Do Performance Targets Affect Financial Support to Schools? Evidence from Discontinuities in Test Scores*

Marcello Sartarelli† Alessandro Tampieri‡

March 20, 2012

Abstract

This paper studies the effect of meeting performance targets in tests by students on the subsequent take-up by schools of government financial support for students with special needs. We use a regression discontinuity design that exploits thresholds in test scores to identify the effect and we estimate it by using administrative data on students in state schools in England. We find that the effect of meeting the target is negative and weakly significant for support programs whose assignment rules are discretionary, it is not significant for programs with non-discretionary rules, and it also varies by students' ability. Naive estimates show instead that the effect is negative, significant, and greater than regression discontinuity estimates. Finally, we build a stylised model to interpret the empirical findings by comparing the optimal levels of financial support that the government and the school would supply respectively, since the two institutions face different costs and incentives. The empirical evidence and the interpretation of the mechanism driving it inform the decisions of policy-makers on the allocation of transfers to schools, or to other public sector institutions.

JEL Classification: C21, H52, H53, I22, I28
Keywords: additional language, education, education finance, performance targets, regression discontinuity, school meals, special education needs, test scores

*We are grateful to Nathan Carroll, Valentina Conti, Gianni De Fraja, Iñigo Iturbe-Ormaetxe, Irene Mammi and Arsen Palestini for their comments. We gratefully acknowledge funding from the ADMIN Node of National Centre for Research Methods (ESRC grant RES-576-25-0014) and from Hera Holding. All errors remain our own.

†Institute of Education, University of London, 20 Bedford Way, London WC1H0AL, UK and University of Alicante, Economics Department, Apartado de Correos 99, 03080 Alicante, Spain, Email: m.sartarelli@ioe.ac.uk

‡Department of Economics, University of Bologna, Strada Maggiore 45, 40125 Bologna, Italy. Email: alessandro.tampieri@unibo.it

1 Introduction

Performance targets are adopted in several dimensions of an individual's life, such as education and the labour market. Targets are important in helping individuals to build human capital or signal ability at school and subsequently in a job. An individual may have little incentive to study or work if performing above the average level does not lead to any rewards in the future. However, performance targets may lead to dysfunctional responses by individuals if the rules or laws that link targets to rewards, or their implementation, are subject to manipulation or strategic behaviour. Individuals may not exert effort to meet the target and focus instead on actions that aim at manipulating their performance or the performance target itself. When such actions are considered undesirable, this may lead to an intervention by administrators or policy-makers to rectify rules, laws, or their implementation.

The effect of incentives in employment contracts has been widely explored.[1] Instead little is known about the potentially dysfunctional responses by individuals to policies that exploit performance targets, since a policy-maker may have not anticipated such actions when designing a policy. For example, if the government offers additional financial support for each student with special education needs in a school, its teachers may have an incentive to define as having special needs those students who performed poorly in tests in the past. What may induce this is the negative association between students having special needs and their performance in tests, which a school may try to mitigate by obtaining additional funding to improve the future performance of those students.

In this paper we study whether meeting performance targets in tests by students has an effect on their assignment by schools to support programs for students with special needs, which lead to additional funding for schools. We set out to answer this empirically by using administrative data on tests scores and on the subsequent assignment to government support programs of a cohort of students in compulsory education in state schools in England. The funding for the support of students with special needs in schools is approximately £4 billion yearly, or 13% of the current expenditure in education in England (House of Commons Education and Skills Committee (2006)). Expenditure in similar programs in the USA is lower, although obtaining a comparable figure is complicated by the federal structure of the education budget in the USA (Figlio (2003)).

In the empirical analysis probit estimates of the effect of test scores on the subsequent as-

[1] See Prendergast (1999) for a review of incentives and performance targets in employer-employee contracts.

signment of students to government financial support programs are negative and statistically significant. However, this may be a spurious correlation that also captures the effect on the assignment to government support programs of unobserved variables that influence students' achievement, e.g. the socio-economic background. Thanks to thresholds in test scores which the Department for Education set as performance targets for students and schools we use a regression discontinuity design to identify the effect of meeting a performance target by students on the probability of the take-up of government support programs by schools. We show that the only reason why one student marginally meets a performance target, while another students with very similar characteristics marginally misses the target, is a source of randomness in the process of marking students' tests. Hence, there is no reason why one of the two students is in greater need of additional teaching support than the other, which leads to the prediction of non-significant estimates of the effect of meeting a target on the assignment to a support program in the absence of unobserved actions by teachers. Conversely, if the probability of assignment of students to a program jumps discontinuously at a target, this suggests that teachers may have an incentive to assign only one of two similar students to a government support program, and obtain additional resources, for reasons that are related to whether a student met or failed to meet the target, rather than to the student's special needs.

We find that the effect of meeting the target is negative and weakly significant for programs whose assignment rules are discretionary, it is not significant for programs with non-discretionary rules, and it also varies by student's ability, thus not entirely rejecting the hypothesis of no dysfunctional responses by schools to the policies on the support for students with special needs.

We interpret the empirical evidence by using a stylised theoretical model, in which a school decides whether to assign students to support programs as this leads to i) additional financial support for the school and ii) potential reputational costs of applying for the support programs since school choices by parents in the future may depend on the share of students in support programs in a school among other factors. The resulting demand for financial support for such programs by the school may be greater than the one arising in the event that the government allocates the funding, since the cost of reputation that the school incurs for obtaining the financial support is marginally lower than the monetary cost of the financial support that the government incurs.

Recent studies show evidence on the implementation of administrative rules or of laws whose

objective is inducing schools to modify their actions in a direction that is arguably socially desirable. Cullen and Rivkin (2003) study empirically the mechanisms linking school choice to programs for special education needs in the USA. The empirical evidence shows that students with special needs may have less choice over schools than other students, and also that the schools with high shares of students with special needs may be of little appeal to other students. Figlio and Getzler (2002) study the effect of policies that reward those schools whose students obtain high test scores, or improve their test scores over time, on actions by schools, i.e. test-based accountability systems. They find that in Florida the schools that were subject to an accountability system increased the assignment of low-performing students to disability programs, since the assignment to the program excluded students from the accountability system, thus increasing the average performance in tests by students in a school.

In related research on education finance Figlio (2003) studies the effect of a school accountability program in the USA, the No Child Left Behind Act, on the allocation of funding to schools. The empirical evidence shows that the program has direct fiscal consequences for certain types of schools, and also indirect ones, for example, on the assignment of students to special education programs. Similarly, Cullen (2003) studies the impact of fiscal incentives on the assignment of students to disability status by schools. Empirical evidence that exploits court-induced variation in education finance rules in Texas shows that the introduction of fiscal incentives increases observed disability rates. Cullen and Reback (2006) study similar incentives in a school accountability system in Texas, and also find that schools try to game the system.[2] Finally, Crawford and Vignoles (2010) and Keslair *et al.* (2011) evaluate a policy for students with special needs in the UK and find that the effect of the policy on students' test score is null or negative, hence casting doubts on the content of the policy or on its implementation.

This paper proposes a novel application of a statistical procedure to assess whether the rules governing the assignment of students to support programs and the allocation of additional funding for students in schools, and their implementation, may lead teachers to apply for support programs also for students without special needs. It also interprets the mechanism at play in the estimates by using a stylised model that compares the optimal level of financial support that the government and the school would supply respectively, since the two institu-

[2]In related studies Urquiola and Verhoogen (2009) assess the effect of class size on test scores in schools in Chile. They find that the administrative rules determining class size induce schools to change fees or enrollment at administrative class-size thresholds, thus leading to spurious estimates of the effect of class size on test scores.

tions face different costs and incentives. This contributes to the literature on the provision of incentives in government interventions in education. In addition, it is helpful to inform public policy decisions on the provision of incentives in education, and in other areas in the public sector.

The structure of the rest of the paper is as follows. Section 2 describes the institutional setting of compulsory education in England, and the summary statistics of the dataset of students that we use in the empirical analysis. Section 3 sketches the econometric strategy and section 4 presents the results from the empirical analysis. Section 5 presents a stylised model to interpret the results, and Section 6 concludes.

2 Institutional setting and data

In this section we describe the source of exogenous variation in test scores that identifies the effect of meeting performance targets in tests in primary school on the subsequent assignment of students to government support programs in secondary school. We estimate the effect by using administrative data, the National Pupil Database (NPD), with information on test scores of students who completed primary education in state schools in England in Summer 2001.[3] The data also contain information on whether students were subsequently assigned to government support programs within 2 years of starting secondary school. The timeline in Figure 1 summarises the sequence of events over time.

In England there are 11 years of compulsory education, which is divided into the Foundation Stage Profiles, plus 4 Key Stages, as Table 1 shows. The Foundation Stage Profile starts at age 3-4. Primary school starts at age 5-6 with Key Stage 1 and it is followed by Key Stage 2, as columns (1)-(3) show. Secondary school starts with Key Stage 3 and students typically complete it at age 15-16 by obtaining the General Certificate of Secondary School (GCSE) after passing the school leaving exam at the end of Key Stage 4.[4] Column (5) shows the type of assessment at each stage, which varies from teacher assessment to national assessment by external examiners. Lastly, column (6) shows the achievement levels or targets that the Department for Education expects students to meet at each Key Stage. Such targets are set out to help students, parents and schools to interpret a student's progress throughout

[3]Private schools account for about 7-8% of students in compulsory education for the period 1990-2006 (Green *et al.* (2010)).

[4]See Bradley *et al.* (2000) for additional information about the institutional setting of secondary education in England.

compulsory education.[5]

2.1 Key Stage 2 tests

The NPD data contain information about students who sit the compulsory tests in English, Maths and Science in year 6, the last in Key Stage 2, when they are 10 or 11 years old. Such tests are set by the Qualifications and Curriculum Development Agency (QCDA (2010)), which is an independent authority from the Department for Education.[6] Students are also assessed by their teachers before results in the Key Stage 2 tests are known.

External examiners mark test scripts by using numerical scores on an integer scale whose range varies by test. Four categorical achievement levels from 2 to 5 are also created as intervals of test scores. In 2001 a score lower than the threshold 22 in the Maths test leads to an achievement level equal to 2 in Maths and a score in the interval 22-48 leads to a level equal to 3.[7] The Department for Education turns raw test scores into "fine grade"test scores, that are decimal numbers in the range 2-6 and they are obtained by weighting test scores by the distance from the nearest threshold to the right of the score. For example, the threshold in the Math test score equal to 22 in the earlier example is equal to 3 in the fine grade point score scale. Test scores equal to 21 and 23 are equal to 2.96 and 3.04 in the fine grade Maths score. In the empirical analysis we use the *fine grade* point score rather than the raw test score as it is advantageous to obtain precise inference.[8]

Examiners know the thresholds for each achievement level when they mark tests. However, they do not the know students and vice versa, and QCDA trains examiners to ensure high standards of consistency in the marking process. This rules out perfect manipulation of test scores as examiners have no information about students. In addition, one examiner gets all test scripts in one type of test, e.g. English, in a school. Hence, a student has his or her tests in English, Maths and Science each marked by a different examiner who only knows the score of one of the three tests by a student. This rules out manipulation both of the other two tests and

[5]DirectGov (2010) is a government-maintained website to inform citizens about the characteristics of services in the public sector in the UK. It motivates the test score targets by the Department for Education at each Key Stage in compulsory education as follows: "Children develop at different rates, but National Curriculum levels can give you an idea of how your child's progress compares to what is typical for their age".

[6]For example the Key Stage 2 Maths test verifies learning of *i*) using and applying numbers such as problem solving and communication, *ii*) numbers and the number system such as counting, percentages and ratios, *iii*) calculations such as mental and written methods and *iv*) solving numerical problems such as combining number operations. See QCDA (2010) for additional information.

[7]QCDA (2010) offers additional information about thresholds in all Key Stage 2 tests.

[8]See Lee and Card (2008) for additional details about inference in a regression discontinuity design with an integer-valued running variable.

of the average test score. Finally, unobserved effort by teachers in helping students with the test preparation is not a serious concern for the research design since teachers' compensation in the UK is independent of students' performance in tests.

Categorical achievement levels in externally marked tests in English, Maths and Science at Key Stage 2, together with teacher assessments in these subjects, are disclosed to students and parents.[9] However, the underlying test scores are not disclosed, which is critical for the research design in this paper.[10] For example, two students whose average fine grade score is 3.03 and 3.97 get level 3. Conversely, two students scoring 3.97 and 4.05 in the same test get level 3 and 4 respectively. When performances in tests are disclosed, students, teachers and parents can compare the students' achievement level with the expected performance target level, which is equal to 4, as column (6) in Table 1 shows. In addition, achievement levels 3 and 5 are implicit targets for low and high ability students respectively. They may be more relevant than the expected target if students' past achievement, e.g. in the teacher-assessed tests at Key Stage 1, was so low that they will be very unlikely to achieve the expected target at Key Stage 2, level 4, or vice versa so high that they will score considerably above that target. Table 2 shows in the second panel that approximately 20% of students do not meet the expected achievement target at level 4 in tests in English, Maths or Science, 50% meet the target, and the remaining ones achieve instead a level above the target. The third and fourth panel in the table show that 67% of students attend Community secondary schools, while the remaining ones attend either Voluntary or Foundation schools. In addition, half of the students who took Key Stage 2 tests in 2001 are males, 86% white and the remaining ones are Asians, Black or belong to other minority groups.

2.2 Government financial support programs in schools

In addition to information on test scores, the NPD data also contains information from the Pupil Level Annual School Census (PLASC) about whether they are eligible for government support programs at school: Free School Meals (FSM), English as Additional Language (EAL) or Special Educational Needs (SEN). Students are eligible for Free School Meals (FSM) if their parents meet eligibility criteria based on income and receipt of social benefits. Similarly, English as Additional Language (EAL) or Special Educational Needs (SEN) programs offer

[9] Additional information about the administration of Key Stage 2 tests is available in the UK Parliament Statutory Instruments 1999 No. 2188, 2001 No. 1286 and 2003 No. 1038.

[10] Sartarelli (2011) describes additional details about the results sheet that schools use to disclose test outcomes to students and parents.

additional support at school to students who meet a number of eligibility criteria that are assessed by teachers and psychologists. The SEN program offers two mutually exclusive types of support: students whose needs are assessed as low are assigned to the non-statemented SEN program, while those with more severe needs are assigned to statemented SEN program (DfE (2010)). Differently from FSM, assignment to EAL and SEN programs is based on criteria that the Department for Education set out and that allow for discretionary assignment by schools, since some of the criteria are subjective. The total yearly government expenditure in EAL and SEN programs in secondary schools in England is approximately £4 billion yearly, or 13% of the current expenditure in education in England (House of Commons Education and Skills Committee (2006)). FSM expenditure is instead considerably lower, as it consists in a voucher to purchase a meal at school.

Table 2 shows in the top panel that 14% of students are assigned to the non-statemented SEN program, while 2% to the statemented SEN. 8% of students are assigned to the EAL program and 15% receive free school meals. The summary statistics by gender show little variation in the assignment to the programs, other than for non-statemented SEN. Figure 2 shows similar information on each government program, but by subgroups of students with different achievement levels in tests at Key Stage 2, going from 2 to 5, on the horizontal axis. 40% of students whose average achievement level in tests is 3 are assigned to the non-statemented SEN, while only 10% of student whose achievement level is 4 are, and for the same groups of students the percentage of students assigned to EAL goes from 10% to 8%. Overall, the mean take-up in these programs decreases with the achievement level in tests at Key Stage 2. However, the figure also shows low variation, for example in EAL, between adjacent achievement levels in tests, thus emphasising the importance of assessing whether these correlations also have a causal interpretation.

3 Research design

We assess the effect of meeting the performance target at Key Stage 2 on the probability that students are subsequently assigned to government support programs at school. We use as outcome variable a dummy G that is equal to one if a student is assigned to a government program in the school year following Key Stage 2 tests, for example, English as Additional Language, and zero otherwise. As independent variables we use a continuous measure T of test score and a dummy $D = I\{T \geq \bar{T}\}$ that is equal to one if a student's test score T is

greater or equal than a performance target $\bar{T}$, and it is equal to zero otherwise.

$$Pr(G = 1|T) \quad = \quad \Phi(\alpha + \beta_{Probit} T) \tag{1}$$

Consider in equation (1) a probit regression of the binary assignment G to a program on test scores T, the average test score over all tests at Key Stage 2, where $\Phi(.)$ is the cumulative distribution function of the standard normal. The marginal effect that is associated to β_{Probit} is interpreted as the change in the probability that after tests a student is, for example, assigned to the English as Additional Language program. An increase in test score from 3 to 4, or similarly from 3.2 to 4.2, leads a student to meet the expected performance target. Similarly, an increase from 4 to 5 leads a student to meet the performance target for high ability students. This gives an insight into the difference in the probability of assignment to support programs by students with different performances in tests. However, unobservable ability of children, parental care and school arrangements contained in the error term in equation (1) may lead to a spurious correlation between test scores and assignment to a program if unobservables correlate with test scores.

$$G \quad = \quad \alpha + \beta_{RD} D + \gamma(T - \bar{T}) + \delta D * (T - \bar{T}) + U \tag{2}$$

Using instead information on students who marginally met an achievement target, e.g. obtain a test score greater than or equal to a threshold value $\bar{T} = 4$, relative to students who marginally failed to meet it, e.g. obtain a score smaller than 4, identifies the effect of meeting a performance target with respect to not meeting it on the probability of assignment to a government support program by exploiting a regression discontinuity (RD) research design.[11] This is the interpretation of the parameter β_{RD} in equation (2), whose independent variables are the dummy D that is equal to one if a student's test score T is greater than or equal to a performance target $\bar{T}$ and zero otherwise, the difference between test scores and the threshold value and their interaction. β_{RD} is, for example, negative if students who met a performance target ($P = 1$) are also less likely to be assigned to a government support program than those who did not meet it ($P = 0$). Three thresholds $\bar{T}$ equal to 3, 4 or 5 in test scores determine whether a student meets a performance target P. The expected performance target that the Department for Education sets for all students at Key Stage 2 is 4. By exploiting this target

[11] Thistlethwaite and Campbell (1960) and Trochim (1984) developed the RD design. Imbens and Lemieux (2008) and Lee and Lemieux (2010) survey the advances in the theory as well the recent increase in the number of applications of the design in economics.

and those for low and high ability students at targets 3 and 5, one estimates the effect of just meeting the expected performance target at Key Stage 2 with respect to just missing the target on the probability that, a student is, for example, truant two to three years after the disclosure of the test results.

The research design is non-experimental. However, it is similar to an experimental one for three reasons, as Lee and Lemieux (2010) suggest. First, students take decisions and act to maximise the probability of meeting a performance target in test scores (the treatment), before the test date and with the aid of parents and teachers. Second, obtaining a test score to the left of a threshold or target (control group) or to the right of it (treatment group) can be interpreted as a stochastic shock to the test score due to nature, as scripts in the three compulsory tests are marked externally. Third, the treatment is assigned on the basis of the value of the test score, i.e. the running variable. In the empirical analysis we use the average test score over all tests as running variables since students have to meet a threshold in all tests, and under the assumption that they put effort in all tests, as the evidence in section 2 suggests. The RD design holds under the identifying assumptions that students on the left of the threshold $\bar{T}$ are similar to those on the right of it, for example, in their socio-economic background that parents' education proxies.[12] This is testable by estimating the mean of conditional residuals in the left neighbourhood of the threshold, $\lim_{T \uparrow \bar{T}} E[U|T]$, and in the right neighbourhood, $\lim_{T \downarrow \bar{T}} E[U|T]$, and it holds if their difference is not statistically significant. Robustness checks in section 4.2 offer evidence that this assumption holds.

Students either obtain an average score in the tests that is greater or equal than 4, thus achieving on average the performance target set by the government, or fail to obtain this. This target as well as targets 3 and 5, that are respectively lower and higher than the expected one, are thresholds at different percentiles of the distribution of the average test score. Hence, we estimate the effect of meeting a target on the assignment to support programs in schools by using a sharp RD design and we use those observations that are in the interval $[\bar{T} - 1, \bar{T} + 1]$, that goes from the threshold $\bar{T} - 1$ that is to the left of $\bar{T}$ to the threshold $\bar{T} + 1$ to the right of $\bar{T}$. A smaller window around the threshold $\bar{T}$ would omit relevant observations of students. A larger window would instead include students whose fine grade test score is either so low, e.g. 2, or high, e.g. 5 that no random shock in test score would be big enough to achieve the target 4, thus estimating a potentially misspecified model. We choose the bandwidth of the

[12] The identifying assumptions in a RD design are testable differently from instrumental variable or matching on observables strategies, as Lee and Lemieux (2010) suggest.

polynomials by using the data-driven choice rule in Imbens and Kalyanaraman (2009) that corrects an asymptotically optimal bandwidth in theory for small sample size and specification problems. In the preferred specification we also add as covariates in the polynomials pre-determined characteristics of students that are listed in Table 2.

4 Empirical results

In section 4.1 we describe probit and regression discontinuity estimates of the effect of meeting a performance target in tests on the probability of assigning students to government support programs. In section 4.2 we assess the sensitivity of the estimates by conducting robustness checks.

4.1 Probit and regression discontinuity estimates

Table 3 shows probit estimates and RD estimates of the effect of meeting a performance target in tests on the probability of being assigned to support programs. Negative and significant probit estimates show that the probability of assignment to a government support program decreases with students' test score. However, the estimate may be spurious as social economic background and unobserved ability tend to be correlated with test scores.

Exploiting instead a discontinuity in test scores that determines whether a student has met a performance target in tests shows a mixed pattern in the sign and in the precision of the effect. The effect of meeting the target on the probability of assignment to non-statemented SEN program is small in magnitude and not significant. The probability of assignment to the EAL or to the statemented SEN programs show instead a significant change at the test score thresholds that the government expects students to meet, in column (4) in the table, and that low and high ability students meet, in columns (3) and (5), respectively. Overall, the sign, size and precision of discontinuous jumps in the probability of assignment to such programs at the thresholds in test scores that the Department for Education set as targets for students, parents and schools varies for students with different ability, i.e. at different thresholds. Since assignment of students to EAL and SEN programs is decided by students' teachers, a jump in the measures of outcome with discretionary assignment may suggest unintended actions by school, since two students on different sides in a small neighbourhood of a test score threshold are similar, and should be treated similarly in the assignment to a program. The probability of assignment to FSM instead does not change discontinuously at a performance target, which

one expects if the assignment to FSM that is based on administrative records of the receipt of benefits by students' is not instead subject to discretionary judgment.

4.2 Robustness checks

A RD design to study the effect of performance targets is similar to the assignment of students to either side in the neighbourhood of test score targets 3, 4 and 5 in a coin-flip experiment. We test empirically whether obtaining a test score to either side of a performance target offers a valid research design to identify the effect of meeting a target on the assignment of students support programs with three robustness checks. This consists of assessing the two testable aspects of the identifying assumption that the design is as if it was locally randomised at the thresholds.[13]

The first one is that the distribution of all pre-determined characteristics of students, such as gender, ethnicity and other variables proxying socio-economic background, is the same just to the left and to the right of a threshold of the fine grade average test score, as these characteristics are determined before the test scores are disclosed and a local randomisation at the threshold does not alter their distribution. A RD design is valid, for example, if the share of students from ethnic minorities who score just to the left of a threshold is not different from the share of ethnic minorities scoring just to the right of it. Failing this, the effect on behaviour that one attributes to meeting a performance target is confounded by the correlation between ethnicity and performance in tests, thus invalidating the randomised design around a threshold. We estimate in Table 4 smooth polynomials in the fine grade average test score of the pre-determined characteristics in column (1), separately for samples of students who score to the left or to the right of thresholds 3, 4 and 5. Insignificant estimates of regression discontinuity regressions of pre-determined characteristics with respect to tests at Key Stage 2 on the running variable at Key Stage 2 lead not to reject the null hypothesis of no significant difference in the value of such characteristics of students, which supports the validity of the design.

The second testable part of the identifying assumption is the inability of students, parents, or schools to perfectly manipulate test scores. This holds if before test scores are disclosed students, parents or schools cannot precisely guess whether their score is to the left or right of a target in any subject, or manipulate it. Marking of test scripts by external examiners, rather

[13] In a related study Sartarelli (2011) finds that estimates obtained by exploiting the same research design are not sensitive to the definition of the variables in the empirical analysis.

than by students' teachers, ensures local randomisation of test scores in the neighbourhood of a threshold.[14] Potential manipulation around a threshold can be detected graphically by visually inspecting an undersmoothed histogram of the fine grade average test score with a small binwidth, such that bins contain an arbitrarily small number of students separately to the left and right of a threshold and no bin contains the threshold value. Evidence in Figure 3 suggests no suspicious jumps in the empirical density around thresholds.

Moreover, McCrary (2008) developed a formal test of the null hypothesis of no manipulation, which is not rejected if the difference in the height of the undersmoothed histogram bins to the left and to the right of a threshold is sufficiently small. The rows in Table 5 show t-statistics of t-tests in McCrary (2008) at each threshold in test scores for the fine grade test scores in each subject and for the average score. The top three rows in the table show t-statistics of scores in each compulsory subject test: English, Maths and Science. The t-statistics are greater than 2 in all cases and often by a large margin, thus implying the rejection of the null of no manipulation in subject tests. However, manipulation is not statistically significant when the fine grade test scores are averaged, as the bottom row in the table shows. This is because manipulation of the running variable is imperfect by design, with different external examiners marking tests in a different subject tests for each student. For example, one examiner may attempt to manipulate the score in English tests in one school although he does not know the students. However, such manipulation attempt may cancel out or be reversed when averaging test scores in English, Maths and Science to compute the fine grade average test score.

5 Theoretical model

The empirical evidence shows that the demand for the funding that is linked to government support programs in schools jumps discontinuously in the neighbourhood of performance targets in students' test scores for certain types of programs while it does not for others. We interpret this evidence by considering the different cost structure that the government and the school face in supplying the additional teaching respectively. The government funds the additional teaching while the school obtaining the fundings bears instead a cost in terms of reputation. Indeed, a high demand for government support can be seen by parents as a signal of low average ability of the pupils within the school. The reputational cost increases with the

[14] See as examples of potential gaming around threshold Jacob and Lefgren (2004) that study the effect of remediation courses on test scores in schools in the USA and Urquiola and Verhoogen (2009) that study the effect of class size on test scores in schools in Chile. In both examples gaming is induced by unintended responses of teachers to incentives that are embedded in the institutional setting.

demand for additional teaching. In addition, the cost increases faster the greater the overall demand, since parents and students prefer a school with a low overall demand for additional funding from a support program *ceteris paribus*. Coversely, the monetary cost increases with the demand of fundings at a decreasing rate, for the presence of economies of scale (hiring a teacher or a psychologist provides help from one to many students in an institution, as well as renting a classroom). In other words, the marginal cost of reputation for the school is lower than the marginal monetary cost that is borne by the government. Hence, the school is willing to ask for more funding than the amount that the government would optimally provide.

The school may ask for funding irrespectively of the students' test score at Key stage 2. However, motivating an application for additional teaching for a student who obtained a score above the performance target is harder than for a student whose score is below the target. Hence, schools may use performance targets in test scores as a rule of thumb to apply for additional teaching resources, rather than equating the marginal monetary cost of teaching with its marginal benefit for a student. The reason is that by using the performance target a school may obtain more financial support for additional teaching than it would by comparing marginal costs and benefits. As a consequence, since the score is a noisy signal of ability, some students obtain the subsidy while others with the same ability level do not. We show this by using a stylised model that characterises the decision by a school to apply to the government to obtain funding for additional teaching.

We study an economy in which students take a school entry test, the school applies for financial support for additional teaching for students with learning difficulties, and the government funds the additional teaching.[15]

Students. We consider a continuum of students with mass normalised to one. Students differ in ability $\theta \in [0,1]$ that is distributed according to a probability density function $f(\theta)$ and a cumulative distribution function $F(\theta)$. We abstract from students' effort and we assume that test scores are a noisy measure of students' ability, e.g. a student may be unlucky and have a bad day in the test, or similarly she may not feel well. Hence, we assume that under uncertainty a student's test score $g \in [0,1]$ is a linear function of ability θ and of an error term ε:

$$g = \theta + \varepsilon \ , \ \varepsilon \sim \Phi\left(0, \sigma^2\right). \tag{3}$$

[15]For simplicity, we abstract from factors such as competition among several schools, that we will address in future research.

We also assume that there exists an exogenously given performance target g_t that splits students into two groups based on their test score. The students in one group met the target while the students in the other group did not. Further to the school application, a student may receive additional teaching. The total benefit of additional teaching for the students is given by $B(\theta) = b[F(\hat{\theta})]$, where $F(\hat{\theta})$ is the quantity of students receiving extra teaching, and $\hat{\theta}$ is the ability level of the most able student provided with additional teaching, that is, the student who receives additional teaching "at the margin". The marginal benefit is a decreasing function of ability in such a way that the least able student obtains a very large benefit from it while the most able student will not receive any benefit, therefore $b' \geq 0$, $b'' \leq 0$, $\lim_{\hat{\theta}\to 0} b'[F(0)] = +\infty$ and $b'[F(1)] = 0$.

The government. It supports additional teaching, although the decision on whether or not providing it is up to the school. The financial cost of additional teaching is denoted by $C = c[F(\hat{\theta})]$. The marginal monetary cost increases with the number of students receiving the support but at a decreasing rate, i.e. $c' \geq 0$, $c'' \geq 0$, $c''' \leq 0$. These assumptions are such that the functional form of the cost function mimics the cost of a public policy intervention. When the policy intervention is introduced it is usually piloted on a small sample, its cost increases with the size of the pilot although at a decreasing rate. Moreover, since the benefit of additional teaching is higher the lower the student's ability, the government will start providing it from the least able student upward, so that $c'[F(0)] = 0$. With these assumptions, we can identify the optimal level of financial support by the government. This is reached when the marginal level of ability is $\hat{\theta} = \bar{\theta}$, where $\bar{\theta}$ is the level of ability such that the marginal cost of provision equates the marginal benefit, $b' = c'$.

The school. It decides for which students to apply for additional teaching support. We assume that the school can apply only if the student achieved a score that is smaller or equal to g_t.[16] If a school decides to apply for the support, the government incurs the monetary costs for it rather than the school itself. However, obtaining the funding lowers the school reputation, by negatively altering the perception about the quality of the institution and about the ability of its students. We denote the school cost in terms of reputation as $R = r[F(\hat{\theta})]$. Since the lower the student's ability the greater the benefit from additional teaching, the school will apply for it starting from the least able student and subsequently for marginally more able

[16]Schools in the UK can apply for additional funding for students with special needs, e.g. EAL and SEN programs irrespective of the student's score. However, it is hard to justify additional funding for students with a score that is considerably greater than the threshold score.

students, so that $r'[F(0)] = 0$. We also assume that R increases in the level of funding ($r' \geq 0$) and that the reputational cost is marginally low if relatively few students are in need of additional teaching and it is instead high otherwise ($r'' \geq 0$ and $r''' \geq 0$). Finally we assume that the marginal monetary cost is higher than the marginal cost of reputation for every $\theta \in [0, 1]$, so that $c' > r'$.

The school applies for the level of fundings for additional teaching that maximises the benefit that students obtain from the additional teaching net of the reputational cost that is associated with the funding, subject to the constraint g_t. Hence the school problem is:

$$\begin{cases} \max\limits_{\hat{\theta}} \; b[F(\hat{\theta})] - r[F(\hat{\theta})] \\ \qquad s.t. \; \hat{\theta} \leq g_t \end{cases} \tag{4}$$

In equilibrium all the students with ability equal or below $\hat{\theta} = \min\{\theta^*, g_t\}$ receive additional teaching, where:

$$\theta^* = \arg\max_{\hat{\theta}} \; b[F(\hat{\theta})] - r[F(\hat{\theta})], \tag{5}$$

is the unconstrained equilibrium in the demand for additional teaching. The foregoing discussion can be summarised in:

Proposition 1 *Let the assumptions on C and R hold. Then $\theta^* \geq \bar{\theta}$ and the demand for additional teaching may lead to an ine cient provision of teaching support.*

Proof. Given $r' \geq 0, r'' \geq 0, r''' \geq 0, r'[F(0)] = 0,\ c' \geq 0, c'' \geq 0, c''' \leq 0, c'[F(0)] = 0$ and the fact that $c' > r'$ for every $\theta \in [0, 1]$, the marginal benefit curve b' will cross the marginal monetary cost curve c' on the left of the level at which it crosses the reputational cost curve r', so that we exclude the case where $\theta^* < \bar{\theta}$ ■

Figure 4 illustrates an example of Proposition 1 in which:

- F being a uniform distribution between 0 and 1 so that $F = \theta$; thus we can rewrite the benefit and cost functions as $B = b(\hat{\theta}), C = c(\hat{\theta})$ and $R = r(\hat{\theta})$, and
- the equilibrium in the demand for additional teaching is reached at g_t.

The optimal provision of additional teaching is the level $\bar{\theta}$ that equalises the marginal benefit for its provision with the associated monetary cost ($b'(\hat{\theta}) = c'(\hat{\theta})$). Nonetheless, the school chooses the performance target level $\hat{\theta} = g_t$ as it is closer to its preferred level θ^*, i.e., where $b'(\hat{\theta}) = r'(\hat{\theta})$. The figure shows that the school may ask for a greater level of additional

teaching than the optimal one ($\hat{\theta} > \bar{\theta}$) since the cost of reputation for the school is marginally lower than the monetary cost that the government pays to offer the the additional teaching. In addition, some students may receive additional teaching, while students with the same ability level may not receive it since test scores are a noisy measure of ability in the neighbourhood of test score thresholds.

6 Conclusion

This paper studies the effect of meeting performance targets in school tests on the subsequent take-up by schools of government financial support for students with special needs. We exploit thresholds in test scores to identify the effect by using a regression discontinuity design, since the discretionary assignment of financial support by schools may confound probit estimates We estimate the effect by using administrative data on students in state schools in England. We find that probit estimates show a negative and significant correlation between financial support and test scores, while the sign and the precision of regression discontinuity estimates vary by type of support program and by students' ability. Finally, we build a stylised model to interpret the empirical results by comparing the optimal levels of financial support that the government and the school would supply respectively, since the two institutions face different costs and incentives. The paper contributes to the literature by proposing a novel application of a statistical procedure to test teachers' behaviour, and by interpreting the mechanisms driving the behaviour thanks to a model.

This type of government financial support to schools for students with special needs is worth approximately £4 billions yearly, or 13% of current expenditure in education in England. Finding that, in addition to the budget that schools are allocated, the probability of assignment of students to a support program jumps discontinuously at a target in test scores, would suggest that teachers obtain additional resources to help only one of two otherwise similar students. Overall, the tax revenues that are potentially misallocated due to teachers' behaviour are a small fraction of the expenditure in education, which does not undermine the value of such interventions in education for students. However, it may suggest to revise certain characteristics in the design and in the implementation of the policies offering additional support to students in education. For example, a policy may jointly aim at delivering funds to support students in need, which is essential for teachers to help such students, and also at minimising potentially dysfunctional responses by teachers to a policy, which helps them to fully focus

their effort on teaching students.

Similar procedures to test potentially dysfunctional actions can be carried out in the future and inform the decisions of policy-makers over education and public policies, such as curriculum design in compulsory education, and the design of incentive schemes for teachers and for civil servants in other areas in the public sector. This can be done by using similar linked administrative data in the UK or in the USA, as well as in other developed and developing countries whose governments collect data on students' achievement at school and actions by schools. Offering guidelines for the setup of an institutional setting that allows one to exploit similar research designs in other policy-relevant applications fits into a broader research agenda to study the role of incentives in the design of public policies.

References

Bradley, S., Taylor, J., Millington, J. and Crouchley, R. (2000). Testing for quasi-market forces in secondary education. *Oxford Bulletin of Economics and Statistics*, **62** (3), 357–90.

Crawford, C. and Vignoles, A. (2010). *An analysis of the educational progress of children with special educational needs*. DoQSS Working Papers 10-19, Department of Quantitative Social Science - Institute of Education, University of London.

Cullen, J. B. (2003). The impact of fiscal incentives on student disability rates. *Journal of Public Economics*, **87** (7-8), 1557–1589.

| and Reback, R. (2006). *Tinkering Toward Accolades: School Gaming Under a Performance Accountability System*. Working Paper 12286, National Bureau of Economic Research.

| and Rivkin, S. G. (2003). *The Role of Special Education in School Choice*, University of Chicago Press, pp. 67–106.

DfE (2010). Special educational needs in england. http://www.dcsf.gov.uk/rsgateway/DB/SFR/s000939/SFR19-2010.pdf.

DirectGov (2010). Understanding the national curriculum. http://www.direct.gov.uk/.

Figlio, D. N. (2003). Fiscal implications of school accountability initiatives. In *Tax Policy and the Economy, Volume 17*, NBER Chapters, National Bureau of Economic Research, Inc, pp. 1–36.

| and Getzler, L. S. (2002). *Accountability , Ability and Disability: Gaming the System*. NBER Working Papers 9307, National Bureau of Economic Research, Inc.

Green, F., Machin, S., Murphy, R. and Zhu, Y. (2010). *The Changing Economic Advantage from Private School*. IZA Discussion Papers 5018, Institute for the Study of Labor (IZA).

House of Commons Education and Skills Committee (2006). Special educational needs, third report of session 200506, volume i, hc 478i. House of Commons , http://www.publications.parliament.uk/pa/cm200506/cmselect/cmeduski/478/478i.pdf.

Imbens, G. and Kalyanaraman, K. (2009). *Optimal Bandwidth Choice for the Regression Discontinuity Estimator*. Working Paper 14726, National Bureau of Economic Research.

| and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, **142** (2), 615–635.

Jacob, B. A. and Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, **86** (1), 226–244.

Keslair, F., Maurin, E. and McNally, S. (2011). *Every Child Matters? An Evaluation of "Special Educational Needs" Programmes in England*. IZA Discussion Papers 6069, Institute for the Study of Labor (IZA).

Lee, D. S. and Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, **142** (2), 655–674.

| and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, **48** (2), 281–355.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, **142** (2), 698–714.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, **37** (1), 7–63.

QCDA (2010). Assessment of subjects in key stage 1 and key stage 2. Qualifications and Curriculum Development Agency, http://curriculum.qcda.gov.uk/.

Sartarelli, M. (2011). *Do Performance Targets Affect Behaviour? Evidence from Discontinuities in Test Scores in England*. DoQSS Working Papers 11-02, Department of Quantitative Social Science - Institute of Education, University of London.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, **51** (6), 309 – 317.

Trochim, W. (1984). *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. Beverly Hills, CA: Sage Publications.

Urquiola, M. and Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, **99** (1), 179–215.

Table 1: Institutional setting: the national school curriculum in England

(1) Primary/ Secondary	(2) Age	(3) Stage	(4) Year	(5) Assessment	(6) Expected achievement level
	3-4	Early Years Foundation Stage (EYFS)			
	4-5		Reception	Tests	6-9/13 elements
Primary School	5-6	Key Stage 1	1		
	6-7		2	Teacher assessments in English, Maths and Science (EMS)	2
	7-8	Key Stage 2	3		
	8-9		4		
	9-10		5		
	10-11		6	National and teacher assessments in EMS	4
Secondary School	11-12	Key Stage 3	7	Teacher assessments	
	12-13		8	Teacher assessments	
	13-14		9	Teacher assessments in EMS and foundation subjects	5 or 6
	14-15	Key Stage 4	10	Some children take GCSEs	
	15-16		11	Most children take GCSEs or other national qualifications	5 A*-C or equivalent including English and Maths

Notes:

i) The table illustrates the stages in which compulsory education is divided in England. Column (1) groups them into primary and secondary school. Column (2) shows the age range at each stage in column (3). Column (4) lists as a count each of the 11 years of schooling. Column (5) shows the type of assessment for students at the end of each stage and column (6) the expected achievement level that the Department for Education set for students and schools at each stage.

ii) Source: DirectGov (2010).

Table 2: Summary statistics by gender

Variable Names	All	Females	Males
Outcome variables			
SEN non-statemented	0.14	0.10	0.17
SEN statement	0.02	0.01	0.02
English additional language	0.08	0.08	0.08
Free school meals	0.15	0.15	0.14
Covariates: Key Stage 2 test scores			
Key stage 2 English score	59.93	62.20	57.68
S.d.	14.17	13.76	14.21
KS2 English teacher assessment level 2	0.02	0.02	0.03
KS2 English teacher assessment level 3	0.21	0.17	0.24
KS2 English teacher assessment level 4	0.50	0.50	0.51
KS2 English teacher assessment level 5	0.23	0.27	0.18
Key stage 2 Maths score	62.46	61.33	63.59
S.d.	20.53	20.38	20.62
KS2 Maths teacher assessment level 2	0.02	0.02	0.02
KS2 Maths teacher assessment level 3	0.20	0.20	0.19
KS2 Maths teacher assessment level 4	0.49	0.50	0.48
KS2 Maths teacher assessment level 5	0.25	0.24	0.27
Key stage 2 Science score	57.63	57.33	57.92
S.d.	12.50	12.50	12.48
KS2 Science teacher assessment level 2	0.01	0.01	0.01
KS2 Science teacher assessment level 3	0.13	0.13	0.13
KS2 Science teacher assessment level 4	0.53	0.54	0.53
KS2 Science teacher assessment level 5	0.29	0.28	0.30
S.d.	0.09	0.08	0.09
School type at Key Stage 2			
KS2 Community school	0.67	0.67	0.67
KS2 Voluntary aided school	0.18	0.19	0.18
KS2 Voluntary controlled school	0.10	0.10	0.10
KS2 Foundation school	0.03	0.03	0.03
KS2 Other schools	0.01	0.01	0.01
Gender and ethnicity			
Male	0.50	0.00	1.00
White	0.86	0.86	0.85
Black	0.03	0.03	0.03
Asian	0.06	0.06	0.06
Other	0.03	0.03	0.03
Observations	531,213	264,433	266,780

Note: The table shows summary statistics for the full sample and separately by gender of outcome variables and covariates in the regressions in the empirical analysis in section 4. The last row shows the number of observations. The outcome variables in the top panel are equal to one if a student is eligible for a certain government financial support program at school. *SEN non-statemented* is a dummy equal to one if a student is assessed as having minor special education needs, thus obtaining additional support by teachers at school, while *SEN non-statemented* is a dummy equal to one if a student is assessed as having greater needs. *Free School Meals (FSM)* is a dummy equal to one if a student gets a free meal at school based on multiple criteria about receipt of social benefits by parents. *English additional language (EAL)* is a dummy equal to one if a student who is not British native gets support in English. The second panel shows summary statistics of scores in English, Maths and Science tests at Key Stage 2, both the continuously measured test score and the share of students by categorical achievement level in tests, going from 2 to 5. The third panel shows shares of students by the type of school that they attended at Key Stage 2. The last panel panel shows the shares of students by gender and ethnicity. Section 2 offers additional information on the institutional setting of compulsory education in England and on the data.

Table 3: Probit and regression discontinuity estimates of the effect of meeting performance targets 3, 4 or 5 in test scores at Key Stage 2 on assignment to government interventions at Key Stage 3

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	All sample				Females				Males			
	Probit	RD 2-3	RD 3-4	RD 4-5	Probit	RD 2-3	RD 3-4	RD 4-5	Probit	RD 2-3	RD 3-4	RD 4-5
SEN non-statemented	-.24	.01	.002	-.002	-.22	.02	.004	-.001	-.26	.02	-.0006	-.003
	(.003)***	(.03)	(.007)	(.002)	(.004)***	(.04)	(.007)	(.004)	(.004)***	(.05)	(.009)	(.005)
Obs.	529951	92815	390445	437125	263800	43886	192975	219911	266151	48929	197470	217214
SEN statemented	-.03	.04	.0004	-.002	-.02	.04	-.0004	-.0005	-.04	.04	.001	-.003
	(.001)***	(.02)**	(.003)	(.0008)**	(.001)***	(.02)**	(.002)	(.001)	(.002)***	(.02)*	(.004)	(.002)
Obs.	529951	92815	390445	437125	263800	43886	192975	219911	266151	48929	197470	217214
EAL	-.01	.002	-.007	.002	-.01	.02	-.01	.002	-.01	-.01	.0006	.002
	(.0009)***	(.01)	(.003)**	(.002)	(.001)***	(.02)	(.004)***	(.003)	(.001)***	(.02)	(.004)	(.003)
Obs.	529951	92815	390445	437125	263800	43886	192975	219911	266151	48929	197470	217214
FSM	-.13	.01	-.009	-.001	-.13	-.04	-.01	-.002	-.12	.06	-.008	-.001
	(.002)***	(.03)	(.007)	(.004)	(.003)***	(.04)	(.009)	(.006)	(.003)***	(.04)	(.008)	(.006)
Obs.	529951	92815	390445	437125	263800	43886	192975	219911	266151	48929	197470	217214

Notes:
i) Estimates in the table are equal to the difference in the probability of assignment to a certain type of government support program column (1) for the students to the left and right of one among three targets $\bar{T}$: 3, 4 and 5 in test scores T that the Department for Education sets at Key Stage 2. The running variable is average fine grade test score. We estimate the probability by using smooth polynomials in test scores and separately for students to the left and right of a threshold in the running variable. We use a window that is centered at a target $\bar{T}$ and contains observations in the interval $[\bar{T}-1, \bar{T}+1]$. $\bar{T}-1$ is the threshold to the left of $\bar{T}$ and $\bar{T}+1$ is the threshold to the right of it. We obtain the bandwidth to estimate the polynomials by using the choice rule in Imbens and Kalyanaraman (2009). Estimates from Probit regressions are marginal effects that are computed at the mean value of the test score, by clustering standard errors at the school level. In all regressions we use as covariates the list in Table 2 which includes gender, dummies for ethnicity, proxies for socio-economic background, dummies for school types and scores in tests that are assessed by teachers. Section offers 3 offers additional information on the research design and section 4 on the results in the empirical analysis. The significance levels are as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
ii) The outcome variables are equal to one if a student is eligible for a certain government financial support program at school. *SEN non-statemented* is a dummy equal to one if a student is assessed as having minor special education needs, thus obtaining additional support by teachers at school, while *SEN non-statemented* is a dummy equal to one if a student is assessed as having greater needs. *Free School Meals (FSM)* is a dummy equal to one if a student gets a free meal at school based on multiple criteria about receipt of social benefits by parents. *English additional language (EAL)* is a dummy equal to one if a student who is not British native gets support in English. Summary statistics of the variables are in Table 2. Section 2 offers additional information on the institutional setting of compulsory education in England and on the data.

Table 4: Probit and regression discontinuity estimates of the effect performance targets 3, 4 or 5 in Key Stage 2 test scores on the value of characteristics that are determined before Key Stage 2

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	All sample			Females			Males		
	RD 2-3	RD 3-4	RD 4-5	RD 2-3	RD 3-4	RD 4-5	RD 2-3	RD 3-4	RD 4-5
Gender and ethnicity									
Male	.07	.01	.02						
	(.03)**	(.008)	(.006)***						
Obs.	92815	390445	437125						
Ethnicity black	-.01	.003	-.0006	-.01	-.0009	.002	-.007	.008	-.003
	(.01)	(.003)	(.002)	(.02)	(.004)	(.003)	(.02)	(.004)*	(.003)
Obs.	92815	390445	437125	43886	192975	219911	48929	197470	217214
Ethnicity Asian	.01	-.0002	-.002	.01	-.0003	-.004	.01	-.0001	.001
	(.02)	(.004)	(.003)	(.03)	(.006)	(.004)	(.03)	(.006)	(.004)
Obs.	92815	390445	437125	43886	192975	219911	48929	197470	217214
Ethnicity Other	-.005	.004	.003	.01	.004	.002	-.02	.003	.004
	(.01)	(.003)	(.002)	(.02)	(.004)	(.003)	(.02)	(.004)	(.003)
Obs.	92815	390445	437125	43886	192975	219911	48929	197470	217214
School type at Key Stage 2									
Voluntary aided	.004	-.005	.005	.008	-.002	.01	-.005	-.008	-.0008
	(.02)	(.006)	(.005)	(.03)	(.009)	(.007)	(.03)	(.009)	(.008)
Obs.	92815	390445	437125	43886	192975	219911	48929	197470	217214
Voluntary controlled	.02	.003	.003	.04	-.006	.004	-.007	.01	.002
	(.02)	(.005)	(.004)	(.03)	(.007)	(.006)	(.03)	(.007)*	(.006)
Obs.	92815	390445	437125	43886	192975	219911	48929	197470	217214
Foundation	-.01	.0001	.0005	-.003	-.002	.004	-.02	.002	-.004
	(.009)	(.003)	(.002)	(.01)	(.004)	(.003)	(.01)	(.004)	(.003)
Obs.	92815	390445	437125	43886	192975	219911	48929	197470	217214

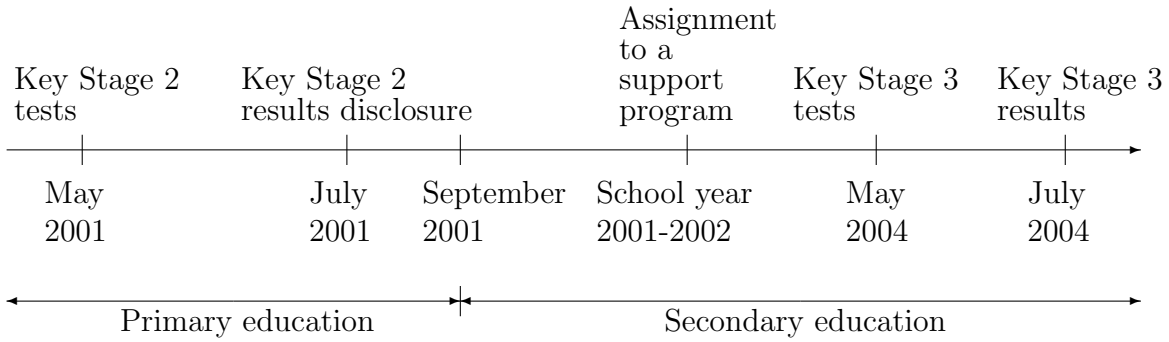
Note: The table shows estimates of the difference between students in the left neighbourhood of a threshold $\bar{T}$ in test score T and those in the right neighbourhood in probability of having a certain characteristic in column (1) that is pre-determined with respect to the disclosure date of test scores at Key Stage 2, e.g. ethnicity. The estimates are obtained separately for students to the left and right of one among three test score targets $\bar{T}$ equal to 3, 4 and 5 that are set by the Department for Education. We estimate the probability of a pre-determined characteristic by using smooth polynomials in test scores and separately for students to the left and right of a threshold. The running variable is the average fine grade average test score over the scores in English, Maths and Science. We use a window that is centered at $\bar{T}$ and contains observations in the interval $[\bar{T}-1, \bar{T}+1]$. $\bar{T}-1$ is the threshold to the left of $\bar{T}$ and $\bar{T}+1$ the threshold to the right of $\bar{T}$. We obtain the bandwidth to estimate the polynomials by using the choice rule in Imbens and Kalyanaraman (2009). Section 2 offers additional information on the institutional setting of compulsory education in England and on the data. Section 4 offers additional information on the empirical analysis.

Table 5: T-statistics of a null hypothesis test of no manipulation of the running variable fine grade test score at a threshold in the regression discontinuity design

Test	Full sample		
	3	4	5
English	5.32	14.40	17.81
Maths	5.27	10.97	7.42
Science	2.71	10.87	16.77
Average	1.26	0.54	0.09

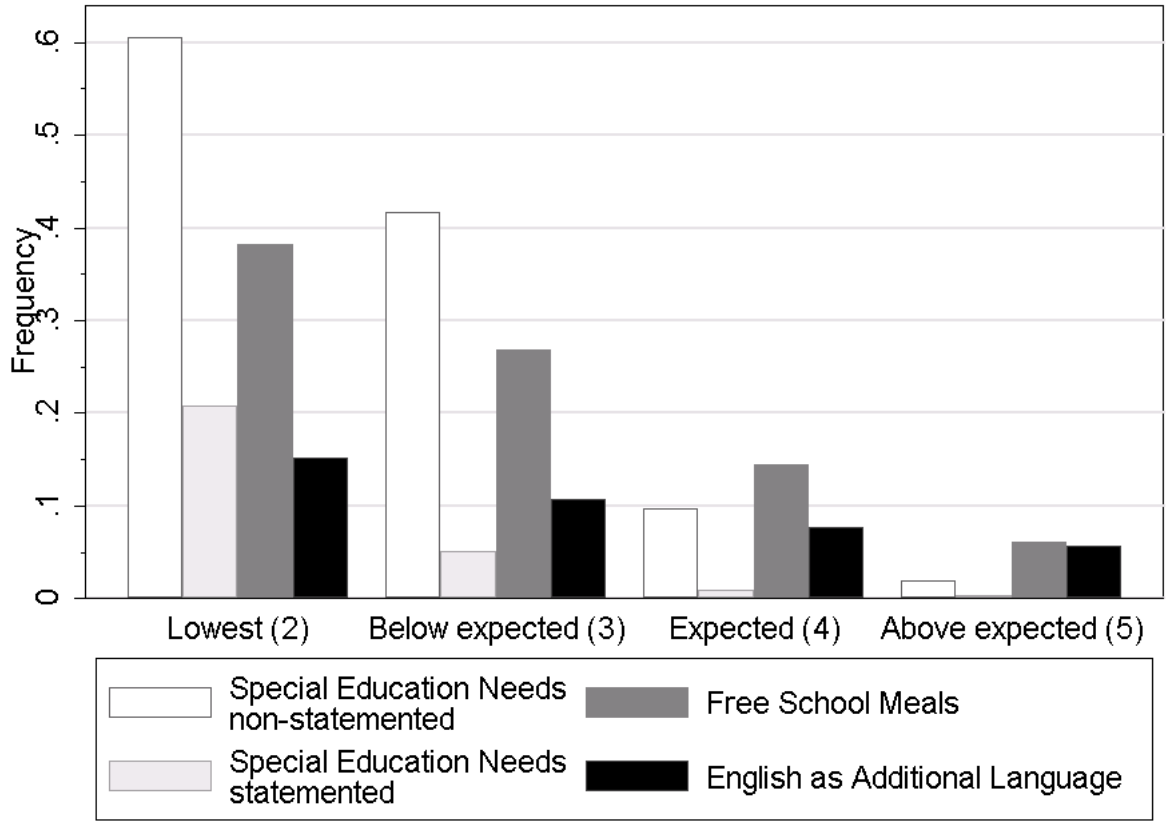
Notes: The table shows t-statistics of the test in McCrary (2008). Its null hypothesis is no manipulation of a running variable at threshold in the regression discontinuity regressions to estimate the effect of meeting performance targets 3, 4 and 5 in test scores on the assignment of students government financial support program at school. Four different fine grade test scores at Key Stage 2 are used as running variables, one test score for each test: English, Maths and Science, and the average test score over all tests. The table shows along columns thresholds and t-statistics. The first three rows show t-statistics for the null hypothesis of no manipulation of test scores in each of subject: English, Maths and Science. The last row shows t-statistics of tests of the fine grade average test score. The test in McCrary (2008) does not reject the null hypothesis if the difference in the probability mass points which is estimated as the height of undersmoothed histogram bins estimated separately for observations to the left and to the right of a threshold is sufficiently small.

Figure 1: Illustration of the timeline of tests at Key Stage 2 and assignment of students to a government support program at school after the disclosure of results in tests



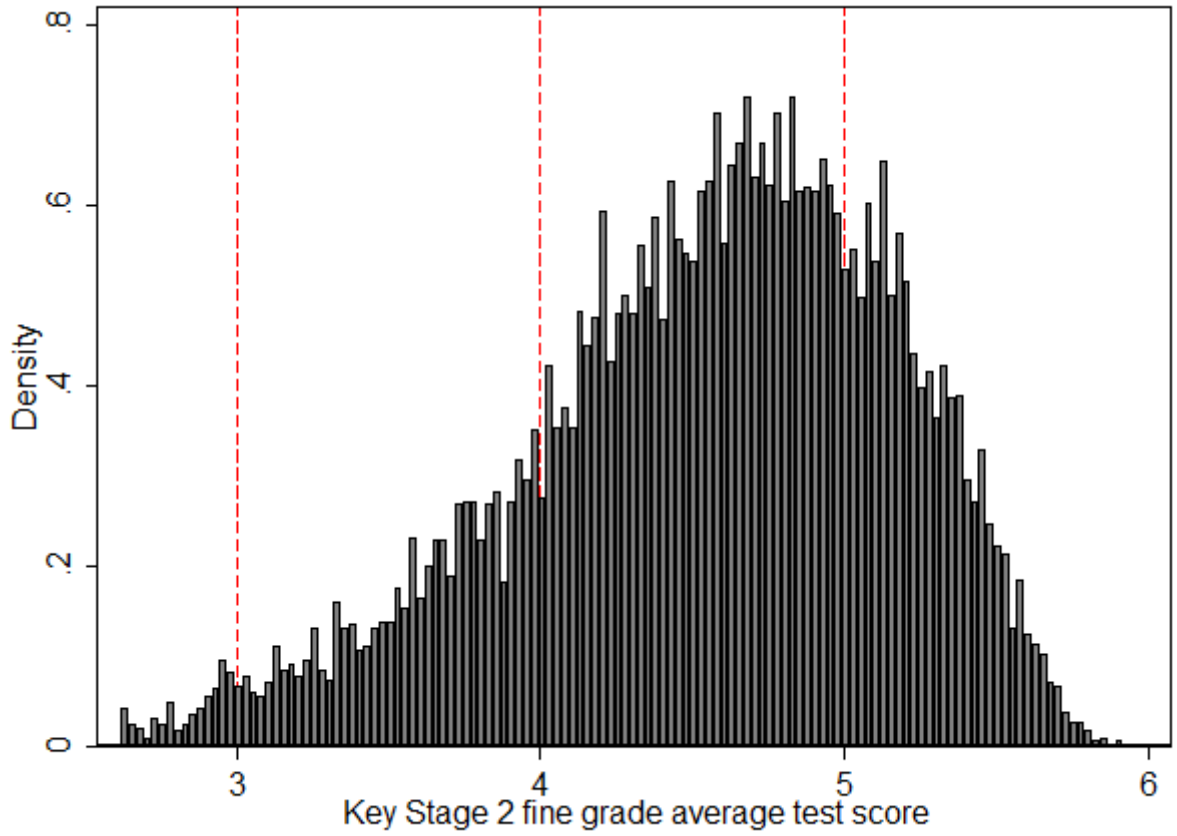
Note: The figure shows the timeline of the events and decisions that students face. A student sits Key Stage 2 tests in May 2001. Test scripts are marked externally and the achievement level is disclosed to students by July 2001. Students start secondary school with Key Stage 3 in September 2001. Assignment of students to a government support program at school among Special Education Needs, English as Additional Language or Free School Meals occurs within 2 years of starting secondary school. Key Stage 3 tests are held in May 2004. Section 2 offers additional information on the institutional setting of compulsory education in England and on the data.

Figure 2: Bar chart of shares of students in different government financial support programs at Key Stage 3 by achievement level in the average fine grade test score at Key Stage 2



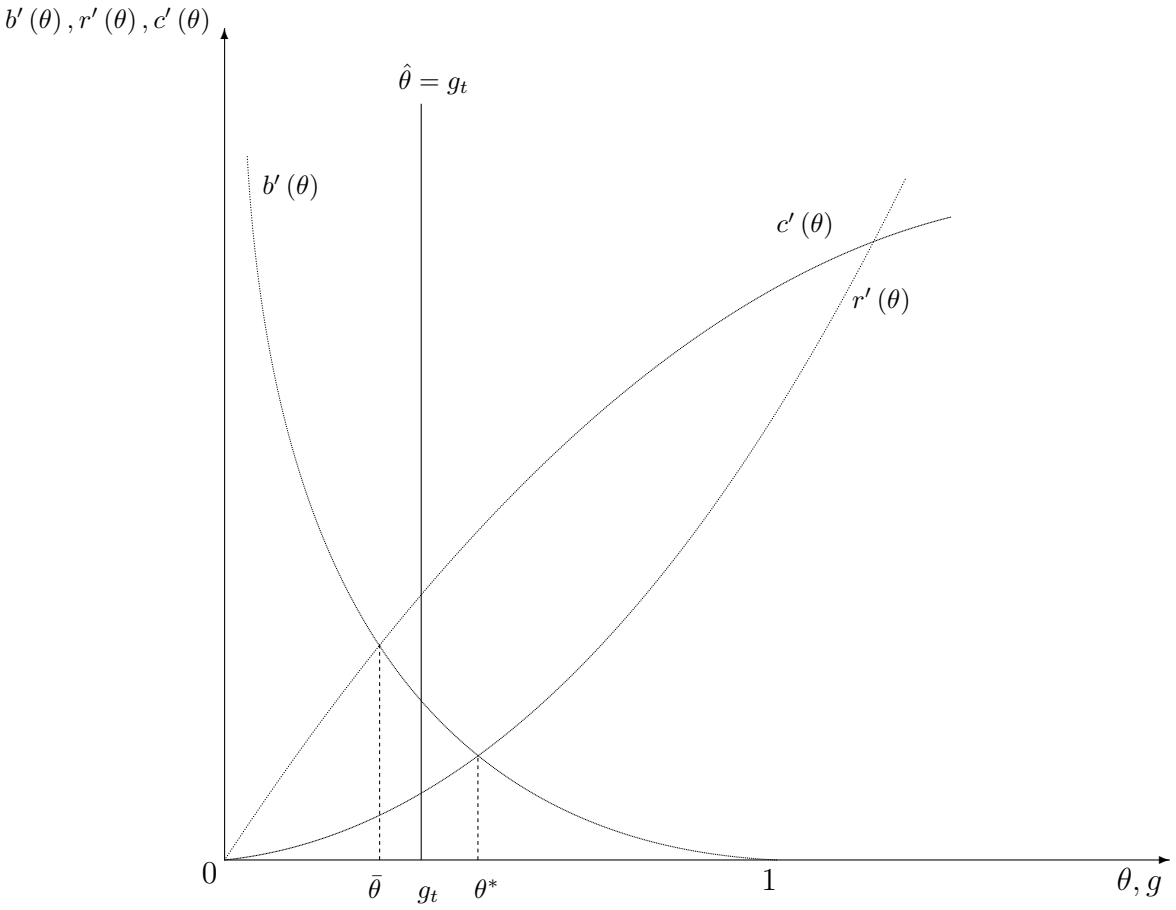
Note: The figure shows the shares of students in different government financial support programs by categorical achievement level from 2 to 5 in tests at Key Stage 2. Each level is defined by using thresholds 3, 4 and 5 in the fine grade average test score at Key Stage 2. The outcome variables are equal to one if a student is eligible for a certain government financial support program at school. *SEN non-statemented* is a dummy equal to one if a student is assessed as having minor special education needs, thus obtaining additional support by teachers at school, while *SEN non-statemented* is a dummy equal to one if a student is assessed as having greater needs. *Free School Meals (FSM)* is a dummy equal to one if a student gets a free meal at school based on multiple criteria about receipt of social benefits by parents. *English additional language (EAL)* is a dummy equal to one if a student who is not British native gets support in English. Summary statistics of the variables are in Table 2. Section 2 offers additional information on the institutional setting of compulsory education in England and on the data.

Figure 3: Undersmoothed histogram of the running variable fine grade average test score and performance targets



Notes: The figure shows an undersmoothed histogram of the fine grade average test score with bin size equal to 0.025. It also shows thresholds or targets in test scores as vertical and dashed lines at values 3, 4 and 5 on the horizontal axis. The fine grade average test score gives an average measure of test score at Key Stage 2 as a decimal number that can take values in the interval $[2.5, 6]$. The plot offers graphical evidence to assess the validity of the regression discontinuity design. Sorting at a threshold may occur if students, schools or teachers with certain characteristics benefit from scoring to the left or right of it, thus invalidating the research design. Visual inspection of the size of the bins at each threshold to assess whether they differ sensibly on either side of a allows one to assess the extent of sorting. Section 4.2 offers additional information about robustness checks to assess the validity of the regression discontinuity design.

Figure 4: Equilibrium in the demand for additional teaching support



Notes: The figure shows the equilibrium level in the demand by a school for additional teaching. It illustrates the case in which (i) F is a uniform distribution between 0 and 1 so that we can rewrite the benefit and cost functions as $B = b(\hat{\theta}), C = c(\hat{\theta})$ and $R = r(\hat{\theta})$, and (ii) the performance target is in between the optimal level of additional teaching, where extra teaching is provided to students with ability smaller or equal to $\bar{\theta}$ and the level where the maximum level of ability of a student receiving it is θ^*. Proposition 1 shows that the equilibrium level is reached at the performance target level $\hat{\theta} = g_t$. Section 5 offers additional information about the characteristics of the theoretical model that we use to interpret the empirical results in section 4.