in school s (the level of randomization); N_S is an indicator variable for the Non-Financial Incentive in school s; X_{ij} contains the characteristics of student i including subject-specific prior attainment, j; Z_s contains the characteristics of school s; and ε_{ijs} is noise.

The second model uses a difference-in-difference strategy by including the prior year cohort. This allows us to control for school fixed effects using the following model at the student level:

$$g_{ijsc} = \alpha + \pi_{Fj}F_{sc} + \pi_{Nj}N_{sc} + \beta X_{ij} + \mu_{js} + \delta_{cj} + \varepsilon_{ijsc} \quad (2)$$

where g_{ijsc} is the score of student i in subject j in school s in cohort c; F_{SC} is an indicator variable for the Financial Incentive treatment in school s in cohort c; N_{SC} is an indicator variable for the Non-Financial Incentive treatment in school s in cohort c (F_{SC} and N_{SC} are zero for all schools in prior cohorts and takes on the assigned status for the trial cohort); μ_{js} is a subject-specific school fixed effect; δ_{cj} is a subject-specific common cohort effect and ε_{ijsc} is noise.

Our third model estimates treatment effects for students with High predicted treatment effects and Low predicted treatment effects (we discuss how we generate the predictions in the next section). We use an interaction approach and separately estimate effects for the Financial Incentive and Non-Financial Incentive treatments (because we predict both a Financial Incentive treatment effect and a Non-Financial Incentive treatment effect for each subject for each student). For the Financial Incentives we use the following model to estimate treatment effects at the individual level for students with High predicted effects π_{FHj} and for students with Low predicted effects π_{FLj}:

$$g_{ijsc} = \alpha + \pi_{FHj}F_{sc}H_{Fij} + \pi_{FLj}F_{sc}L_{Fij} + \rho H_{Fij} + \beta X_{ij} + \mu_{js} + \delta_{cj} + \varepsilon_{ijsc} \quad (3)$$

where H_{Fij} is an indicator variable for High predicted treatment effects for Financial Incentives for the student i in subject j; and L_{Fij} is an indicator variable for Low predicted treatment effects for Financial Incentives for the student i in subject j. For Non-Financial Incentives, we estimate treatment effects at the individual level for students with High predicted effects π_{NHj} and for students with Low predicted effects π_{NLj} using the same model as in (3), except that we replace H_{Fij} with H_{Nij}, an indicator variable for High predicted treatment effects for non-financial incentives for student i in subject j; and we replace L_{Fij} with L_{Nij}, an indicator variable for Low predicted treatment effects for Non-Financial Incentives for student i in subject j.

We normalize the GCSE scores year by year over the whole national cohort. Estimated effects are therefore interpretable as units of student

[20] The realised ICC was 0.099 in math, 0.119 in English, 0.171 in Science 1, 0.097 in total capped GCSE points score and 0.067 in achieving the GCSE benchmark of at least 5 passes (grade C or higher).

[21] The small number of schools limited the number of covariates we could block the matched pairs on. We categorized pre-existing reward schemes in the school based on responses to the school survey, pooling schools that did not have reward schemes with schools that were non-responsive. See Wilson et al. (2011) for an analysis of the attainment of different ethnic groups.

[22] We randomized using the characteristics of the year 10 cohort in the year prior to the start of the experiment – i.e., the rising year 11 cohort. The baseline characteristics of the realized year 11 cohort are used in the analysis.

[23] For space, in Table 1 we do not include school level characteristics that are controlled for by the inclusion of fixed effects and do not enter our primary analysis of heterogeneity. The full set of pre-treatment characteristics used in the randomization are reported in Appendix Table 2. The full list includes: the proportion of pupils eligible for Free School Meals; the proportion of female students; the proportion of pupils of white ethnicity; the proportion of pupils of black ethnicity; the proportion of pupils of Asian ethnicity; the proportion of pupils with Special Educational Needs; the proportion of pupils with English as an Additional Language; whether the largest ethnic group in school is black; whether the largest ethnic group in school is Asian; whether the largest ethnic group in school is white; a measure of pupil neighbourhood poverty; value added (best 8 results) for low attainers; school average GCSE math score; school average GCSE English score; school average GCSE science score; school average capped GCSE score trend 2009-2011; cohort average prior attainment, math; cohort average prior attainment, English; cohort average prior attainment, science; cohort average prior attainment, average; the proportion of pupils achieving 5 A*-C GCSEs; the average capped GCSE points; whether the school is in London; single sex school; Academy; whether school has a sixth form; total number of pupils in school; school cohort size; total expenditure per pupil; Headteacher hired either September 2010 or September 2011; school has own incentive scheme: yes versus missing or no; school has own incentive scheme: no versus missing or yes.

level standard deviations (SD). Normalizing using the full population allows us to estimate the impact of our intervention on GCSE scores in the national distribution (see Kraft, 2020 for discussion). One student level SD in math is 1.8 grade points; that is, almost the two-point grade difference between an A and a C. Because the intervention is delivered at the school level, we cluster standard errors in all models at that level.

3.4. Attrition and compliance

We have very low attrition rates for our main outcome, GCSE scores, because these exams are compulsory and the data are available through a census national database. We have GCSE scores for 98.3% of our sample in math, 97.5% in English and 100% in science (all took some form of science, as explained in Appendix D), observing no difference between treatment and control schools.[24]

We have higher attrition rates for our secondary outcomes, the behavioral measures that we incentivize, because participating schools were responsible for collecting these data. All schools in the Financial and Non-Financial Incentive treatments provided complete data on the behavioral measures. However, of the 33 control schools, only 18 provided full behavioral data for the entire year. As shown in Appendix Table 2, the non-attriters are balanced on baseline characteristics across the treatment and control groups. We analyze the characteristics of those leaving in Appendix Table 3 and find no evidence of differential attrition. While these variables are useful for understanding the process by which the incentives worked, they are not required for our primary outcome.

Compliance in this context means schools acting against the intervention protocol. Complying treatment schools provided incentives for all of their pupils to respond to. The strength of pupils' responses is part of the heterogeneous effectiveness we estimate, it is not a measure of compliance; for example, pupils missing out on some of their reward does not indicate non-compliance, but perhaps simply that the cost of effort was too high.

One potential problem is non-compliance by control schools. All schools had to be told what the schemes entailed as part of the recruitment process, so those later randomized into control would know what was happening in the other schools, and could try to do the same. There are two counter-arguments to this. First, we informed schools as late as we could about their status, just before the start of the term. While we helped treatment schools to prepare and we had procedures ready, this would not have been the case for control schools wishing to imitate the incentives. Secondly, the Financial treatment is quite costly. Although we believe that schools could afford this on a continuing basis, it is far too large an amount of money for a school to have to find from an already committed budget at that point in the year. We were in contact with the control schools throughout the year and there was no indication that they were implementing any version of the treatment.

Finally, two treatment schools did not fully comply with the treatment protocol. One school in the Financial Incentive treatment stopped distributing the feedback letters after the first half-term. One school in the Non-Financial Incentive treatment did not explain the incentive to students and allowed all students to attend the school-wide events. We include all schools in our Intention-to-Treat Analysis, regardless of compliance.[25]

[24] Results available upon request.

[25] Estimates that are limited to intact matched triplets – i.e., triplets in which treatment schools complied with the protocol and control schools provided behavioral data throughout the year – are similar to the results for the full sample discussed in the next section, but are underpowered (available upon request).

4. Results

We first present the results for our primary question, the impact of behavior incentives on high-stakes test score outcomes. We then examine whether we can successfully identify students who differentially benefit from treatment. We then discuss potential mechanisms for the differential effects by examining the impact on the incentivized behaviors themselves. Finally, to examine its feasibility for policy, we examine whether we can identify schools that should be targeted for intervention.

4.1. The impact of financial and non-financial incentives on high stakes test score outcomes

Table 2 reports estimated treatment effects on standardized GCSE grades in math, English, and science. For each subject, we first report the treatment effect using only the experimental cohort (odd-numbered columns) and then add the prior year cohort in order to estimate the difference-in-difference model that allows us to include school fixed effects (even-numbered columns).[26] Standard errors are clustered at the school level in this table and throughout. All regressions include a full set of baseline student characteristics: gender, race/ethnicity, English as an additional language (EAL) status, Free School Meal (FSM) status, any stated notice of special educational needs (SEN), month of birth, and the baseline score in the relevant subject.[27]

We find no statistically significant impact of either intervention.[28] Focusing on the analyses with school fixed effects, the estimated effects of both incentives are generally small and positive, but in no case do they approach statistical significance. The point estimates for financial incentives of 0.053 standard deviations (0.047 standard error) in math and 0.082 standard deviations (0.067) in Science are higher than the meta-coefficients of approximately 0.02 standard deviations for incentive interventions estimated by Fryer (2017). And, given the size of the standard errors, we are not able to rule out economically meaningful impacts. There is similarly suggestive evidence of modest overall effects on science performance in response to non-financial incentives. There is little estimated impact in English, though standard errors are also large.

4.2. Heterogeneous treatment effects

We now turn to examining whether there is a subgroup of students who demonstrate economically and statistically significant treatment effects. In Table 3, we split the sample by EAL status. This split is motivated in part by our prior work, which finds that EAL classification – which is a proxy for immigrant status – has a high positive correlation with motivation and performance (Burgess et al., 2009; Burgess & Heller-Sahlgren, 2018).[29] To increase precision, all regressions include a full set of baseline covariates, as well as school fixed effects using the difference-in-difference approach discussed above. We find a significant impact of the financial incentives among non-EAL students, that is, native students: an effect size of 0.136 standard deviations in math

[26] Regressions including 2, 3 and 4 years of prior data yield similar results (available upon request).

[27] As discussed above, we use Keystage 2 scores taken in grade 6 as the baseline score. The coefficients in specifications excluding school and student characteristics follow a similar pattern and magnitude to Table 2 and are available upon request.

[28] These results are consistent with the findings of the funder's independent evaluation, which used the intervention year of data only, and focused on the overall effect. See the report here: https://educationendowmentfoundation.org.uk/uploads/pdf/Final_Copy_EEF_Evaluation_Report_-_Pupil_Incentives.pdf

[29] EAL is highly correlated with ethnicity. Approximately 92% of Asian students are classified as EAL students while only 21% of white students are also EAL. Interestingly, EAL status among black students mirrors the overall sample with about 47% classified as EAL.

Table 2
Effects of incentives on high stakes exam scores.

	Math	English	Science			
	1	2	3	4	5	6
Financial Incentive	-0.006 (0.074)	0.053 (0.047)	-0.072 (0.099)	-0.001 (0.052)	-0.049 (0.113)	0.082 (0.067)
Non-Financial Incentive	0.091 (0.064)	0.014 (0.034)	-0.012 (0.078)	0.029 (0.060)	0.059 (0.106)	0.054 (0.053)
Pupil Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	No	Yes	No	Yes	No	Yes
p-value: Financial = Non-Financial	0.244	0.396	0.594	0.592	0.332	0.630
Observations	9827	20,058	9750	19,791	10,238	20,713
Schools	63	63	63	63	63	63

Note: OLS estimates. Dependent variable is normalized exam score. Standard errors clustered by school in parentheses. The sample includes students in the experimental cohort and in the prior cohort in the prior year. All estimates include school and year fixed effects and the following baseline characteristics: gender, race/ethnicity, Free School Meal status, English as an Additional Language status, subject-specific baseline test score, and month of birth. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3
Effects of incentives by EAL status.

	Math	English	Science			
	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive
Financial Incentive * EAL = 0	0.136** (0.058)		0.029 (0.053)		0.133* (0.071)	
Financial Incentive * EAL = 1	-0.018 (0.054)		-0.028 (0.057)		0.038 (0.075)	
Non-Financial Incentive * EAL = 0		0.011 (0.049)		0.071 (0.063)		0.073 (0.058)
Non-Financial Incentive * EAL = 1		0.017 (0.049)		-0.016 (0.068)		0.033 (0.058)
Pupil Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
p-value: EAL=0 = EAL = 1	0.040	0.919	0.161	0.138	0.098	0.382
Observations	14,947	15,551	14,773	15,359	15,386	16,069
Schools	63	63	63	63	63	63

Note: OLS estimates. Dependent variable is normalized exam score. Standard errors clustered by school in parentheses. The sample includes students in the experimental cohort and in the prior cohort in the prior year. All estimates include school and year fixed effects and the following baseline characteristics: gender, race/ethnicity, Free School Meal status, English as an Additional Language status, subject-specific baseline test score, and month of birth. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

(standard error 0.058) and 0.133 standard deviations in Science (0.058). For EAL students, the children of immigrants, we find nothing of any size or significance. Again, we find no significant effects of either incentive for English for either group. Since the estimation sample splits about 50:50 on EAL, the overall results in Table 2 (effect size around 0.06) are a simple average of zero effect for EAL students and an 0.13 effect size for non-EAL students. The estimated impact on the math scores of non-EAL students remains significant at the 10% level after correcting for multiple hypothesis testing across demographic subgroups (EAL, gender and race/ethnicity) using the method described in Anderson (2008); the estimated impact in science is not robust to MHT corrections.[30]

A concern about the subsample analysis presented above is that the differential effects we find by EAL status may be spurious, simply picking up noise. We therefore examine whether our finding holds up to alternative methods for examining heterogeneous treatment effects. In particular, we use cross-validation methods in order to use a richer set of observable characteristics to predict which students will experience the largest treatment effects. The literature on machine learning in economics is still developing and so we use different techniques including a well-tried approach as well as a recently-developed approach. First, we use a linear leave-one-out analysis, which we argue is a straightforward tool for policymakers interested in using estimates of heterogeneity to target interventions. In particular, they can be used to predict likely effectiveness for targeting in very large scale (national) datasets. We supplement this analysis by adopting the recent approach of Athey and Imbens (2016b) and Wager and Athey (2018) providing what Davis & Heller (2017) describe as a "principled way" of identifying treatment heterogeneity using causal forests (CF). We describe our approach in more detail in Appendix F[31]; we largely follow the approach set out very clearly in Davis & Heller (2017); that paper and Davis and Heller (forthcoming) provide more general details and implementations of the approach.

We implement the leave-one-out approach as follows. We first run the difference-in-difference specification in Table 2 with the addition of interaction terms for each incentive treatment in each subject with indicator variables for each of the following baseline characteristics: female, Asian, black, English as an Additional Language (EAL), Free School Meals (FSM), school is an Academy, and school is in London.[32] We use this specification to generate estimates of predicted treatment

[30] We report effects by gender and race/ethnicity in Appendix Tables 4a and 4b respectively. The gender subsamples do not include the full set of 63 schools because 8 schools are single sex girls' schools and 3 schools are single sex boys' schools. Similarly, the subsamples by race/ethnicity do not include the full set of schools because 3 schools contain no Asian students, and 1 school contains no black students. The Financial Incentive has the largest consistently positive impact on black students, significant at the $p < 0.05$ level in math. In the Non-Financial Incentive, the largest effects are among girls and white students, the latter significant in science at the $p < 0.05$ level.

[31] We also implement a Least Absolute Shrinkage and Selection Operator (LASSO) approach that also yields similar results (available upon request). For further discussion of approaches to treatment effect heterogeneity, see Athey and Imbens (2016a).

[32] We do not include an interaction term for Special Education Needs due to small sample sizes (2% of students). As discussed above, Academies are similar to charter schools in the U.S.

effects at the individual level using the leave-one-out approach. That is, for each student we separately estimate the regressions described above, excluding the focal individual, and use those estimated coefficients to predict the treatment effect for that individual (as detailed in Abadie et al. (2018) including the focal student introduces bias into the estimates due to the mechanical correlation between the student's predicted and actual outcomes). We calculate a student's predicted treatment effect by applying the coefficients for the incentive and each incentive-characteristic interaction effect to the individual's own characteristics. Note that leave-one-out estimation is a k-fold cross-validation method where k = n (i.e., the number of observations).

We calculate a predicted treatment effect for each subject and incentive type. We therefore calculate six predicted treatment effects for each student in our sample: the predicted treatment effects of Financial Incentives on math test scores, English test scores, and science test scores; and the predicted treatment effects of Non-Financial Incentives on math test scores, English test scores, and science test scores. To show a 'typical' set of results, Appendix Table 5 presents the coefficient estimates when we include the full sample (i.e., without leaving-one-out).

To summarize the quantitative and statistical significance of the results we again estimate summary regressions based on these predicted treatment effects, shown in Table 4. We split the sample by above and below median predicted treatment effects, separately for Financial and Non-Financial Incentives and by subject. Because we generated our predictions using a leave-one-out approach, this is essentially an out-of-sample test of our predicted effects.

For math GCSE scores, the interaction of treatment with Predicted High is significant for both the Financial Incentive and Non-Financial Incentive treatment ranging from 0.12–14 standard deviations. The point estimates for the un-interacted Predicted High term are negative, suggesting that students who are most responsive to incentives are those with lower scores at baseline. Taken together, the estimated impact of treatment among Predicted High students is approximately 0.16–0.2 standard deviations, significant at the $p < 0.05$ level. For comparison, the attainment gap between poor and non-poor students in our sample is 0.32 standard deviations in math, and 0.34 in Science. Among Predicted Low students, the estimated treatment effects are small and not statistically significant. A similar pattern holds for the estimated impact of the financial incentive on science with an estimated treatment effect among Predicted High students of 0.2 standard deviations significant at the $p < 0.1$ level. We do not find differential effects of the non-financial incentive among Predicted High students in science. And we find no impact of either treatment in English. We note that these effect sizes are similar to the estimated effects among non-EAL students reported above.

The results of the leave-one-out approach are similar to our Causal Forest (CF) approach. The results of the CF analysis are shown in Appendix Table 6. The estimated effects for high effectiveness students are generally of economic significance, between 0.15–0.4 SD, and are differentiated from the estimated effects for "Predicted Low" students, which are generally negative (the exception to this pattern is the non-financial incentive in math, which has little estimated impact on either "Predicted High" or "Predicted Low" students). In English and science, the high effectiveness students also have lower performance at baseline.

In Table 5 we analyze outcomes on two measures of overall performance: total capped GCSE points (a student's best 8 subjects), and whether a student met the school accountability benchmark of at least 5 good passes (grade C or higher in at least five GCSEs including English

Table 5
Effects of incentives on overall high stakes exam performance: by predicted treatment effects.

	Total GCSE points	Meet benchmark of 5 good passes		
	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive
Predicted High	-0.025 (0.044)	-0.067 (0.046)	-0.052* (0.030)	-0.048** (0.024)
Predicted High * Treated	0.082* (0.043)	0.152** (0.062)	0.050* (0.029)	0.031 (0.037)
Predicted Low * Treated	-0.025 (0.032)	0.029 (0.085)	-0.002 (0.038)	-0.019 (0.034)
Pupil Characteristics	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes
*p-value: Predicted High =Predicted High * Treated*	0.0283	0.0154	0.0127	0.0989
Observations	14,744	15,373	14,744	15,373
Schools	48	48	48	48

Note: OLS estimates. Dependent variable is normalized exam score. Standard errors clustered by school in parentheses. The sample includes students in the experimental year and in the prior cohort in the prior year. Estimates for Financial Incentives include schools assigned to the Financial Incentive and schools assigned to Control. Estimates for Non-Financial Incentives includes schools assigned to the Non-Financial Incentives and schools assigned to Control. Predicted High (Low) is an indicator variable for an above-median (below-median) predicted treatment effect for math in the relevant treatment. All estimates include school and year fixed effects and the following baseline characteristics: gender, race/ethnicity, Free School Meal status, English as an Additional Language status, subject-specific baseline test score, and month of birth. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4
Effects of incentives on high stakes exams: by leave-one-out predicted treatment effects.

	Math	English	Science			
	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive
Predicted High	-0.025 (0.063)	-0.083 (0.055)	-0.022 (0.044)	0.075 (0.048)	-0.082* (0.043)	0.062* (0.034)
Predicted High * Treated	0.135** (0.054)	0.121* (0.069)	-0.002 (0.048)	0.079 (0.064)	0.117 (0.082)	0.081 (0.057)
Predicted Low * Treated	-0.024 (0.051)	-0.050 (0.038)	-0.002 (0.066)	-0.027 (0.058)	0.053 (0.065)	0.031 (0.060)
Pupil Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
*p-value: Predicted High = Predicted High * Treated*	0.0408	.0436	0.7145	0.9581	0.0683	0.7996
Observations	14,947	15,551	14,773	15,359	15,386	16,069
Schools	48	48	48	48	48	48

Note: OLS estimates. Dependent variable is normalized exam score. Standard errors clustered by school in parentheses. The sample includes students in the experimental year and in the prior cohort in the prior year. Estimates for Financial Incentives include schools assigned to the Financial Incentive and schools assigned to Control. Estimates for Non-Financial Incentives include schools assigned to the Non-Financial Incentives and schools assigned to Control. Predicted High (Low) is an indicator variable for an above-median (below-median) predicted treatment effect for the relevant subject and treatment. All estimates include school and year fixed effects and the following baseline characteristics: gender, race/ethnicity, Free School Meal status, English as an Additional Language status, subject-specific baseline test score, and month of birth. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

and math). The structure of the table is similar to Table 4 with students categorized as Predicted High or Predicted Low based on their predicted treatment effects in math. For both the financial and non-financial incentives we find a significant impact on performance among students with Predicted High effects improving overall scores by 0.1–0.22 standard deviations, and increasing the proportion of students meeting the GCSE benchmark by 8 – 10 percentage points.[33] These effect sizes are similar to the +7 ppt of being female, and the -11 ppt effect of being poor (coefficients from the same regression but not reported).

Finally, in Table 6, we present the baseline characteristics of students predicted to have High vs. Low treatment effects, along with predicted math GCSE score and predicted behavior.[34] The most striking finding is that for both Financial and Non-Financial Incentives, students predicted to have High treatment effects have significantly lower attainment as measured by baseline scores in year 6, predicted GCSE scores, and predicted behavior. The gap in predicted math GCSE scores is 0.38–0.4 standard deviations. Our estimated treatment effects on math scores of 0.16–0.2 standard deviations among these students would close about half of the predicted attainment gap. This focus of the impact on low-attaining students is also illustrated in Fig. A1, which plots actual GCSE scores against predicted scores for both treatment and control groups, separately for each subject and both treatments. It is clear that the bulk of the effects are in the lowest two to three quintiles of predicted scores.

Table 6
Composition of high and low predicted incentive effectiveness.

Financial Incentives	Predicted Effectiveness		
	High	Low	*p-value* High = Low
Female (%)	53.8	51.0	0.51
FSM (%)	40.0	36.9	0.28
EAL (%)	15.7	83.9	0.00
Asian ethnicity (%)	3.7	47.5	0.00
Black ethnicity (%)	37.6	12.7	0.00
White ethnicity (%)	53.4	28.9	0.00
Baseline attainment	3.33	3.70	0.00
Predicted Math GCSE	-0.461	0.081	0.00
Predicted Behavior	-0.199	-0.004	0.00
N	4966	4861	
Non-Financial Incentives	Predicted Effectiveness		
	High	Low	*p-value* High = Low
Female (%)	68.2	37.6	0.00
FSM (%)	30.0	46.4	0.00
EAL (%)	52.5	46.4	0.25
Asian ethnicity (%)	28.8	22.1	0.16
Black ethnicity (%)	24.5	26.0	0.60
White ethnicity (%)	41.4	41.1	0.94
Baseline attainment	3.06	3.93	0.00
Predicted Math GCSE	-0.397	0.000	0.00
Predicted Behavior	-0.151	-0.057	0.00
N	4773	5054	

Note: The table reports means by Predicted High (above-median) or Low (below-median) predicted treatment effects for math in the relevant treatment. Predicted Math GCSE is calculated based on personal characteristics and school attended from estimation using four pre-treatment years; Predicted Behavior calculated based on personal characteristics from estimation using control schools in the treatment year. The final column reports the p-value for a test of equality between the High and Low groups with standard errors clustered at the school level.

Turning to demographic characteristics, we find different patterns for the Financial and Non-Financial Incentives treatments. For Financial Incentives, there are large differences in the ethnic composition of the High and Low groups. High effectiveness students are significantly more likely to be white or black and significantly less likely to be Asian. The largest difference between the High and Low groups is EAL status. Only 15.7% of students predicted to have High treatment effects are classified as EAL, compared to 83.9% of students predicted to have Low treatment effects. We find similar results using the Causal Forests approach. Because of the high-dimensional complexity of the forest, isolating the impact of one particular variable in this analysis is not straightforward. To gage the overall contribution that each variable makes we re-ran the analysis, sequentially dropping each covariate.[35] Dropping EAL status produced the largest change in the variance, suggesting that this variable accounts for a lot of the heterogeneity. Taken together, the results of our machine learning approaches support the results of the subsample analysis – that non-EAL students particularly benefit from the financial incentives treatment.

For Non-Financial Incentives, there are no significant differences in ethnic composition or EAL status. Instead, the High Effectiveness students are more likely to be female and are less likely to qualify for Free School Meals (FSM), which is a proxy for family income. As discussed in the next section, these are the students whose effort and behaviors improve in response to the non-financial rewards.

4.3. *Mechanisms–the role of behaviors and student characteristics*

Why do we find an overall null effect? And why are incentives effective at improving test scores for some students and not for others? In this section we discuss several potential mechanisms for our results. We first examine the extent to which the results we discussed above map into the impact of our treatments on the behaviors we directly incentivized: attendance, class conduct, classwork and homework–measured by the number of times a student met the behavior threshold over the course of the year (normalized to have mean 0 and standard deviation 1). We then discuss additional mechanisms including communication of the incentives, structure of the incentives, measurement of test scores and student motivation.

As discussed in Section 2, teachers were responsible for evaluating students' behaviors (other than for attendance). The judgements of all the teachers are subjective, and it is certainly possible in treatment schools that teachers could "game" the process in some way – grading their favorite students more generously, giving them fewer assignments, etc. It is worth re-emphasizing here that the key exam scores used in our analysis do not come from the class teachers; these exams are marked remotely from the school to a consistent national standard.

Table 7 has a similar structure to Tables 4 and 5 except that the sample only includes students in the experimental cohort (because we do not have behavioral measures for prior cohorts) and so the regression does not include school fixed effects (we instead include the school value added in math estimated using the four prior cohorts). As in Table 5, we estimate the impact of incentives split by predicted treatment effects

[33] The effects on overall GCSE performance are primarily driven by the impact of treatment on math and science scores, as we find little impact on English scores. We also find no evidence of wider spillovers, either positive or negative, to grades on non-incentivized subjects – for example, on French and History as popular options (see Appendix Table 7). Similar to the results in Table 2, there is no impact on overall GCSE performance in the full population (results available upon request).

[34] We estimated coefficients for predicted GCSE scores using the specification in Equation 2 and four pre-treatment cohorts. We then applied these coefficients, including school fixed effects, to our experimental sample to generate predicted GCSE scores. We estimated coefficients for predicted behavior (average of the four behavior measures, attendance, conduct, classwork and homework) using the specification in Equation 1 and students in the control group only. We then applied these coefficients to our experimental sample to generate predicted behavior. Predicted measures are in normalized units.

[35] Thanks to Jonathan Davis for this suggestion. (results available on request).

Table 7
Effects of incentives on behaviors: by predicted treatment effects.

	Attendance	Conduct	Classwork	Homework	Overall
Panel A: Financial					
Predicted High * Treated	0.068 (0.142)	-0.021 (0.145)	0.424** (0.203)	0.053 (0.216)	0.176 (0.168)
Predicted Low * Treated	0.302 (0.241)	0.084 (0.162)	0.294 (0.211)	0.052 (0.191)	0.238 (0.218)
Predicted High	0.218 (0.185)	-0.040 (0.145)	-0.077 (0.182)	0.001 (0.159)	0.032 (0.166)
*p-value: Predicted High = Predicted High * Treated*	0.842	0.707	0.064	0.455	0.222
Observations	4476	4476	4476	4476	4476
Schools	30	30	30	30	30
Panel B: Non-Financial					
Predicted High * Treated	0.010 (0.185)	0.247 (0.165)	0.515** (0.225)	0.344* (0.198)	0.369* (0.191)
Predicted Low * Treated	0.226 (0.221)	0.188 (0.203)	0.072 (0.189)	-0.025 (0.203)	0.143 (0.223)
Predicted High	0.060 (0.174)	0.082 (0.180)	-0.267 (0.180)	-0.012 (0.200)	-0.050 (0.199)
*p-value: Predicted High = Predicted High * Treated*	0.652	0.098	0.014	0.022	0.024
Observations	4716	4716	4716	4716	4716
Schools	30	30	30	30	30
Pupil Characteristics	Yes	Yes	Yes	Yes	Yes
School Value Added	Yes	Yes	Yes	Yes	Yes
Control group proportion:					
Achieved threshold	0.676	0.767	0.627	0.651	0.680
Achieved threshold every term	0.480	0.600	0.432	0.463	0.201

Note: OLS estimates. Dependent variable is the normalized behavioral outcome reported in each column. Standard errors clustered by school in parentheses. The sample includes students in the experimental year. Estimates for Financial Incentives (Panel A) include schools assigned to the Financial Incentive and schools assigned to Control. Estimates for Non-Financial Incentives (Panel B) include schools assigned to the Non-Financial Incentives and schools assigned to Control. Predicted High (Low) is an indicator variable for an above-median (below-median) predicted treatment effect for math in the relevant treatment. All estimates include school-level value added in math and the following baseline characteristics: gender, race/ethnicity, Free School Meal status, English as an Additional Language status, subject-specific baseline test score, and month of birth. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

using the predicted effects on GCSE math scores.[36] The dependent variables are attendance, conduct, classwork, homework and overall behavior (average of the four measures) normalized using the control group. Among students with Predicted High effects, for Non-Financial Incentives the estimated treatment effects are positive and significant for all behaviors except attendance. For the Financial Incentive, the only significant impact is on completing classwork.

The differences in the pattern of results for Financial and Non-Financial Incentives may be due to noise resulting from measurement error, inaccurate teacher reporting, or multiple hypothesis testing. We also note that only eighteen of the thirty-three control schools tracked behavioral measures and so we interpret the treatment effects with caution. However, to the extent these estimates reflect the true impact of the intervention on student behavior, they suggest that there is greater heterogeneity in the responsiveness to non-financial incentives – i.e., some students are motivated by the event rewards and others are not.

If behavioral responsiveness to financial incentives is more homogeneous, why do we see heterogeneous effects on exam scores? One possibility is that across different students, the same marginal increase in effort can have different marginal impacts on exam performance. For example, a unit increase in effort by highly motivated, high performing students who are already working hard may have less of an impact on exam scores than an increase in effort by under-motivated, low-performing students. A related possibility is that the particular effort behaviors we incentivized may have differential impacts on student performance – e.g., improving classwork may not improve test scores if the marginal classwork for that student is not related to material on the test. We did not design our experiment to identify the causal linkages between the incentivized behavioral inputs into the human capital production function and test score outputs. Separately identifying not only the marginal impacts of behaviors, skills and motivation, but also heterogeneity in those impacts, could inform which inputs policy interventions should target, and for which students (Cotton, Hickman, List, Price, & Roy, 2020).

We next consider three additional mechanisms that could be partially driving our results: awareness of the incentives, structure of the incentives, and measurement of test scores. First, it is possible that (some) students were not fully aware of the incentives or did not understand what they needed to do to earn incentives. A lack of understanding could help explain our overall null effects. The schools shared information with students, though as noted in Section 3.4 two treatment schools did not perfectly comply, which may attenuate our effects. We also sent letters home via the schools that explained the program and provided feedback on students' performance every half term. It is possible that households responded differently to the letters (see e.g., Berry, 2015 for discussion). For example, if some parents of EAL students were less likely to understand the letters, this could be a potential reason why the incentives had smaller impacts on the performance of EAL students. Because we do not have a measure of students' or parents' knowledge of the incentives, we are not able to address this hypothesis directly.

Second, the threshold structure of our incentives may have differentially motivated students. As discussed in Section 2.1, we set individual targets for homework and classwork and thresholds of no absences and no more than one sanction for attendance and conduct respectively. We expect the incentives to have the largest impact among students who are not meeting the thresholds at baseline but view them as attainable (see Levitt et al., 2016b; Campos-Marcade and Wengstrom, 2020 for discussion). If there is a large share of students already meeting the thresholds in the absence of incentives, or a large share who view the incentives as unattainable, this could help explain our overall null effects. The bottom of Table 7 reports the proportion of the control group that met each behavioral incentive threshold (i.e., in the absence of incentives). Each pupil had 16 targets over the year, one for each of the four incentivized elements of behavior in the four half-terms. On average, control students met about two-thirds of the targets, suggesting the thresholds are attainable. However, only one-fifth of students met every threshold, suggesting that most students had room for improvement.

Third, we consider the role of test score measurement. As noted in Section 3.1, 60 percent of the English GCSE score is determined by the controlled assessment which includes one year of coursework prior to

[36] In Appendix Table 8, we report estimates for the full sample with and without value added controls. The estimated effects are all positive but are largely not statistically significant. Excluding the value-added measures does not affect the results.

our intervention and the year of coursework during our intervention. We are not able to separately identify the weight given to the pre-intervention year of coursework, but if the controlled assessment is split evenly across the two years, then 30 percent of the GCSE grade in English could be due to grades students received prior to our intervention. The controlled assessment receives less weight in Science (25 percent) and no weight in math. This may be one reason why we find smaller treatment effects in English even among those predicted to have high treatment effects. However, it may also be the case that it is more difficult to improve scores in English than in math.[37]

Finally, we return to the role of pupil motivation. As discussed above, our incentives are most likely to be effective for students who lack motivation at baseline. While we did not measure motivation directly, Burgess & Heller-Sahlgren (2018) show that in England children of immigrants have more positive attitudes to school as a way of progressing in life than do native students. If immigrant students – as proxied by EAL status – are already putting forth high levels of effort in the absence of our incentives, they may have little room to move in response to them. This may be one reason why non-EAL status emerges from both the subsample and the machine learning analyses as a leading predictor of responsiveness to incentives. However, other mechanisms including those discussed above may also be driving our results.

4.4. Identifying high effectiveness schools

With an eye to policy implementation, we examine whether individual level heterogeneity can be used to predict treatment effects at the school level. This is of particular policy relevance for targeting between schools; within-school targeting of this intervention is not attractive to schools.

There are three conditions for the identification of heterogeneous responses to be practically applicable for policy at a cluster (i.e., school) level. First, as at the individual level, there needs to be sufficient heterogeneity that can be identified using baseline characteristics; small between-school differences would not be worthwhile exploiting. Second, the "high effectiveness" students need to be of significant policy interest; for example, disadvantaged or low-performing students. Third, in a cluster setting, the high effectiveness students need to be "separable" from the low effectiveness students. That is, there needs to be sufficient "segregation" between schools in terms of those characteristics. If students with high predicted effects are evenly distributed across schools, then there will be little scope for targeting an intervention at the school level. For example, if gender is the primary driver of heterogeneity, then the only source of school targeting would be the small minority of single sex schools.

To test our predictive power at the school level, we estimate our difference-in-difference specification from Table 2 but limit the sample to schools with above-median proportions of students with Predicted High treatment effects on the relevant outcome. The estimates reported in Table 8 suggest that there is indeed scope for targeting incentives at the school level as a means of improving math and science scores, again finding effect sizes of around 0.10–0.15 standard deviations. Reassuringly, there is high correlation between the predicted ``high effectiveness'' schools using the causal forest and leave-one-out approaches, particularly in math where the correlation is 0.89 for Financial incentives and 0.78 for Non-financial Incentives.

We can apply our estimated profile of "high effectiveness students" (i.e., above median predicted treatment effects) to the national data to produce a suggestive estimate of the number of such students would be eligible for targeting, and what therefore is the scope for targeting at the school-level.[38] These suggestive estimates obviously need to be treated with caution as we are extrapolating from a sample of schools in the poorest neighborhoods. There are 1300 high schools (out of about 3000) in which the fraction of high effectiveness pupils is at least 75%. Focusing down further, there are 240 of those schools with high school-level poverty rates (above 20%), containing more than 37,000 students. Furthermore, across all high poverty schools, about half of students are estimated to be high effectiveness.

5. Conclusion

We report results from a large field experiment with 63 high schools and over 10,000 students where we randomized incentives for students to increase their inputs to the education production function. We measure the impact of such incentives on high-stakes assessments taken by all students. The objective of the incentives was to raise pupils' effort and engagement in school, and thereby increase their performance on the high-stakes assessment. We implemented two systems of behavior incentives: a Financial treatment that rewarded pupils with cash, and a Non-Financial treatment that offered high-value trips of their own choosing for successful students.

There is little average impact of the incentives. We investigate the distribution of treatment effects using the large sample and rich dataset, and identify a "right tail". We show that half of the students have economically meaningful positive effects. For students with above-median predicted effects, we estimate that exam scores improve by 10% to 20% of a standard deviation and that the standard benchmark "pass rate"[39] increases by 8–10 percentage points. We confirm simple subgroup findings using machine learning techniques: the "right tail" of highly-responsive students is well proxied by a single demographic characteristic: those who are native English speakers.

The particular nature of the estimated heterogeneity helps to further our understanding of pupil responses to incentives. Clearly, students already have large incentives to invest in education: the returns they will experience in later life, including increased earnings, better health, longer life expectancy, and higher self-reported well-being (Oreopoulos, 2007; Oreopoulos & Salvanes, 2011). For students who have already internalized the inherent incentives for working hard in school, additional rewards may add little further motivation.[40] But other students may underestimate or give little weight to the future benefits of education.[41] They may not fully understand the role of effort in the education production function (rather than say innate ability or parental

[37] Much of the literature finds that educational interventions have larger effects on math than on English achievement (see for example, Decker, Mayer & Glazerman, 2004; Sadoff, 2014; Jacob, 2005; Dobbie & Fryer, 2011; Levitt et al., 2016b). One explanation for this result is that math scores are more sensitive to effort in school than reading. Curto and Fryer (2014), *p.* (80) point out that almost all of a student's math experience is within the classroom (so a change there has a big overall effect), whereas English skills – reading and writing – are also developed throughout non-school life. However, as noted above, Fryer (2017) estimates similar meta-coefficients for the effect of incentives on math and reading, 0.024 and 0.021 respectively.

[38] We use the full experimental sample to estimate the coefficients to predict treatment effects on math scores for the financial incentive. We then apply those coefficients to every student in the national cohort to generate a predicted treatment effect for each student.

[39] Getting at least 5 grade C score or better across all subjects taken.

[40] Students may internalize the returns to education directly, or as in Becker's seminal model (1981,1991) of the family, parents can induce children's investment in schooling through parental transfers.

[41] Work in psychology, neurology and behavioral economics has shown that children and adolescents tend to focus on the present and give little weight to the consequences their decisions will have in the future (Lavecchia, Lui & Oreopoulos, 2014) provide a review). Recent studies have linked this behavior to educational investment, finding that impatience and high discount rates are negatively correlated with educational outcomes (Kirby et al., 2002; Kirby, Winston and Santiesteban, 2005; Castillo Ferraro, Jordan & Petrie, 2011; Cadena & Keys, 2015). These students are also more likely to regret dropping out of school (Cadena and Keys, 2015).

Table 8
Effects of incentives in high predicted effectiveness schools.

	Math		English		Science	
	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive
Financial Incentive	0.106 (0.075)		0.056 (0.034)		0.131 (0.099)	
Non-Financial Incentive		0. (0.048)		0.077 (0.083)		0.(0.080)
Pupil Chars.	Yes	Yes	Yes	Yes	Yes	Yes
School F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6949	7548	6611	8836	7636	7870
Schools	27	26	27	27	22	27

Note: OLS estimates. Dependent variable is normalized exam score. Standard errors clustered by school in parentheses. The sample includes students in the experimental year and in the prior cohort in the prior year in schools with above-median fractions of students with Predicted High (above-median) predicted treatment effects in the relevant subject and treatment. Estimates for Financial Incentives include schools assigned to the Financial Incentive and schools assigned to Control. Estimates for Non-Financial Incentives include schools assigned to the Non-Financial Incentives and schools assigned to Control. All estimates include school and year fixed effects and the following baseline characteristics: gender, race/ethnicity, Free School Meal status, English as an Additional Language status, subject-specific baseline test score, and month of birth. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

resources), or they may perceive effort in school to be very costly (Fryer, 2011; Levitt et al., 2016a; Levitt et al., 2016b). It seems likely then that there will be diverse responses to incentives: powerful for some, irrelevant for others who are already well motivated.

The effect sizes we estimate are similar to the impact of a one standard deviation improvement in teacher quality (Chetty, Friedman, & Rockoff, 2014b; Slater and Burgess, 2012). The economic impact of the intervention is also likely to be large because of the high estimated earnings rate of return for passing the benchmark of 5 good GCSEs (Battistin et al., 2012; McIntosh, 2006).[42] The treatment effects are stronger and more robust for the Financial incentives than the Non-Financial incentives. However, given their low cost and the ease of administration, our results suggest that non-financial rewards may provide a feasible and cost-effective alternative to financial incentives. At approximately one-quarter to one-third the cost of financial incentives (when including administrative costs), back-of-the-envelope calculations using GCSE pass rates suggest that non-financial rewards could be two to three times more cost effective than financial incentives.

Finally, our results suggest that targeting incentives on schools with high fractions of students with high predicted treatment effects could help in closing achievement gaps. In our sample, the estimated impact of the incentives would close about half of the predicted attainment gap. School-targeting is also attractive from a cost-effectiveness perspective. Implementing the Financial Incentive treatment corresponds to an increase of approximately 3.5 percent in per-pupil school spending, or approximately $1500 per standard deviation increase.[43] While it may be difficult to justify an across the board spending increase of this size, a targeted approach that increases low-performing students' high stakes exam scores by 10–15% of a standard deviation much more easily satisfies cost-effectiveness calculations.

As we discuss above, targeting at the school level is a more feasible policy than targeting at the individual student level. We demonstrate that it is possible to identify schools that are likely to differentially benefit from the incentives. However, there may be political constraints on the criteria used as the basis for targeting. For example, it may be acceptable to target schools with high proportions of low income or low performing students; but it may not be palatable to target schools based on the proportion of native-born students. Incorporating political constraints with machine learning analysis could allow for broader exploration of how to target and scale up experimental interventions.

Acknowledgments

Many thanks to the Education Endowment Foundation (EEF) for funding this project, and to Kevan Collins, Milly Nevill and Dan Sinnott at EEF for all their support. Many thanks also to: Julia Carey, senior project manager, and her team Zennon Sherley-Dale, Jamie Atkins and Christine Spencer; our lead education consultants Carole Baker and Rhona Sevieur. Many thanks to Jon Davis, Justin Holz, Jennifer Mayo, and Matthew Burgess for assistance during the project; and to Rebecca Allen, Steven Levitt and John List for very helpful discussions at the set-up of this project. Thanks for comments on earlier drafts to Eric Taylor, Frank Windmeijer and to seminar attendees at NHH in Bergen and the University of Sussex. Many thanks also to the Department for Education for supplying the National Pupil Database. Finally, many thanks also to all the Headteachers, administrators and teachers for their part in implementing this project.

Appendix

[42] Using observational data, Battistin et al. (2012) find a 26% penalty in earnings at age 33 to leaving school with no qualifications as opposed to some, and McIntosh (2006) finds a 27%–29% return for 5 good GCSE passes using different data.

[43] Average cost of £211per pupil per year, against an approximate average of over £6000 per pupil per year going to schools and an average estimated increase among high effectiveness pupils of 0.14 standard deviations in math (Table 4). See https://www.ifs.org.uk/bns/bn121.pdf

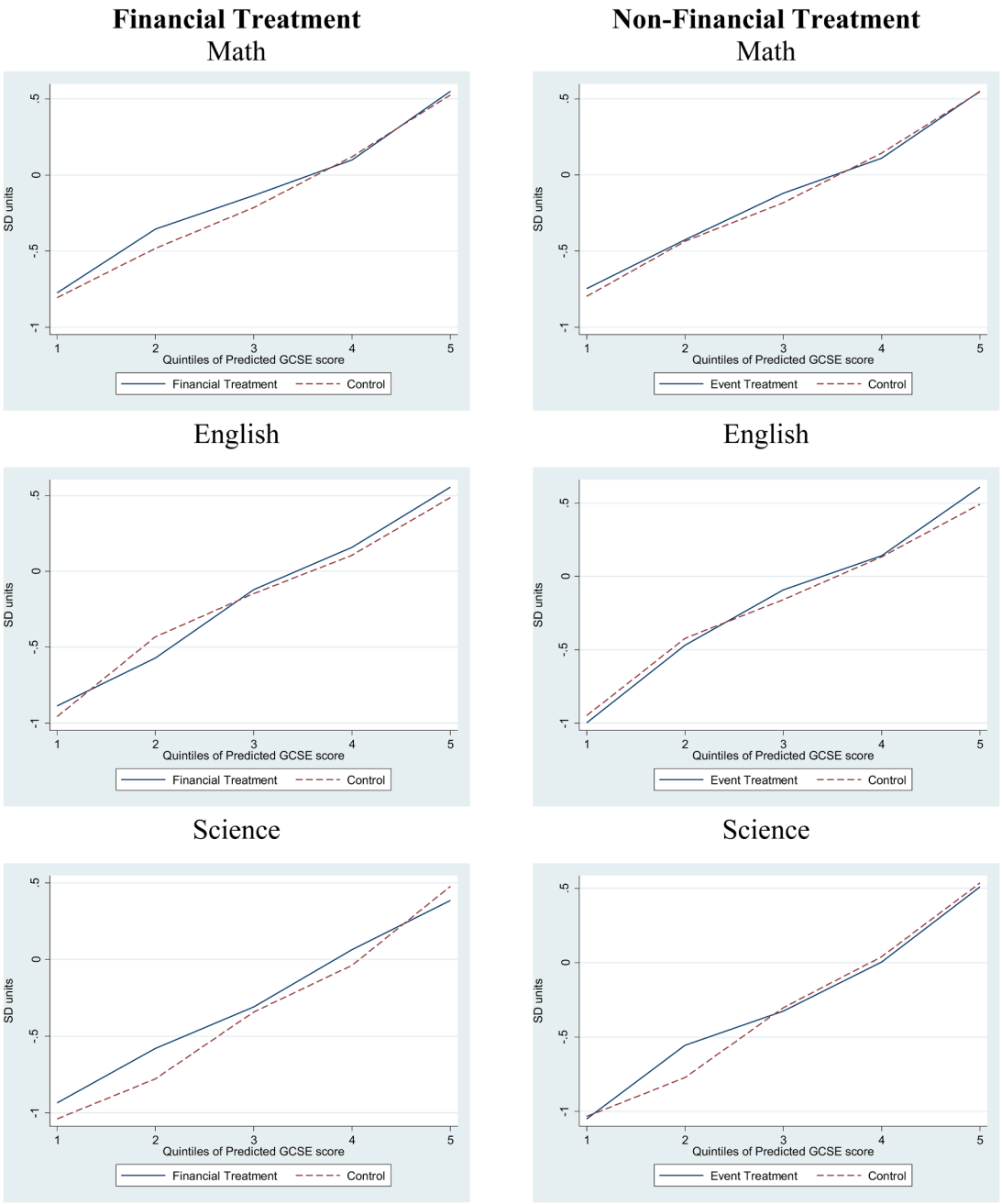

Fig. A1. Treatment Effects by Predicted GCSE Score Note: Vertical axis: actual normalized GCSE grades in experimental year, by treatment status and subject. Horizontal axis: predicted grade estimated using all prior years of data, pupil characteristics and school fixed effects.

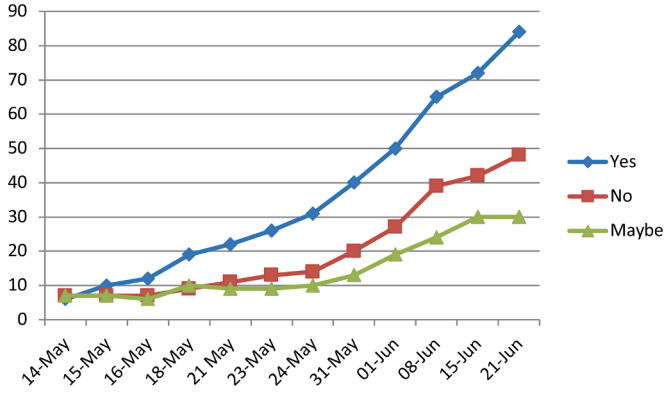

Fig. C1. Recruitment Trends.

References

Abadie, A., Chingos, M. M., & West, M. R. (2018). Endogenous stratification in randomized experiments. *Review of Economics and Statistics, 100*(4), 567–580.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: a reevaluation of the abecedarian, Perry preschool, and early training projects. >*Journal of the American Statistical Association, 103*(484), 1481–1495.

Angrist, J., Hudson, S., & Pallais, A. (2016). *Evaluating post-secondary aid: enrollment, persistence, and projected completion effects*. No. w23015. *National Bureau of Economic Research*.

S. Athey, & G.W. Imbens (2016a). "The econometrics of randomized experiments." arXiv preprint arXiv:1607.00698.

Athey, S., & Imbens, G. W. (2016b). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113*(27), 7353–7360.

Austen-Smith, D., & Fryer, R. G. (2005). An economic analysis of `acting white. *Quarterly Journal of Economics, 119*(2), 551–582.

Battistin, E., Nadai, M. D., & Sianesi, B. (2012). Misreported Schooling, multiple measures and returns to educational qualifications. *Journal of Econometrics, 181*(2), 136–150.

Becker, G. (1981). *A treatise on the family* (Enl. ed. 1991). Harvard University Press.
Beilock, S. (2010). *Choke: What the secrets of the brain reveal about getting it right when you have to*. Simon and Schuster.
Berry, J. (2015). Child control in education decisions: An evaluation of targeted incentives to learn in India. *Journal of Human Resources, 50*(4), 1051–1080.
Bruhn, M., & McKenzie, D. (2009). In Pursuit of balance: randomization in practice in development field experiments. *American Economic Journal: Applied Economics, 1*(4), 200–232.
S. Burgess, & G. Heller-Sahlgren (2018). Motivated to succeed? Attitudes to education among native and immigrant pupils in England. IZA DP No. 11678, IZA.org.
Burgess, S., Wilson, D., & Piebalga, A. (2009). Land of hope and dreams: Education aspirations and parental influence among England's ethnic minorities. *Mimeo, CMPO*.
Cadena, B. C., & Keys, B. J. (2015). Human capital and the lifetime costs of impatience. *American Economic Journal: Economic Policy, 7*(3), 126–153.
Castillo, M., Ferraro, P. J., Jordan, J. L., & Petrie, R. (2011). The today and tomorrow of kids: Time preferences and educational outcomes of children. *Journal of Public Economics, 95*(11), 1377–1385.
Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review, 104*(9), 2593–2632.
Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review, 104*(9), 2633–2679.
Cotton, C., Hickman, B. R., List, J. A., Price, J., & Roy, S. (2020). *Productivity versus motivation in adolescent human capital production: evidence from a structurally-motivated field experiment*. No. w27995. *National Bureau of Economic Research*.
Curto, V., & Fryer, RG., Jr. (2014). The potential of urban boarding schools for the poor: Evidence from seed. *Journal of Labor Economics, 32*(1), 65–94.
Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review, 64*, 313–342. in press.
Davis, J. M. V., & Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review Papers and Proceedings, 107*(5), 546–550.
J.M.V. Davis, & B. Sara Heller (forthcoming) rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. Forthcoming, Review of Economics and Statistics.
Dearden, L., Emmerson, C., Frayne, C., & Meghir, C. (2009). Conditional cash transfers and school dropout rates. *Journal of Human Resources, 44*(4), 827–857.
Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). The effects of teach for America on students: Findings from a national evaluation. *Mathematica Policy Research Report No., 8792*, 750.
Dobbie, W., & Fryer, R. G. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the harlem children's zone. *American Economic Journal: Applied Economics, 3*(3), 158–187.
Fryer, G. R., Jr. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *The Quarterly Journal of Economics, 126*(4), 1755–1798.
Fryer, R. G. (2017). The Production of human capital in developed countries: Evidence from 196 randomized field experimentsa. *Handbook of Economic Field Experiments, 2*, 95–322.
R. Fryer, & R. Holden (2013). "Multitasking, dynamic complementaries, and incentives: A cautionary tale." Working Paper.
R. Fryer (2013). "Information and student achievement: evidence from a cellular phone experiment" NBER Working Paper no. 19113.
Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review Insights, 1*, 291–308. no. 3.
Handel, B., & Kolstad, J. (2017). Wearable technologies and health behaviors: New data and new methods to understand population health. *American economic review, papers and proceedings*. Forthcoming.
Hanushek, EA. (2009). The economic value of education and cognitive skills. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 39–56).
S. Hirshleifer (2017). "Incentives for effort or outputs? A Field experiment to improve student performance" University of California, Riverside, Mimeo. https://economics.ucr.edu/pacdev/pacdev-papers/incentives_for_effort.pdf.
Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics, 7*(1), 443–470.
Jacob, B. (2005). Accountability, incentives and behavior: evidence from school reform in Chicago. *Journal of Public Economics, 89*(5-6), 761–796.
Jalava, N., Schrøter Joensen, J., & Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization, 115*, 161–196.
Kirby, K. N., Godoy, R., Reyes-Garcia, V., Byron, E., Apaza, L., Leonard, W., Perez, E., Vadez, V., & David and Wilkie. (2002). Correlates of delay-discount rates: Evidence from Tsimane'Amerindians of the Bolivian rain forest. *Journal of Economic Psychology, 23*(3), 291–316.
Kirby, K. N., Winston, G. C., & Santiesteban, M. (2005). Impatience and grades: Delay-discount rates correlate negatively with college GPA. *Learning and Individual Differences, 15*(3), 213–222.
Kohn, A. (1999). *Punished by rewards: The trouble with gold stars, incentive plans, a's, praise, and other bribes*. Houghton Miin Harcourt.
Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253.
Lavecchia, AM., Liu, H., & Oreopoulos, P. (2014). Behavioral economics of education: Progress and possibilities. *NBER Working Paper*, 20609.
Levitt Steven, D., John, A., & List, Susanne Neckermann and Sally Sadoff. (2016a). The Behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy, 8*(4), 183–219.
Levitt, S. D., List, J. A., & Sadoff, S. (2016b). The Effect of performance-based incentives on educational achievement: Evidence from a randomized experiment. *NBER Working Paper*, 22107.
McIntosh, S. (2006). Further analysis of the returns to academic and vocational qualifications. *Oxford Bulletin of Economics & Statistics, 68*(2), 225–251.
K. Michelmore, & S. Dynarski (2016). "The gap within the gap: Using Longitudinal data to understand income differences in student achievement," NBER Working Paper No. 22474.
Ofqual, (2013). "Review of controlled assessments in GCSEs," Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/377903/2013-06-11-review-of-controlled-assessment-in-GCSEs.pdf.
Oreopoulos, P., & Salvanes, K. G. (2011). Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives, 25*(1), 159–184.
Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics, 91*(11), 2213–2229.
Sadoff, S. (2014). The role of experimentation in education policy. *Oxford Review of Economic Policy, 30*(4), 597–620.
Slater, D., & Burgess, S. (2012). Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics, 74*(5), 629–645.
Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242.
V. Wagner, & G. Riener (2015) Peers or parents? On non-monetary incentives in schools. No. 203. DICE Discussion Paper.
Wilson, D., Burgess, S., & Briggs, A. (2011). The dynamics of school attainment of England's ethnic minorities. *Journal of Population Economics, 24*(2), 681–700.