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

Simon Burgess
Robert Metcalfe
Sally Sadoff

Discussion Paper 16 / 678

17 October 2016

Department of Economics
University of Bristol
Priory Road Complex
Bristol BS8 1TU
United Kingdom

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

Simon Burgess
University of Bristol

Robert Metcalfe
University of Chicago

Sally Sadoff
University of California, San Diego

September 2016

Abstract

We analyze the impact of incentivizing students' effort during the school year on performance on high-stakes assessments in a field experiment with 63 low-income high schools and over 10,000 students. We contribute to the literature on education incentives by incentivising inputs rather than output, by focusing on high stakes outcomes, and by comparing financial and non-financial rewards. We take advantage of our large sample and rich data to explore heterogeneity in the effects of incentives, and identify a "right tail" of underperforming students who experience a significant impact on high stakes assessments. Among students in the upper half of the distribution of incentive effectiveness, exam scores improve by 10% to 20% of a standard deviation, equal to about half the attainment gap between poor and non-poor students.

Many thanks 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; Justin Holz and Jennifer Mayo for research assistance during the set-up of 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 Education Endowment Foundation for funding, and to Kevan Collins, Milly Nevill and Dan Sinnott at EEF for all their support; and 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.

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 (Michelmore and Dynarski, 2016; Hanushek, 2009). Increasingly, many governments are turning to new ideas in an attempt to deal with this problem. One of these is the potential for incentives to increase student motivation and change behaviors in schools. As we discuss in more detail below, there is now a growing body of research in both developing and developed countries on the overall impact of educational incentives. However, not a great deal is known about two critical questions within this literature: What types of incentives? And for which students?

To address these questions, we first ask, why might incentives for student effort and engagement help? 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 and Salvanes, 2011). For students who have already internalized the inherent incentives for working hard in school, additional rewards may add little further motivation.[1] But other students may be responsive to short term incentives for effort. These students may underestimate or give little weight to the future benefits of education.[2] They may not fully understand the role of effort in the education production function (rather than say innate ability or parental resources), or they may perceive effort in school to be very costly (Fryer, 2011; Levitt et al., 2016). It seems likely then that there will be diverse responses to incentives: powerful for some, irrelevant for others who are already well motivated.

In this paper we report the results of a large scale field experiment offering students repeated near-term incentives in the final year of compulsory schooling leading up to the high stakes assessments in England, called the General Certificate of Secondary Education (GCSE). GCSEs serve as the primary gatekeeper for students to stay in school and progress to

[1] Students may internalize the returns to education directly, or as in Becker's seminal model (1974, 1991) of the family, parents can induce children's investment in schooling through parental transfers.

[2] 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 (Gruber, 2001; Bettinger and Slonim, 2007; Steinberg et al., 2009; Lavecchia, Lui and 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, 2005; Castillo et al., 2011; Cadena and Keys, 2015). These students are also more likely to regret dropping out of school (Cadena and Keys, 2015).

university.[3] The GCSEs are also key qualifications in the labor market for those not continuing in education. GCSEs are typically two-year courses, taken in the final two years of compulsory schooling (year 10 and 11) when students are 15 to 16 years old.

The incentives we test are short-term and designed to motivate student effort and engagement during the final year of GCSE courses. They are based on multiple dimensions of effort (attendance, behaviour, classwork, and homework) and are measured over a five-week period (a half-term), with four periods between September and April. Within this structure, we compare the impact of financial incentives and non-financial incentives. The financial incentives offered students up to £80 per half-term (for a total of £320 over the year).[4] The non-financial incentives offered students the chance to qualify for a high-value event determined jointly by the school and students such as a trip to the national soccer stadium or a theme amusement park.

Our primary outcome of interest is the GCSE assessment scores on the subjects for which behavior was incentivised: English, math and science. We find little average impact in the full sample of either the financial or non-financial incentives on GCSE performance in the intention-to-treat analysis. However, as discussed above, we expect incentives to primarily have an impact among those students who are not responding to the much larger inherent returns to education. We therefore focus much of our analysis on identifying the "right tail" of students whose performance can be moved by relatively small, short-term rewards. To do this, we take advantage of our large and diverse sample of students, and highlight the upper half of the distribution of incentive effectiveness.

Our experiment includes over 10,000 students in 63 schools spread across England, as shown in Figure 1. As discussed in more detail in Section 3, we targeted schools serving students living in neighborhoods in the bottom decile of neighborhood poverty. Within this constraint, there is substantial diversity at the student level. As shown in Table 1, about half

[3] Appendix A provides details on the structure of the education system in the England.
[4] This is equivalent to about $520 using July 2014 exchange rates and equates to just over 4 hours 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 enrollment rates by 4-7 percentage points (authors' calculations based on Dearden et al. 2009, Table 1) . The incentives are offered only after students complete year 11 and take the GCSE exam.

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 eligibility for free school meals (FSM) or neighbourhood poverty is 40-45%. This is different from experiments in the US targeting low-income neighborhoods or low-performing schools, which have been more highly segregated.[5]

Using a rich set of observable characteristics, we generate individual-level predicted treatment effects using a leave-one-out approach. We find evidence that there is a "right tail" of students – the upper half – who experience significant impacts on assessment scores. Among students predicted to have above median treatment effects, the financial incentives improve math and science GCSE scores by 0.16 and 0.20 standard deviations respectively. The effects of the non-financial incentives are similar to the financial incentives in math, but do not have a significant impact in science. We find no treatment effects of either incentive on performance in English.

Importantly, we find that the students who are predicted to be most responsive to treatment are also those with lower performance at baseline. Our results suggest that targeting incentives towards the most responsive students could help close achievement gaps on high stakes examinations by about half.[6] This achievement effect significantly increases the attractiveness of targeting financial incentives to these students as it raises cost-effectiveness (because a larger portion of the incentives are paid out on the margin rather than for baseline performance), and also from the perspective of policy makers aiming to close achievement gaps. Building on these results, we examine the extent to which individual-level predicted treatment effects can be used to target interventions at the school level, and find evidence that we are able to identify schools that would benefit differentially from incentives.

There is now a large literature on financial incentives in education (see e.g., Sadoff, 2014 for further discussion). Our results contribute to this literature by incentivizing inputs (effort and

[5] 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).
[6] Comparing students predicted to have High (above-median) vs. Low (below median) treatment effects, the predicted attainment gap in math GCSE is 0.38-0.4SD and the effect size of the intervention in maths is 0.16-0.2SD.

engagement) not outputs (test scores), by focusing on effort leading to high stakes exams, and by comparing the impact of financial and non-financial rewards.

Our program is closest in design to a randomized experiment conducted by Levitt et al. (2016) and a series of randomized experiments conducted by Fryer (2011). Fryer (2011) tests incentives for grade performance among ninth graders in Chicago; regular test performance among fourth and seventh graders in New York City; and reading books among second graders in Dallas. These programmes all used monthly (or near monthly) financial rewards given to students. Fryer finds that incentives for reading books (Dallas) have the largest effects (only among English speaking students). The test-based incentives (New York City) and grade-based incentives (Chicago) have little or no effect on achievement outcomes. Levitt et al. (2016) offer ninth grade students in Chicago Heights (a suburb of Chicago) monthly incentives for meeting an achievement standard based on grades, attendance, behaviour and diagnostic test scores. They find an impact on grades that carries over into the year after the program ends (but fades after that). Unlike our study, however, the tests in these contexts are generally not high stakes for the students.[7]

To our knowledge, all previous studies of incentives involving high stakes exams have offered rewards solely based on performance on the exam itself rather than on effort investment through the year. Angrist and Lavy (2009) study the impact of large incentives of up to $2,400 for passing a high stakes exam required for university admission in Israel. They find effects on pass rates but only for girls. In a developing country context, Kremer, Miguel and Thornton (2009) find that offering girls large scholarship incentives (covering two years of schooling) for performance on the secondary school admissions exam in Kenya improves scores in one of the two school districts where the program was tested. In related work, Jackson (2010) finds that adoption of a program in Texas that pays both students and teachers for passing grades on Advanced Placement exams is associated with increased scores. As discussed above, if students are not forward-looking, have inconsistent time preferences, or do not fully understand the production function, they may fail to put in the effort necessary

[7]In some cases, the exams are relatively high stakes for the schools in that the results measure school performance, which may help determine school resources.

throughout the year to earn an incentive that they will only receive at the end of the year or even several months after the exam takes place.[8]

As far as we know, no previous study has demonstrated that incentives for effort investment through the year can have an impact on high stakes exam performance that is not itself directly incentivized. Other work has demonstrated the importance of student effort on test performance and student sensitivity to delays in rewards. Levitt et al. (forthcoming) find that offering students incentives on a low-stakes tests improves scores ten to twenty percent of a standard deviation, purely through increased effort on the test, when rewards are given immediately after the test ends. However, there is no impact when rewards are delayed by a month. Metcalfe et al. (2011) find a similar sized impact of effort on the same high stakes assessments we study. They show that when the value of leisure increases during the exam period (due to the occurrence of the world's major soccer tournaments), test scores fall, particularly among male and disadvantaged students. These studies however focus on short-term effort during the exam period. They do not examine whether sustained incentives for student effort can increase achievement.

There is also little research on the impact of financial incentives relative to non-financial incentives. We compare financial and non-financial incentives in this context for several reasons. A growing body of research in behavioral economics demonstrates the power of non-financial rewards (e.g., Levitt et al., forthcoming provide a discussion). 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. And, the collective[9] nature of the non-financial reward in our program potentially allows us to tap into positive peer effects (e.g., wanting to attend the school-wide event with friends).[10]

Finally, in the context of education incentives, non-financial rewards have only been tested on a very short timescale. The Levitt et al. (forthcoming) study discussed above compares financial and non-financial incentives for test performance and finds that non-financial

[8]As we discuss below, the high-stakes exams in our context (as in similar settings) are marked remotely from the school and benchmarked across the entire country in a thorough process.

[9] To be clear, access to the trip depends on each pupil's own individual behaviour; it is not a collective criterion such as "all go or no-one goes".

[10] 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)

incentives can be significantly more cost effective than financial incentives. However, this is in a very different context of a short time frame (lasting between 15 minutes and an hour), small non-financial rewards (a trophy worth $3) and with little scope for peer effects. Related work, also in the context of motivating short-term effort on tests, has examined the role of feedback, grading structure and symbolic rewards with no financial value (e.g., Jalava et al., 2014 provide a discussion). As far as we know, ours is the first study to offer large non-financial rewards, for example a school trip, over the course of a full year. Ours is also the first to compare such rewards to financial incentives, and to test these treatments at the school level in a large-scale policy level experiment.

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). But have only recently gained attention as a tool for predicting responsiveness to interventions (Imai and Ratkovic, 2011; Athey and Imbens, 2015; Allcott and Kessler, 2015).

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 some wider conclusions.

2. Program details

Incentive design

The experiment 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: maths, 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.[11] Using target grades allows the performance measures to depend primarily on student effort relative to baseline ability – rather than setting a single quality 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 summarised below:

	Criterion	Financial incentives group	Non-financial incentives group
Attendance	No absences	£10	1 award token
Conduct	No more than one sanction	£10	1 award token
Homework	Complete all work on time at or above target level	£30	3 award tokens
Classwork	Complete all work on time at or above target level	£30	3 award tokens

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.[12] In order to participate in the trip, a student needed to earn at least three-quarters of the tickets over the two half-terms (i.e., 12 out of the maximum of 16).[13] The events were chosen by the

[11] Target grades are set by teachers for each pupil in each subject based on their interactions with the pupils in class and 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).

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

[13] 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 term 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 must earn at least 12 tickets to qualify for the event, students earn at least 4 tickets per half-term to qualify.

students and school administration collectively at the start of term, given a budget constraint.[14]

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.[15] 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 behaviour items 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.

Overall, payments to schools totalled £729k. This included a compensation payment to each school of £2k, as well as £540k to financial treatment schools and £119k to event treatment schools. The monetary value of the financial treatment was greater than that of the event treatment. This was partly by design and partly because the schools and students made choices well within the budget.[16] 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.

[14] 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).

[15] 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".

[16] 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. The information we gave to schools was: "We will allocate funding for the events in the following way: £1,000 fixed amount per term for the first 30 pupils who achieve their targets. As well as the fixed amount of £1,000 for the first 30 pupils, we will pay an £25 per head for each additional pupil who meets his/her targets, up to a maximum total amount of £6,000 per event."

Sampling frame and recruitment

The sampling frame was composed of regular state secondary schools in very disadvantaged areas, defined as the highest decile of neighbourhood poverty as measured by the Income Deprivation Affecting Children Index (IDACI), yielding a total of 296 schools.[17] 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.[18] The remaining sample included 279 schools covering 60 Local Authorities (out of 150 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 (<.2%) opted their student 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 Figure 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

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

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

[18] 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. \

GCSEs in English, math and science. Students must generally achieve a good pass (a grade of C grade 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. A smaller portion of the score depends on performance during the two years of coursework. 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.[19] Therefore, our treatment effect estimates may be an underestimate of the impact of incentives on GCSE performance.

We focus on GCSE scores in the core subjects: math, English (language), and science.[20] 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).[21] In the cohorts in our data, students receive one of the following grades on the GCSE, with A* being the highest: A*, A, B, C, D, E, F, 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 behaviours we incentivized: attendance, conduct, classwork and homework. The school administration reported the attendance and conduct measures. Classroom teachers reported students'

[19] 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 and given the multiplicity of exam boards, it is not straightforward to find out exactly what proportion of the overall grade is determined by coursework for each student.

[20] 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).

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

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).[22] 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 principal (called the Headteacher in England), 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[23]), 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.

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

[22] 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.

[23] 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 Whites in the regressions.

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 design effect of 0.19 – 0.32.[24]

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.[25] 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 randomisation procedure are given in Appendix E). We then used a straightforward re-randomisation procedure 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.[26] We re-randomized until all p-

[24] The realised ICC was 0.099 in maths, 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).

[25] 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, Burgess and Briggs (2011) for an analysis of the attainment of different ethnic groups.

[26] The full list is: 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 (excl Chinese); the proportion of pupils with Special Educational Needs; the proportion of pupils with English as an Additional Language; whether largest ethnic group in school is Black; whether largest ethnic group in school is Asian; whether largest ethnic group in school is White; a measure of pupil neighbourhood poverty; Value added (best 8 results) for low attainers; school average GCSE maths score; school average GCSE English score; school average GCSE science score; school average capped GCSE score trend 2009-2011; cohort average prior attainment, maths; cohort average prior attainment, English; cohort average prior attainment, science; cohort average prior attainment, average; the proportion of pupils achieving 5A*-C GCSEs; the average capped GCSE points; the school is in London; single sex school; Academy; has a sixth form; total number of pupils in school; school cohort size; total expenditure per pupil; Headteacher hired either Sep 2010 or Sep 2011; School has own incentive scheme: yes versus missing or no; school has own

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 in the experimental cohort.[27] 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 and McKenzie (2009).[28]

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.

incentive scheme: no versus missing or yes. Since this is a table with no significant effects by design, we do not present it here, but it is available from the authors.

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

[28] All of the school level characteristics are controlled by the inclusion of school fixed effects, and the full set of pupil level characteristics are included in the regressions.

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_{NH} and for student with Low predicted effects π_{NL} 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 normalise the GCSE scores year by year over the whole national cohort. Estimated effects are therefore interpretable as units of student level standard deviations. One student level SD in Maths is 1.8 grades; that is, almost the two-point grade difference between an A and a C. Because the intervention is delivered at school level, we cluster standard errors in all models at that level.

Compliance and Attrition

We have very low attrition rates for our main outcome, GCSE scores, because these exams are compulsory and the data are available through a 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), no different between treatment and control schools.[29]

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 behavioural data for the entire year. We analyze the characteristics of those leaving in Appendix Table 2 and find no evidence of differential attrition.

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 randomised 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 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 will argue at the end that schools can 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

[29] Results available upon request.

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.[30]

Our matched triplets design allows us to estimate treatment effects excluding non-compliers (in both treatment and control). In Appendix table 3 we present estimates of model 3 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. The estimates are very similar to the results for the full sample discussed in the next section.

4. Results

We first present the results for our primary question, the impact of behaviour incentives on high-stakes test score outcomes. We then examine whether we can successfully identify students who differentially benefit from treatment. Finally, we discuss two potential mechanisms for the differential effects: the impact on the incentivized behaviours themselves and the characteristics of students who experience the largest impacts.

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

Table 2 reports estimated treatment effects on 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).[31] 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 statemented notice of special educational needs (SEN), month of birth, and the baseline score in the relevant

[30] Simply dropping the two non-complying treatment schools in fact strengthens the main results (available upon request).
[31] Regressions including 2, 3 and 4 years of prior data yield similar results (available upon request).

subject.[32] We find no statistically significant impact of either intervention.[33] 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.

In Tables 3A and 3B we split the sample by gender and by race/ethnicity, which is the covariate we used to generate the matched triplets in the randomization. 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.[34] Looking across subjects, 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.

Predicting treatment effects

Next we turn to using a richer set of observable characteristics to predict which students will experience the largest treatment effects. To do this, we 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.[35]

We use this specification to generate estimates of predicted treatment effects at the individual level using a leave-one-out approach. For each student, we separately estimate the regression described above, excluding the individual whose treatment effects we are aiming to predict. We calculate a student's estimated treatment effect under financial incentives by applying the coefficients for the financial incentive and each financial incentive interaction effect to an individual's own characteristics. We repeat the procedure for the non-financial incentive to

[32] 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.

[33] 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

[34] 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.

[35] 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. We choose 'London or not', and 'Academy or not' out of many possible school characteristics as the most interesting for policy.

generate an individual's predicted treatment effect under non-financial incentives. We calculate the predicted treatment effect separately for each subject. 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. Appendix Table 4 presents the coefficient estimates when we include the full sample (i.e., without leaving-one-out).

In Table 5, we split the sample by above and below median predicted treatment effects, separately for financial and non-financial incentives and by subject. We then estimate the treatment effects for those with "Predicted High" (above median) effects and those with "Predicted Low" (below median) effects. Because we generated our predictions using a leave-one-out approach, this is essentially an out-of-sample test of our predicted effects. Odd numbered columns present estimates for the Financial Incentive treatment and even-numbered columns present estimates for the Non-financial Incentive treatment. Each regression includes non-treated students and treated students for the relevant incentive (i.e., students who received non-financial incentives are excluded from the Financial Incentive regression and vice versa). The omitted group is non-treated students with Predicted Low treatment effects for the relevant incentive and subject. We estimate coefficients for the Predicted High treatment effects subgroup and the interaction effects of treatment with the Predicted High and Predicted Low subgroups. Therefore, our estimate of the treatment effects among Predicted Low students is the coefficient on "Predicted Low*Treated." Our estimate of the treatment effects among Predicted High students is the difference between the coefficients for "Predicted High" and "Predicted High*Treated" – the row below the regression estimates reports the p-value for the test of equality of these coefficients.

For math GCSE scores, the interaction of treatment with Predicted High is large and significant for both the financial incentive and non-financial incentive treatment ranging from 0.12-0.14 standard deviations. The point estimates for the uninteracted 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.

Much of the literature finds that educational interventions have larger effects on math than on English achievement (see for example, Decker et al., 2004; Rockoff, 2004; Jacob, 2005; Dobbie and Fryer, 2011; Levitt et al., 2012). The most likely explanation for this result is that math scores are more sensitive to effort 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.

In Table 5 we analyse 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 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.[36] This is a substantial effect, comparable to the +7 ppt of being female, and the -11 ppt effect of being poor (coefficients from the same regression but not reported).

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 because schools often require interventions to be implemented for all of their students. 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

[36] 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 5). Similar to the results in Table 2, there is no impact on overall GCSE performance in the full population (results available upon request).

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 6 suggest that there is 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 to 0.15 standard deviations.

We can apply our estimated profile of "high effectiveness students" (i.e., above median predicted treatment effects) to the national data to estimate how many such students would be eligible for targeting, and what the scope is for targeting at the school-level.[37] These estimates need to be treated with caution as we are extrapolating from a sample of schools in the poorest neighborhoods. About half of the national cohort are above the experimental sample median for predicted treatment effects, 280k students out of about 550k. There are 1300 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. Across all high poverty schools, about half of students are estimated to be high effectiveness.

Mechanisms – the role of behaviors and student characteristics

Finally, we 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). As discussed in Section 2, teachers were primarily responsible for evaluating students' behaviors (the exception is attendance). Because we randomized treatment at the school level, it is possible that teachers in treatment schools could "game" the evaluation in some way – grading students more generously, giving them fewer assignments, etc. However, we think it is less likely that this kind of gaming would occur differentially at the individual level

[37] 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.

– i.e., the students we predict to have high treatment effects compared to those predicted to have low treatment effects.

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. As in table 5, we estimate the impact of incentives split by predicted treatment effects using the predicted effects on GCSE math scores.[38] Students with Predicted High effects for the non-financial incentive demonstrate significantly higher treatment effects on completing class-work, and also on overall behaviour and homework. 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. However, to the extent these estimates reflect the true impact of the intervention of student behaviour, 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 behavioural 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, an increase in effort by highly motivated, high performing students may have less of impact on exam scores than an increase in effort by under motivated, low-performing students.

To explore this hypothesis, we examine the characteristics of the students who experience the largest treatment effects under Financial and Non-financial Incentives. Table 8 presents the baseline characteristics of students predicted to have High vs. Low treatment effects, along with predicted math GCSE score, predicted behavior, and rewards received in the program (the latter among the subgroup of students assigned to the relevant treatment).[39] The most striking finding is that for both Financial and Non-financial incentives, students predicted to

[38] There are no treatment effects on behaviors in the full population (results available upon request).

[39] We estimated coefficients for predicted GCSE scores using the specification in Eq. 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 using the specification in Eq. 1 and students in the control group only.

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 SD among these students would close about half of the predicted attainment gap. This is also illustrated in Figure 2, which plots actual GCSE scores against predicted scores for both treatment and control groups, separately for each subject and both treatments.

Turning to demographic characteristics, we find different patterns for Non-Financial and Financial Incentives. 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 or speak English as an Additional Language (EAL), which is a proxy for immigrant status and has been shown in previous research to have a high positive correlation with motivation and performance (Burgess et al., 2009).[40] 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 shown in Table 7, these students are those whose effort and behaviours significantly improve in response to the non-financial rewards.

5. Conclusion

We report results from a large field experiment with over 60 schools and 10,000 students where we randomized incentives for increases in inputs to the education production function. A novel part of our experiment is that we were able to tie the impact of such incentives to extremely high-stake assessments taken by all students. The objective of the incentives was to raise pupils' effort and engagement in schooling inputs, and thereby increase their performance on the high-stakes assessment. We implemented two systems of incentives: a financial treatment that rewarded pupils with cash, and a non-financial incentive that offered high-value trips of their own choosing for successful students.

[40] 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 blacks students mirrors the overall sample with about 47% classified as EAL.

The overall impact of the incentives on achievement is low, with small, positive but insignificant effects on exam performance. However, we use the large sample and rich dataset to investigate the distribution of treatment effects, and identify a "right tail": we show that half of the students have economically meaningful positive effects. Students with above-median predicted effects show very substantial and statistically significant effects: we estimate that exam scores improve by 10% to 20% of a standard deviation and that pass rates increase by 8 to 10 percentage points. These effect sizes are similar to the impact of a one standard deviation improvement in teacher quality (Chetty et al., 2014b; Slater et al., 2012). The economic impacts could also be large because of the high estimated earnings rate of return for passing the benchmark of 5 good GCSEs (Battistin et al., 2012; McIntosh, 2004).[41] The treatment effects are stronger for the financial incentives than the non-financial incentives (particularly in science) but not dramatically so. Given their low cost and the ease of administration, our results suggest that non-financial rewards provide a feasible and cost-effective alternative to financial incentives.

Our results suggest that targeting incentives to individuals with high predicted treatment effects would help achieve the policy goal of closing achievement gaps. In our sample, the estimated impact of the incentives would close about half of the predicted attainment gap. Targeting incentives is also a promising approach in contexts where there are concerns that extrinsic incentives can crowd out intrinsic motivation. If there are some students who are highly motivated at baseline then providing incentives likely has little (or potentially negative) impact on performance. Targeting avoids these students while focusing on the subgroup of underperforming students who have little motivation at baseline – and therefore are less susceptible to crowd out. For these students, incentives can have a significant impact on performance, which can help close achievement gaps.

More generally, our study demonstrates that there should be greater attention to examining not only the average impact of an intervention in a field experiment, but also its distributional effects as well. In educational interventions in particular, it is important to recognize that while there may be little effect in the overall population, there may be a significant

[41] 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 (2004) finds a 27-29% return for 5 good GCSE passes using different data.

subpopulation of students who experience meaningful benefits.[42] A better understanding of how to target educational interventions will improve the efficiency of spending on social programs and help craft policies that meet the needs of individual students.

[42] For example, in a review of 77 randomized controlled trial evaluations commissioned by the U.S. Institute for Education Sciences (IES) between 2002-2013, 9% of interventions demonstrated statistically significant positive average treatment effects (Institute of Education Sciences, 2013).

References

Allcott, Hunt, and Judd B. Kessler. *The welfare effects of nudges: A case study of energy use social comparisons*. No. w21671. National Bureau of Economic Research, 2015.

Angrist, Joshua D., Eric Bettinger, and Michael Kremer. 2006. ``Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review*, 96(3): 847-862.

Angrist, Joshua D., Daniel Lang, and Philip Oreopoulos. 2009. ``Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics*, 1(1): 136-163.

Angrist, Joshua D., and Victor Lavy. 2009. ``The Effect of High-Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review*, 99(4): 1384-1414.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. ``Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review*, 99(1): 544--555.

Ashraf, Nava, Oriana Bandiera, and Kelsey Jack. 2012. ``No margin, no mission? A field experiment on the role of incentives in the distribution of public goods." Working Paper.

Athey, Susan, and Guido W. Imbens. "Machine learning methods for estimating heterogeneous causal effects." *stat* 1050 (2015): 5.

Austen-Smith, David and Roland G. Fryer Jr. (2005) An Economic Analysis of "Acting White" *Quarterly Journal of Economics* Vol 119(2) Pp. 551 – 582

Ball, Sheryl, Cathrine C. Eckel, Philip J. Grossman, and William Zame. 2001. ``Status in Markets." *The Quarterly Journal of Economics*, 116(1): 161-188.

Barankay, Iwan. 2011. ``Rankings and Social Tournaments: Evidence from a Crowd-Sourcing Experiment." Working Paper.

Barrera-Osorio, Felipe, Marianne Bertrand, Leigh L. Linden, and Francisco Perez-Calle. 2011. ``Improving the Design of Conditional Transfer Programs: Evidence from Randomized Education Experiment in Colombia." *American Economic Journal: Applied Economics*, 3:167-195

Barrow, Lisa, Lashawn Richburg-Hayes, Cecilia Elena Rouse, and Thomas Brock. 2014."Paying for Performance: The Education Impacts of a Community College Scholarship Program for Low-income Adults," *Journal of Labor Economics,* forthcoming July 2014.

Barrow, Lisa, and Cecilia Elena Rouse. 2013. ``Financial Incentives and Educational Investment: the Impact of Performance-Based Scholarships on Student Time Use," NBER Working Paper No. 19351.

Basit, T. (2012) 'My parents have stressed that since I was a kid': Young minority ethnic British citizens and the phenomenon of aspirational capital. *Education, Citizenship and Social Justice* July vol. 7 no. 2 129-143

Battistin, E. De Nadai, M. and Sianese, B. (2012) Misreported Schooling, Multiple Measures and Returns to Educational Qualifications IZA DP No. 6337

Becker, Gary. 1981 (Enl. ed. 1991). A Treatise on the Family. Cambridge, MA: Harvard University Press.

Behrman, Jere R., Piyali Sengupta, and Petra Todd. 2005. ``Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico. *Economic Development and Cultural Change*, 54(1): 237-275.

Benabou, Roland and Jean Tirole. 2006. ``Incentives and Prosocial Behavior." *American Economic Review*, 96(5): 1652-1678.

Berry, James. 2009. ``Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India." Working Paper.

Bettinger, Eric. 2010. ``Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." NBER Working Paper 16333.

Blanes i Vidal, Jordi and Mareike Nossol. 2011. ``Tournaments Without Prizes: Evidence from Personnel Records." *Management Science*, 57(10): 1721-1736.

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non. 2013. ``Employee Recognition and Performance -- A Field Experiment." CESifo Working Paper Series No. 4164

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4): 200–232.

Burgess, S. (2014) Understanding the Success of London's Schools. CMPO WP 2014/333, CMPO University of Bristol.

Burgess, S., Wilson, D. and Piebalga, A. (2009) Land of hope and dreams: education aspirations and parental influence among England's ethnic minorities. Mimeo, CMPO

Bursztyn, L. and Jensen, R. (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics*, 130(3):1329-1367.

Carpenter, H., Ivy Papps, Jo Bragg, Alan Dyson, Diane Harris & Kirstin Kerr, Liz Todd and Karen Laing (2013) Evaluation of Pupil Premium. DFE Research Report DFE-RR282

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." *The American Economic Review* 104.9 (2014): 2593-2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *The American Economic Review* 104.9 (2014): 2633-2679.

Curto, Vilsa and Roland Fryer (2014) The Potential of Urban Boarding Schools for the Poor: Evidence from SEED' *Journal of Labor Economics* vol. 32(1) pp. 65 – 94.

Dearden, Lorraine, Carl Emmerson, Christine Frayne, and Costas Meghir. "Conditional cash transfers and school dropout rates." *Journal of Human Resources* 44, no. 4 (2009): 827-857.

Decker, Paul T., Daniel P. Mayer, and Steven Galzerman. 2004. ``The Effects of Teach for America on Students: Findings from a National Evaluation." Mathematica Policy Research Report No. 8792-750.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. ``Testing for Altruism and Social Pressure in Charitable Giving." *The Quarterly Journal of Economics*, 127(1): 1-56.

Dobbie, Will and Roland G. Fryer. 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.

Dolton, P.J. and Vignoles, A. (2002) The Return on Post-Compulsory School Mathematics Study *Economica*, Vol. 69, No. 273, pp. 113-141

Dynarski, Susan. 2002. "The Behavioral and Distributional Implications for Aid for College." *American Economic Review*, 92(2): 279-285.

Frey, Bruno S. 2007. ``Awards as Compensation." *European Management Review*, 4(1): 6-14.

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

Fryer, Roland G. 2012. "Aligning Student, Parent and Teacher Incentives: Evidence from Houston Public Schools." NBER Working Paper 17752.

Gerber, Alan S. and Donald P. Green (2012) Field Experiments: Design, Analysis and Interpretation. W.W. Norton and Company, New York.

Hanushek, E. A. (2009). The economic value of education and cognitive skills. In: Sykes, G., Schneider, B., Plank, D.N. (Eds.), *Handbook of Education Policy Research*, 39-56.

Hanushek EA. (2011) The economic value of higher teacher quality. *Economics of Education Review* vol(30) pp. 466–79

Hanushek, E. and Woessmann, L. (2015) The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving Pisa Outcomes. OECD, Paris.

Huberman, Bernardo A., Christoph H. Loch, and Ayse Onculer. 2004. ``Status As a Valued Resource." *Social Psychology Quarterly*, 67(1): 103-114.

Imai, Kosuke, and Marc Ratkovic. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7.1 (2013): 443-470.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, 47(1): 5–86.

Institute of Education Sciences. 2013. "Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects", Available at http://coalition4evidence.org/wp-content/uploads/2013/06/IES-Commissioned-RCTs-positive-vs-weak-or-null-findings-7-2013.pdf.

Jackson, C. Kirabo. 2010. ``A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program." *Journal of Human Resources*, 45(3): 591-639.

Jacob, Brian. 2005. ``Accountability, Incentives and Behavior: Evidence from School Reform in Chicago." *Journal of Public Economics*, 89(5-6): 761-796.

Kosfeld, Michael and Susanne Neckermann. 2011. ``Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics*, 3(3): 86-99.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. ``Incentives to Learn." *The Review of Economics and Statistics*, 91(3): 437-456.

Leuven, Edwin, Hessel Oosterbeek, and Bas van der Klaauw. 2010. ``The Effect of Financial Rewards on Students' Achievements: Evidence from a Randomized Experiment." *Journal of the European Economic Association*, 8(6): 1243-1265.

Levitt, Steven D., John A. List, and Sally Sadoff. 2010. ``The Effect of Performance-Based Incentives on Educational Achievement: Evidence from a Randomized Experiment." Working Paper.

Levitt, Steven D., John A. List, and Sally Sadoff. 2012. ``The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." NBER Working Paper 18165.

Loewenstein, George and Samuel Issacharoff. 1994. ``Source Dependence in the Valuation of Objects." *Journal of Behavioral Decision Making*, 7(3): 157-168.

McIntosh, S. (2004) Further Analysis of the Returns to Academic and Vocational Qualifications. CEE DP 35, LSE.

Metcalfe, Robert, Burgess Simon and Stephen Proud (2011) Student effort and educational attainment: Using the England football team to identify the education production function CMPO WP 11/276. CMPO

Michelmore, Katherine and Susan Dynarski. (2016). "The Gap within the Gap: Using Longitudinal Data to Understand Income Differences in Student Achievement," NBER Working Paper No. 22474.

OFSTED (2013) The Pupil Premium: how schools are spending the funding successfully to maximise achievement. http://www.ofsted.gov.uk/resources/pupil-premium-how-schools-are-spending-funding-successfully-maximise-achievement.

Oreopoulos, Philip. "Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling." *Journal of Public Economics* 91.11 (2007): 2213-2229.

Oreopoulos, Philip, and Kjell G. Salvanes. "Priceless: The Nonpecuniary Benefits of Schooling." *Journal of Economic Perspectives* 25.1 (2011): 159-184.

Rockoff, Jonah E. 2004. ``The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review*, 94(2): 247-252.

Schultz, T. Paul. 2004. ``School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program." *Journal of Development Economics*, 74(1): 199-250.

Sharma, Dhiraj. 2010. "The Impact of Financial Incentives on Academic Achievement and Household Behavior: Evidence from a Randomized Trial." Manuscript.

Shah, B., Dwyer, C. and Modood, T. (2010) Explaining educational achievement and career aspirations among young British Pakistani Women: Mobilising 'ethnic capital'? *Sociology*, 44(6) 1109-1127.

Slater, Davies and Burgess (2012) Do Teachers Matter? Measuring the Variation in Teacher Effectiveness in England. *Oxford Bulletin of Economics and Statistics*, vol. 74, issue 5, pages 629-645.

Tran, Anh and Richard Zeckhauser. 2012. ``Rank as an Inherent Incentive: Evidence from a Field Experiment." *Journal of Public Economics*, 96(9-10): 645–650.

Watkins, M. and Noble, G. (2013) Disposed to Learn: Schooling, Ethnicity and the Scholarly Habitus. London: Bloomsbury.

Wilson, Deborah, Simon Burgess and Adam Briggs (2011) "The dynamics of school attainment of England's ethnic minorities" *Journal of Population Economics*, 24(2): 681-700.

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)
Pupil eligible for Free School Meals (FSM)	0.387	0.374	0.441
	(0.487)	(0.484)	(0.497)
Pupil speaks English as an additional language (EAL)	0.478	0.535	0.491
	(0.500)	(0.499)	(0.500)
Pupil has statemented 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) Maths 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

The table reports means and standard deviations of each variable. There are no statistical differences between the control group and either treatment group at the 10/5/1 percent levels. The statistical tests were conducted using robust standard errors, clustered on the school level.

Table 2. Effects of incentives on high stakes grades

	Math		English		Science	
	1	2	3	4	5	6
Financial Incentive	-0.006	0.053	-0.072	-0.001	-0.049	0.082
	(0.074)	(0.047)	(0.099)	(0.052)	(0.113)	(0.067)
Non-Financial Incentive	0.091	0.014	-0.012	0.029	0.059	0.054
	(0.064)	(0.034)	(0.078)	(0.060)	(0.106)	(0.053)
Pupil Chars?	Yes	Yes	Yes	Yes	Yes	Yes
School FE?	No	Yes	No	Yes	No	Yes
p-value Financial = Non-Financial	0.244	0.396	0.594	0.592	0.332	0.630
R^2	0.215	0.208	0.244	0.235	0.198	0.170
Observations	9827	20058	9750	19791	10238	20713
Schools	63	63	63	63	63	63

- Dependent variable is the pupil's test score. Test scores are normalised at pupil level (over the whole national cohort year-by-year).
- The sample is pupils in 63 schools, in just the treatment year in odd-numbered columns and the treatment and prior year in even-numbered columns.
- Estimation is by OLS, with school fixed effects in even-numbered columns.
- Variables included in pupil characteristics are gender, poverty status, broad ethnic groups, whether language spoken at home is English, subject-specific prior attainment, presence of special education needs, month of birth, and year dummies.
- Standard errors are clustered at school level. Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 3A. Effects of incentives on high stakes grades: By gender

	Math		English		Science	
	Female	Male	Female	Male	Female	Male
Financial Incentive	0.028	0.076	-0.002	-0.011	0.099	0.062
	(0.054)	(0.056)	(0.059)	(0.057)	(0.077)	(0.076)
Event Incentive	0.041	0.008	0.104	-0.027	0.087	0.035
	(0.058)	(0.040)	(0.067)	(0.072)	(0.066)	(0.062)
Pupil Chars	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.206	0.213	0.217	0.228	0.171	0.173
Observations	10457	9601	10455	9336	10790	9923
Schools	60	55	60	55	60	55

Table 3B. Effects of incentives on high stakes grades: By race/ethnicity

	Math			English			Science		
	Asian	Black	White	Asian	Black	White	Asian	Black	White
Financial Incentive	-0.070	0.169**	0.072	-0.025	0.067	-0.037	0.063	0.126	0.081
	(0.059)	(0.066)	(0.082)	(0.071)	(0.059)	(0.082)	(0.111)	(0.098)	(0.070)
Event Incentive	0.013	0.028	0.044	0.024	0.056	0.041	0.044	-0.043	0.143**
	(0.061)	(0.051)	(0.054)	(0.079)	(0.064)	(0.093)	(0.061)	(0.092)	(0.069)
Pupil Chars	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.231	0.160	0.227	0.286	0.186	0.246	0.202	0.130	0.174
Observations	5073	5008	8379	5054	4975	8228	5165	5162	8742
Schools	60	62	63	60	62	63	60	62	63

- Dependent variable is the pupil's test score. Test scores are normalised at pupil level (over the whole national cohort year-by-year).
- The sample is pupils in the number of schools shown over the treatment and the prior year.
- Estimation is by fixed effects OLS.

- Variables included in pupil characteristics are gender, poverty status, broad ethnic groups, whether language spoken at home is English, subject-specific prior attainment, presence of special education needs, month of birth, and year dummies.
- Standard errors are clustered at school level. Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 4. Effects of incentives on high stakes grades: By 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	-0.083	-0.022	0.075	-0.082*	0.062*
	(0.063)	(0.055)	(0.044)	(0.048)	(0.043)	(0.034)
Predicted High * Treated	0.135**	0.121*	-0.002	0.079	0.117	0.081
	(0.054)	(0.069)	(0.048)	(0.064)	(0.082)	(0.057)
Predicted Low * Treated	-0.024	-0.050	-0.002	-0.027	0.053	0.031
	(0.051)	(0.038)	(0.066)	(0.058)	(0.065)	(0.060)
Pupil Chars	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
R^2	0.206	0.213	0.217	0.228	0.171	0.173
Observations	14947	15551	14773	15359	15386	16069
Schools	48	48	48	48	48	48

- Dependent variable is the pupil's test score. Test scores are normalised at pupil level (over the whole national cohort year-by-year).
- The sample is pupils in 15 treatment schools and 33 control schools, over the treatment year and the prior year.
- Estimation is by fixed effects OLS.
- Predicted treatment effects estimated on LOO basis as function of pupil characteristics.
- Variables included in pupil characteristics are gender, poverty status, broad ethnic groups, whether language spoken at home is English, subject-specific prior attainment, presence of special education needs, month of birth, and year dummies.
- Standard errors are clustered at school level. Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 5. Effects of incentives on overall high stakes 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	-0.067	-0.052*	-0.048**
	(0.044)	(0.046)	(0.030)	(0.024)
Predicted High * Treated	0.082*	0.152**	0.050*	0.031
	(0.043)	(0.062)	(0.029)	(0.037)
Predicted Low * Treated	-0.025	0.029	-0.002	-0.019
	(0.032)	(0.085)	(0.038)	(0.034)
Pupil Chars	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
R^2	0.149	0.162	0.151	0.160
Observations	14744	15373	14744	15373
Schools	63	63	63	63

- Dependent variable is the pupil's test score. Test scores are normalised at pupil level (over the whole national cohort year-by-year).
- The sample is pupils in 15 treatment schools and 33 control schools, over the treatment year and the prior year.
- Estimation is by fixed effects OLS.
- Predicted treatment effects estimated on LOO basis as function of pupil characteristics.
- Variables included in pupil characteristics are gender, poverty status, broad ethnic groups, whether language spoken at home is English, subject-specific prior attainment, presence of special education needs, month of birth, and year dummies.
- Standard errors are clustered at school level. Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 6: School-level designation, pupil level analysis, high predicted effectiveness schools only

	Math		English		Science	
	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive
Financial Incentive	0.106		0.056		0.131	
	(0.075)		(0.034)		(0.099)	
Event Incentive		0.098*		0.077		0.145*
		(0.048)		(0.083)		(0.080)
Pupil Chars	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.204	0.230	0.220	0.242	0.164	0.154
Observations	6949	7548	6611	8836	7636	7870
Schools	27	26	27	27	22	27

- Dependent variable is the pupil's test score. Test scores are normalised at pupil level (over the whole national cohort year-by-year).
- The sample is pupils in schools with greater than median fractions of high predicted effectiveness pupils, over the treatment year and the prior year.
- Estimation is by fixed effects OLS.
- Predicted treatment effects estimated on LOO basis as function of pupil characteristics.
- Variables included in pupil characteristics are gender, poverty status, broad ethnic groups, whether language spoken at home is English, subject-specific prior attainment, presence of special education needs, month of birth, and year dummies.
- Standard errors are clustered at school level. Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 7: Effects of incentives on Behaviours: By predicted treatment effects

Financial Incentive	Attendance	Conduct	Classwork	Homework	Overall
Predicted High *	0.068	-0.021	0.424**	0.053	0.176
Treated	(0.142)	(0.145)	(0.203)	(0.216)	(0.168)
Predicted Low * Treated	0.302	0.084	0.294	0.052	0.238
	(0.241)	(0.162)	(0.211)	(0.191)	(0.218)
Predicted High	0.218	-0.040	-0.077	0.001	0.032
	(0.185)	(0.145)	(0.182)	(0.159)	(0.166)
Pupil Chars?	Yes	Yes	Yes	Yes	Yes
p-value: Predicted High = Predicted High * Treated	0.842	0.707	0.064	0.455	0.222
R^2	0.051	0.055	0.122	0.083	0.084
Observations	4476	4476	4476	4476	4476
Schools	30	30	30	30	30

Non- financial Incentive	Attendance	Conduct	Classwork	Homework	Overall
Predicted High *	0.010	0.247	0.515**	0.344*	0.369*
Treated	(0.185)	(0.165)	(0.225)	(0.198)	(0.191)
Predicted Low * Treated	0.226	0.188	0.072	-0.025	0.143
	(0.221)	(0.203)	(0.189)	(0.203)	(0.223)
Predicted High	0.060	0.082	-0.267	-0.012	-0.050
	(0.174)	(0.180)	(0.180)	(0.200)	(0.199)
Pupil Chars?	Yes	Yes	Yes	Yes	Yes
p-value: Predicted High	0.652	0.098	0.014	0.022	0.024

= Predicted High * Treated					
R^2	0.030	0.053	0.098	0.072	0.069
Observations	4716	4716	4716	4716	4716
Schools	30	30	30	30	30

- Dependent variable is the pupil's behaviour score, normalised at pupil level.
- The sample is pupils in 15 treatment schools and 15 control schools over the treatment year.
- Estimation is by OLS.
- Predicted treatment effects estimated on LOO basis as function of pupil characteristics. This uses the Math predicted treatment effect.
- Variables included in pupil characteristics are gender, poverty status, broad ethnic groups, whether language spoken at home is English, subject-specific prior attainment, presence of special education needs, month of birth, and year dummies. Also included a measure of school effectiveness, the school maths fixed effect.
- Standard errors are clustered at school level. Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

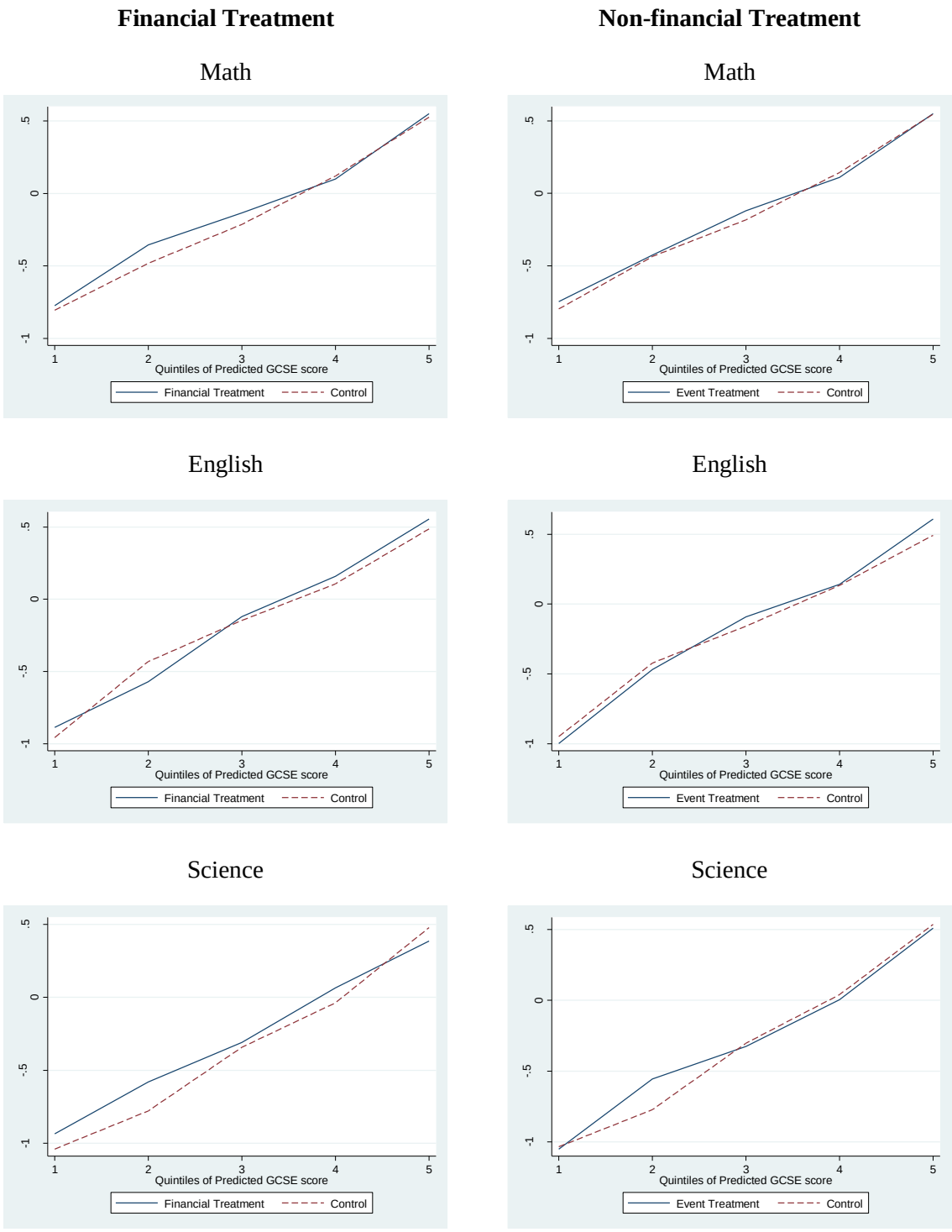
Table 8: Composition of high and low predicted incentive effectiveness

Financial Incentive	Predicted effectiveness		
	High	Low	p-value Hi = Lo
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
Pred'd Maths GCSE	-0.461	0.081	0.00
Pred' Behaviour	-0.199	-0.004	0.00
Financial reward	210.7	238.2	0.00
N	4966	4861	
Event Incentive			
	Predicted effectiveness		
	High	Low	p-value Hi = Lo
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
Pred'd Maths GCSE	-0.397	0.000	0.00
Pred' Behaviour	-0.151	-0.057	0.00
Non-financial reward	11.26	11.26	0.99
N	4773	5054	

These are the groups for maths, computed for all students in 2013, whether T1 or T2 or C group. Baseline attainment is KS2 Maths; Predicted Maths GCSE based on personal characteristics and school attended from estimation using four pre-treatment years; Predicted Behaviour based on personal characteristics from estimation using just control schools. Data for 2013. P-values take account of clustering. Financial rewards are mean actuals in £ (event rewards converted to £).

Figure 1: Location of experimental schools across England

Figure 2: Treatment effects by predicted GCSE score



Vertical axis: actual normalised 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.

Appendix A: The nature of the school system in the England

Compulsory education in England runs from age 4 to 16. School education can continue up to age 18 (roughly half of pupils do this), followed by university or college. School education is split up into 5 Key Stages, each ending with a set of exams. Key stage 1 (KS1) runs from 4 to 7, KS2 from 8 to 11, KS3 from 12 to 14, KS4 15 to 16 and KS5 17 to 18. Typically, a pupil will attend primary school unitl age 11 and then move to a much bigger secondary school until age 16 or 18.

The key exams are the KS4 ones, also known as GCSEs, and the main outcome used here, which act as a gatekeeper to further schooling and are also key qualifications for jobs. In the last few years, KS3 tests have been replaced by teacher assessments so the best prior attainment data before GCSEs is the KS2 tests. These are compulsory for all pupils in all state schools, are taken at the end of primary school at age 11 and are taken in maths, English and science.

There is virtually no grade repetition in England so almost all pupils are in the age-appropriate school year.

In terms of scale, there are about 550k – 600k pupils in a cohort and around 3000 secondary schools in England.

Appendix B: Details of the Treatments

Behaviours

A. Attendance

- **Criteria**: Based on the official definition of attendance – attendance at school, not lessons.
- **Measurement**: Measured twice a day at registration in school
- **Threshold**: No un-authorised absences per half-term

B. Behaviour

- **Criteria**: An instance of poor behaviour is recorded on the basis of the student either (a) arriving late to a lesson (more than 5 minutes late); or (b) exhibiting behaviour in the lesson resulting in a sanction.
- **Measurement**: (a) measured by the teacher each lesson; (b) either measured by the teacher in the class, or recorded centrally by the school
- **Threshold**: No more than one instance per subject per half-term.

C. Class-work/coursework

- **Criteria**: Completion of work on time and at a level consistent with the student's target (at a lower level only with strong acceptable justification).
- **Measurement**: marked and judged by the teacher, and recorded centrally by the school.
- **Threshold**: All work in all three subjects meet this criterion per half-term.

D. Homework

- **Criteria**: Completion of work on time and at a level consistent with the student's target (at a lower level only with strong acceptable justification). This includes participation in any out-of-school learning/homework club as required by the school.
- **Measurement**: marked and judged by the teacher, and recorded centrally by the school.
- **Threshold**: All work in all three subjects meet this criterion per half-term.

Scheme Details

A. Financial incentives

1. Reward

Each student starts with an allocation of £320, and £80 of this is available each of the first four half-terms. Essentially, it is theirs to lose or keep. The rule is that if you miss your threshold as specified above, you lose money as follows: £10 for missing the attendance threshold; £10 for behaviour; £30 for class-work; £30 for homework.

2. Feedback

Each child will receive a letter shortly after each half-term reporting how they did in the preceding half-term on all the four behaviour items. It will also inform them about the outcome for their money. The letter will be addressed to them (not their parents) and posted to their home address.

3. Set up

This is straightforward in this scheme. We still need to finalise the operational details of paying the students, but we are committed to ensuring that it is done very promptly after the end of the relevant half-term.

B. Non-financial incentives

1. Reward

Each student is able to participate in two events per year, in December after the second half-term and again in April after the fourth half term. Participation depends on having retained enough 'tickets'. Again this works by loss aversion: each student starts off with 8 tickets per half-term, and may lose them as follows: 1 for missing the attendance target, 1 for behaviour; 3 for class-work and 3 for homework. So by the end of the two half-terms before Christmas they could have a maximum of 16 tickets. To go on the trip they need 12 tickets. The two rounds are separate and distinct: in January, each student starts off afresh with the same number of tickets for the April trip.

What are the events? We will not prescribe this, and encourage the schools to utilise some forum to let the year 11 students collectively decide at the start of term. This would typically involve a choice of three different events. We will provide the framework, state the budget available and can help to organise it, seek group discounts, and try to involve organisations to donate tickets to things.

2. Feedback

Each child will receive a letter shortly after each half-term reporting how they did in the preceding half-term on all the four behaviour items. It will also inform them about the outcome for their money. The letter will be addressed to them (not their parents) and posted to their home address.

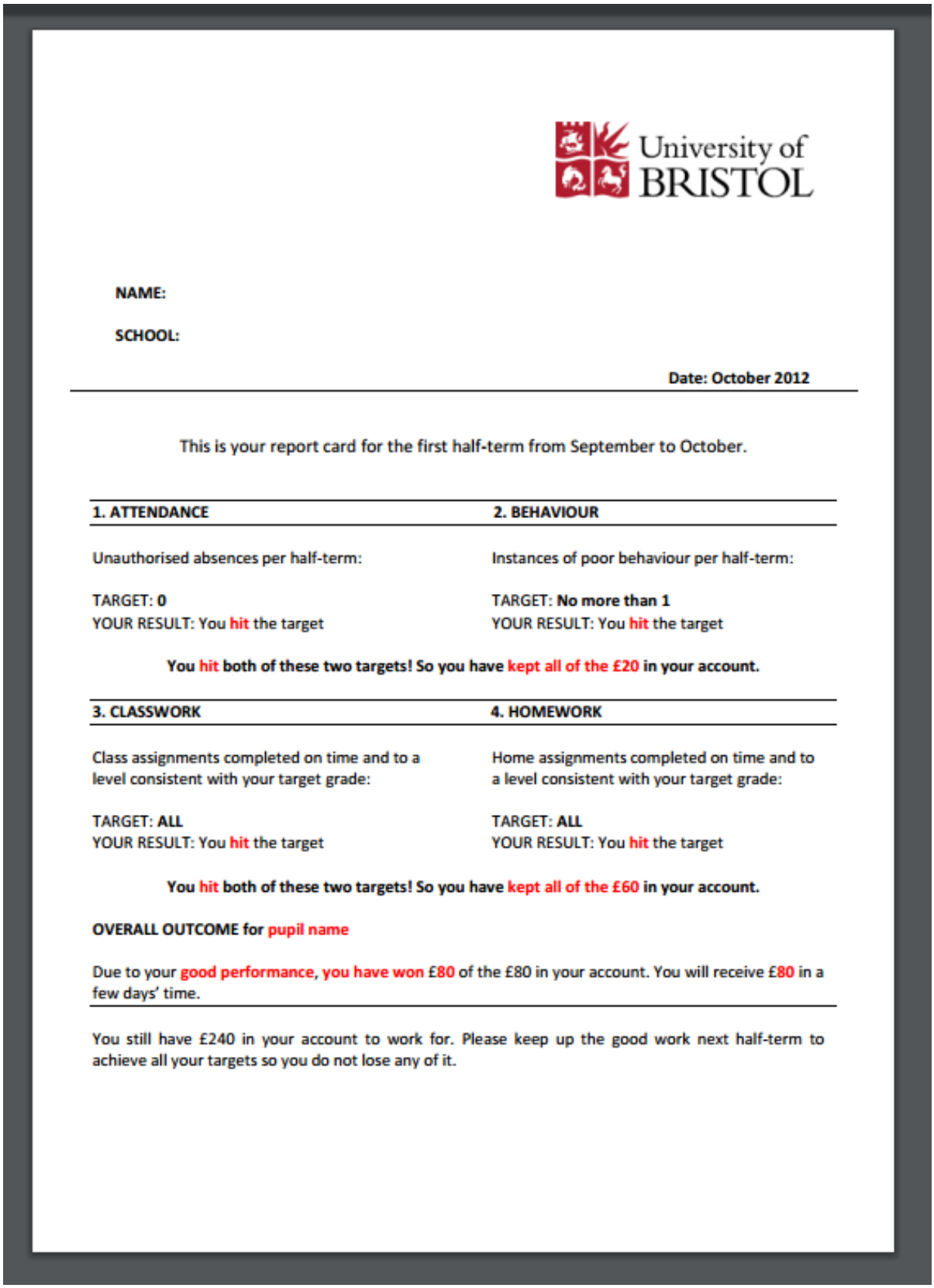
3. Set up

This is straightforward. We can help to organise the booking of the events.

C. Control

The same four items are measured in the same way. There is no incentive payment for the students, and we are not going to send feedback letters to students. We will write a comprehensive report for each school at the end of the year on the outcome of the four items, and how this compares to similar schools, and how it relates to student progress.

Sample Feedback Letter, Financial Treatment:



University of BRISTOL

NAME:

SCHOOL:

Date: October 2012

This is your report card for the first half-term from September to October.

1. ATTENDANCE	**2. BEHAVIOUR**
Unauthorised absences per half-term:	Instances of poor behaviour per half-term:
TARGET: **0** YOUR RESULT: You **hit** the target	TARGET: **No more than 1** YOUR RESULT: You **hit** the target

You hit both of these two targets! So you have kept all of the £20 in your account.

3. CLASSWORK	**4. HOMEWORK**
Class assignments completed on time and to a level consistent with your target grade:	Home assignments completed on time and to a level consistent with your target grade:
TARGET: **ALL** YOUR RESULT: You **hit** the target	TARGET: **ALL** YOUR RESULT: You **hit** the target

You hit both of these two targets! So you have kept all of the £60 in your account.

OVERALL OUTCOME for pupil name

Due to your **good performance**, **you have won £80** of the £80 in your account. You will receive **£80** in a few days' time.

You still have £240 in your account to work for. Please keep up the good work next half-term to achieve all your targets so you do not lose any of it.

Appendix C: Recruitment and communications

Key features and rationale for adopting the recruitment policy

We sought expert advice from consultants with senior leadership positions in the schools' sector, incuding lead inspectors (Ofsted), former head teachers, former education advisors. We adopted this approach to enhance our prospects of securing conversations with School Leadership Teams, and also helped us to understand much more intimately the pressures on schools, and identify key strategies that were likely to secure sign up and engagement with the project.

Approach to schools, including piloting

Two formal meetings between the two liaison consultants and members of the University of Bristol project team were held (i) to clarify the project to the consultant team and to discuss school contexts relevant to the project; and (ii) to refine practical aspects of the application of the project methodology in schools. The liaison consultants were used as professional sounding boards to support the communications strategy, refine transposition of method into practice, and to train schools. The project manager coordinated communication strategy and dissemination of recruitment materials to schools with the lead consultant, who, in turn, coordinated approaches to schools across the wider team.

A pilot of the communications strategy and guidance , leading to minor adjustments, was conducted with six schools by each liaison consultant to trial the guidance and template. "Cold calling" was avoided by sending a letter of invitation and FAQ sheet by post and email to each school just prior to each telephone contact. A common approach to managing the telephone discussion and the main messages to be communicated to schools was agreed.

Weekly updating summaries were provided by each consultant to support the project manager in tracking outcomes (i.e. the response of each school with which contact was made that week) and costs (i.e. hours worked aligned with activities undertaken).

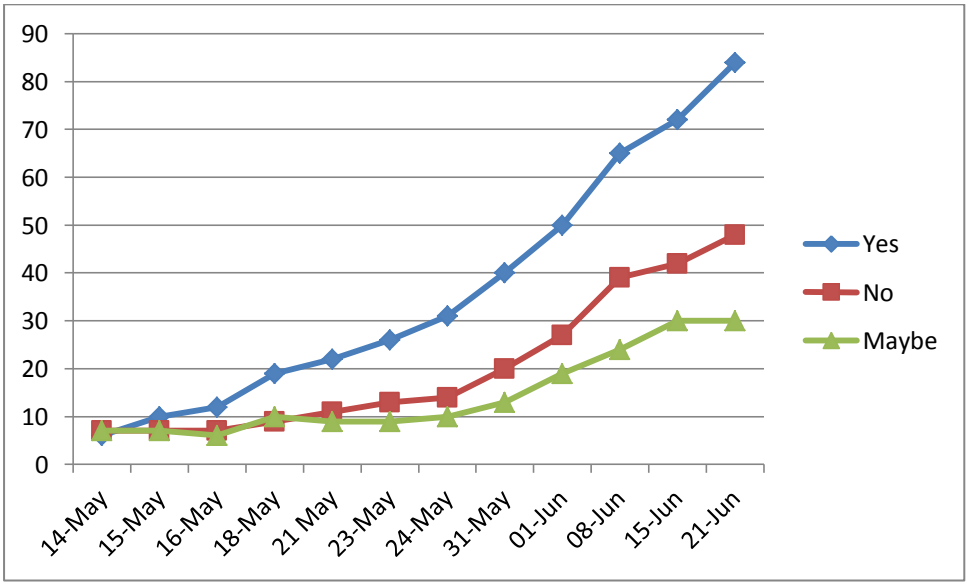


Figure A1. Recruitment trends

This figure demonstrates recruitment to the point where it was agreed we should cease recruitment because of potential cost implications.

Consent

Ethical clearance was granted by the University of Bristol's Faculty of Social Sciences and Law Ethics Committee. This process covered all data protection and safeguarding matters. Consent was secured at school level, to assure agreement to access NPD data, and parental level through reverse consent. The only real issue we experienced with adopting an approach of reverse consent was where parents failed to read the form correctly, and assumed it was an opt-in. We were first alerted to this by a parent calling the Project Team direct. As a result of that, the project team called schools to ask them to confirm with the parent that they really wanted their child opted out of the programme before removing these pupils from the data sets uploaded to the University.

Information to school liaison

The school liaison was appointed by the Head Teacher, and was usually, but not exclusively, a member of the Senior Leadership Team. All schools' liaison staff were offered a choice of three training events. In cases where attendance was not possible, the project manager provided face-to-face or telephone training.

Information ready to engage teachers, parents and pupils was despatched to project liaisons in the last week of August, as indicated in a 'Welcome Letter' sent before the end of the summer term. The letter contained the following information:

- sample letter to parents;
- sample letter to pupils;
- login to the University of Bristol data depository;
- supplementary information relating to the use of the data depository;
- updated FAQs to include those arising in training events and recruitment campaign;
- criteria for determining behaviour in line with project methodology;
- Summary details of the scheme as presented at the training day.

The remainder of the shipment contained:

- Sealed envelopes for distribution to parents containing details of the scheme, opt-out form, and freepost envelope to return opt-out, if appropriate.
- Letters to pupils about the scheme

Information to pupils

Each child was given a letter about the scheme, and indicating that a certain amount of money, or points towards an event had been awarded, and what they needed to do to retain that money and the value of loss attributed to each threshold missed. Schools disseminated information about the project, and marketed the scheme to pupils during an assembly in the first week of term; many schools mediated packs to parents during a parents evening, and took the opportunity to encourage them to engage with their child in working hard to retain their rewards.

Outcome reporting system – what measured, reason for choices, instructions to schools, communication/upload methods

The Project Team devised an excel spreadsheet for schools to complete and return to the University, recording behaviour, attendance, coursework/classwork, and homework. Full instructions were provided to the school in the form of: training events, written instructions, screenshots and telephone support from the project team.

Feedback letters

Feedback letters were generated to pupils ready for their return to school following the holiday after the data upload. It contained information about reward retention and loss as the related to each threshold measures; overall retention and loss; a motivational sentence relating to their outcome to encourage effort in the following period. The final letter excluded the motivational sentence.

Supplementary feedback letters

Event treatment pupils were provided with an additional letter at Christmas and at Easter, detailing the overall outcome of their two terms' effort, and indicating whether they were entitled to take part in the event linked to that term's attainment.

Appendix D: Dealing with science qualifications

(*With very many thanks to Dave Thomson of FFT and Claire Crawford and Ellen Greaves of IFS)*

There are a number of different science qualifications that need to be taken into account. These include single science GCSEs (e.g., biology), Double Science GCSEs (Core Science and Applied Science), BTECs in Science, OCR National Awards and Certificate in Science, and Intermediate GNVQs in Science, as well as Vocational GCSE Single Award Additional Applied Science.

There is no single measure summarising performance across the universe of science qualifications. There appears to be no single accepted way of combining these. Many students take a few qualifications in combination: for example, core science plus applied science, or biology, physics and chemistry. Some summary statistics on patterns or taking:

- Separate biology, chemistry and physics -- this accounts for 30%
- Core science and Additional science -- this accounts for 53%
- Just Core Science -- this accounts for 11%

But 23% do no GCSE science at all.

We adopt this choice: Highest point score achieved in Science (using official GCSE "equivalency" values for different qualifications). The potential set of qualifications includes regular full GCSEs, Full Intermediate or Foundation level General National Vocational Qualification (GNVQ) or Vocational GCSE. The variable is called "KS4_KS4SCI" in the National Pupil Database (NPD).

Appendix E: Randomisation

Method

The randomisation was at school-cohort level for the year 11 cohort in schools. We chose to adapt pair-wise matching as our method for randomisation. The simulation results of Bruhn and McKenzie (2009) support this in the case of small samples, which is relevant given our 63 schools; see also the argument in Imbens and Wooldridge (2009). We had three groups to randomise to: two treatments and a control, so we use a procedure to generate matched triples.

Process

The task was to assign 63 confirmed schools to three groups: 15 schools in T1 (financial reward), 15 schools in T2 (event reward) and the remaining 33 schools in C (monitoring only treatment, acting as control). We placed schools into matched triplets before assigning treatments. A triplet is a group of schools with similar test scores, majority ethnic group and the presence of an existing reward structure in the school (see below); if this latter information was unknown, we assumed the school had no reward structure. We randomly chose fifteen of these triplets to be "A" type schools and the rest were designated "B" type schools. "A" type schools had one member assigned to each treatment (control, financial reward, or event reward). We placed all randomly chosen "B" type schools into the control group.

With a sample size of just 63 schools, there is not a great deal of scope for blocking. We wanted to pick blocking variables to create groups in which the heterogeneity is minimised within the groups, relative to between groups, and to minimise attrition. One pragmatic factor in choosing the blocking variables is the sample sizes within each split defined by those variables. We needed reasonable sizes in each split, and needed all to contain at least 3 schools. Also, if the sample size was not exactly divisible by 3, all those in the remainder are necessarily assigned to C. So organising the splits to minimise the remainders was important.

We considered the following candidates for blocking: whether the school is in London; whether the school is an Academy; quantiles of the school's size; measures of ethnic composition; whether the school is a single sex school; whether the school has its own reward scheme in place already; quantiles of school performance. (We also considered the following but there was no variation: all schools are in urban areas; no schools are grammar schools; no schools have boarders).

We believed that the most important factors were whether the school already had its own scheme, whether it was in London, school size, school performance and ethnic mix, and experimented with different combinations of these. We focussed on the presence of an existing scheme because our own intervention was layered on top of what is already there. We gathered this information from a brief web-based questionnaire given to schools in July, just before the end of term. Of the 63 schools, 33 had replied by the time of the randomisation and of those 18 did have a scheme, all bar one of those being 'trips or trophies' (we grouped the missings with the "no"s). We considered London because of the substantial differential progress made by London secondary schools in recent years, possibly deriving from the 'London Challenge' policy. Size and performance are likely to be important, though we can control for differences in these in very flexible ways. It is well established that different ethnic groups in England have very different school performance and trajectories (see Wilson, Burgess and Briggs, 2011). Ethnicity has also been regularly used as a blocking variable in US studies, and while acknowledging that ethnicity places different roles in education in the US and England, the known differences in England mean that this is likely to be an important source of difference. Most

problematic were single sex schools, with 3 all-boys schools and 8 all-girls schools. There was no practical way of blocking on this and any other variable that yielded cell sizes of at least 3.

We chose to block on the presence of the school's own reward scheme and which ethnic broad group is the majority group in the school (Asian (ie Bangladeshi, Indian or Pakistani ethnicity), Black (Black African or Caribbean), White). The programme blocks on these and sorts on a random variable and school performance. The resulting cell sizes were as follows:

Blocking variables		No. of schools			
Largest broad ethnic group	Own reward scheme?		Number assigned to Financial treatment	Number assigned to Event treatment	Number assigned to control
White	No or missing	22	5	6	11
Asian	No or missing	11	3	3	5
Black	No or missing	12	3	2	7
White	Yes	9	2	1	6
Asian	Yes	5	1	1	3
Black	Yes	4	1	2	1

We adopted a simple system for re-randomisation. A random allocation of schools to treatment status was proposed and the balance checked. If any variables were significantly unbalanced at 10%, then the allocation was rejected and another was drawn. This procedure continued until an allocation was not rejected.

Appendix Table 1: Representativeness of the sample

	Coefficient	t-statistic
Proportion female students	0.129	(1.11)
Proportion white students	0.163	(0.38)
Proportion Black students	0.248	(0.53)
Proportion Asian students	0.256	(0.74)
School mean IDACI score	-0.547	(0.93)
Proportion students eligible for FSM	0.200	(0.59)
Proportion students with SEN	0.125	(0.29)
Proportion students speaking EAL	-0.153	(0.53)
School %5A*-C(EM)	0.047	(0.15)
School value added	-0.001	(0.84)
Trend GCSE points	0.000	(0.23)
London	0.136	(1.64)
Single sex	0.013	(0.26)
Academy	-0.007	(0.11)
Number of pupils	0.000	(0.07)
HT newly hired	0.439	(5.53)
Mean KS2 intake	-0.324	(1.46)
Constant	2.145	(1.70)
N		263
F(17, 245)		2.50
p-value		0.001

- Dependent variable is binary, equal to 1 for the 63 participating schools.
- A probit regression shows the same patterns

Appendix Table 2 - Probability of attrition reporting behavioral measures, school level

	Coefficient	t-statistic
Proportion female students	0.206	(0.65)
Proportion white students	-1.671	(1.52)
Proportion Black students	-1.297	(1.06)
Proportion Asian students	-0.793	(0.80)
School mean IDACI score	-1.852	(1.03)
Proportion students eligible for FSM	0.004	(0.43)
Proportion students with SEN	-0.005	(0.45)
Proportion students speaking EAL	-0.013	(1.43)
School %5A*-C(EM)	-0.393	(0.52)
School value added	0.003	(1.43)
Trend GCSE points	0.003	(0.81)
London	0.260	(1.06)
Single sex	0.127	(0.84)
Academy	-0.121	(0.71)
Number of pupils	0.000	(0.21)
HT newly hired	0.056	(0.37)
Mean KS2 intake	-0.695	(1.02)
Constant	2.633	(0.78)
N		63
F(17, 45)		0.95
p-value		0.526

OLS regression of (0, 1) attrited. (Probit shows the same, individually and collectively insignificant)

Appendix Table 3: Effects of incentives on high stakes exam scores: By predicted treatment effects. Intact Triplets only

	Math		English		Science	
	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive
Predicted High	0.059	-0.141	-0.026	0.054	-0.075	-0.059
	(0.160)	(0.133)	(0.093)	(0.088)	(0.183)	(0.060)
Predicted High * Treated	0.181	0.289**	-0.035	0.124	0.148	0.094
	(0.173)	(0.094)	(0.090)	(0.146)	(0.232)	(0.147)
Predicted Low * Treated	0.119	0.066	-0.053	-0.053	0.070	0.044
	(0.088)	(0.095)	(0.089)	(0.088)	(0.158)	(0.139)
Pupil Chars	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
*p-value: Predicted High = Predicted High * Treated*	0.537	0.073	0.932	0.733	0.594	0.452
R^2	0.227	0.235	0.272	0.276	0.172	0.208
Observations	3163	3365	3121	3347	3254	3463
Schools	10	10	10	10	10	10

Robust standard errors clustered at the school level in parentheses
source: analysis tables 20151212.do
* p<0.10, ** p<0.05, *** p<0.01

Appendix Table 4: Effect of treatment interacted with pupil characteristics

	Math	English	Science
Financial Incentive	0.154	0.036	-0.029
Financial Incentive * Female	0.019	0.049	0.073
Financial Incentive * Asian	-0.005	-0.045	0.032
Financial Incentive * Black	0.119	0.042	0.053
Financial Incentive * EAL	-0.156	-0.042	-0.111
Financial Incentive * FSM	0.011	0.041	0.039
Financial Incentive * Baseline Test	-0.026	-0.019	-0.004
Financial Incentive * London school	-0.016	-0.008	0.179
Financial Incentive * Academy school	0.095	0.083	0.046
Non-Financial Incentive	0.181	0.089	0.010
Non-Financial Incentive * Female	0.073	0.106	0.099
Non-Financial Incentive * Asian	0.070	-0.011	-0.032
Non-Financial Incentive * Black	0.081	-0.017	-0.037
Non-Financial Incentive * EAL	0.022	-0.083	0.025
Non-Financial Incentive * FSM	-0.048	0.048	0.083
Non-Financial Incentive * Baseline Test	-0.027	0.002	0.014
Non-Financial Incentive * London school	-0.178	0.002	-0.066
Non-Financial Incentive * Academy school	-0.041	-0.170	0.098
R^2	0.077	0.104	0.148
Observations	20058	19791	20713
Schools	63	63	63

Coefficients only fir brevity; few of the interactions are statistically significant.

Appendix Table 5: Impact on scores on non-incentivized subjects – French, history

	French		History	
	Financial Incentive	Non-Financial Incentive	Financial Incentive	Non-Financial Incentive
Predicted High	-0.002	-0.097	0.075	0.029
	(0.079)	(0.106)	(0.075)	(0.063)
Predicted High * Treated	-0.008	0.257	-0.193	0.009
	(0.199)	(0.204)	(0.127)	(0.097)
Predicted Low * Treated	0.026	0.010	0.006	-0.075
	(0.150)	(0.195)	(0.115)	(0.134)
Pupil Chars	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes
*p-value: Predicted High = Predicted High * Treated*	0.978	0.180	0.091	0.875
R^2	0.046	0.048	0.093	0.093
Observations	2917	3303	3792	3984
Schools	44	46	47	47

Robust standard errors clustered at the school level in parentheses
source: analysis tables 20151212.do
* p<0.10, ** p<0.05, *** p<0.01