Department of Quantitative Social Science

Do Performance Targets Affect Behaviour? Evidence from Discontinuities in Test Scores in England

Marcello Sartarelli

DoQSS Working Paper No. 11-02
March 2011

Leading education and social research
Institute of Education
University of London

DISCLAIMER

Any opinions expressed here are those of the author(s) and not those of the Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science. Institute of Education, University of London. 20 Bedford way, London WC1H 0AL, UK.

Do Performance Targets Affect Behaviour? Evidence from Discontinuities in Test Scores in England

Marcello Sartarelli[‡]

Abstract. Performance targets are ubiquitous in all areas of an individual's life such as education, jobs, sport competitions and charity donations. In this paper I assess whether meeting performance targets in tests at school has an effect on students' subsequent behaviour. This is helpful to test whether motivation and effort by students, parents and schools that the targets may induce, contribute to explain observed behaviour. I address potentially spurious correlations between test scores and behaviour by exploiting a regression discontinuity design in tests and a linked dataset of test scores and subsequent behaviour by students in compulsory education in England. I find that meeting a target that the government sets for students at age 11 has an insignificant effect on outcomes such as the probability of absence from school or of a police warning. I also find that meeting other targets for high and low ability students decreases the probability of being bullied by up to 34% with respect to the mean probability of such outcomes. The effects are heterogeneous as they vary by gender, parents' education level and type of behaviour. Overall, the research design offers a valuable test to assess unintended consequences that meeting the target or failing to meet it may lead to. The lack of a significant effect of targets on suspension and expulsion from school, as well as police warnings, suggests no adverse behavioural effect of performance targets, which is reassuring evidence on the design of tests in compulsory education. By using Probit estimates, one would conclude that meeting a target has an impact on behaviour. Regression discontinuity estimates show instead an insignificant effect at the expected target and a significant one at other targets for certain outcomes, although smaller than Probit estimates.

JEL classification: C21, I20, I21, I28.

Keywords: Absence, bullying, education, performance targets, police warning, regression discontinuity, suspension, test scores.

---

[*]Department of Quantitative Social Science, Institute of Education, University of London, 20 Bedford Way London, WC1H 0AL. E-mail: (m.sartarelli@ioe.ac.uk)

[†]Thanks for helpful comments to Peter Backus, Lorraine Dearden, Christian Dustmann, John Micklewright, Jeff Smith, Anna Vignoles, seminar participants at the Institute of Education, IMT Lucca, Lancaster University, Rutgers University, Università di Cagliari, the 2nd Meeting of the Danish Microeconometric Society, the 2009 NASM of the Econometric Society, and to the Department for Education for access to the data. This research is funded by the ADMIN Node of National Centre for Research Methods (ESRC grant RES-576-25-0014). This paper was previously circulated as "Do achievement labels affect the well-being of children? Evidence from discontinuities in test scores".

1 Introduction

Performance targets are important in helping individuals to build human capital or signalling ability in education and in a job. However, they may have unintended consequences. For example, rewarding individuals only if they perform above the average level may increase the average performance in a school or a firm although it may also change individuals' beliefs about their ability by overstating true ability for those meeting a target and viceversa for those missing the target. In addition, high ability individuals may exert little effort as the payoff is not proportional to one's performance or effort. This may occur in any agency relationship in which an individual's effort is not typically observed and may result in suboptimal effort provision and outcome in a production process. Potentially low motivation or performance may lead to actions by parents or teachers to help students, and by employers to help their employees. Failing this, policy interventions may help to rectify incentives and performance.[1]

Performance targets in employment contracts have been widely explored.[2] Instead little is known about performance targets as incentives for students. If a student fails to meet an important performance target in test scores, this may decrease the motivation for studying. On the contrary, meeting the target may instead make the student more motivated than before to study and obtain a high test score in the future. Self-confidence and beliefs about ability are additional channels through which feedback can foster positive behaviour, such as leading a healthy and safe life in youth as well as in adulthood. For example, an unauthorised absence from school one day may lead to no consequence in the future. Alternatively, the student may engage in risky activities and, perhaps, be caught by the police. Understanding whether meeting performance targets in education has an impact on students' behaviour is helpful to inform policy decisions about education, such as the design of school curricula.

In this paper I study whether meeting performance targets in tests has an effect on subsequent behaviour by students. I set out to answer this empirically by using linked administrative data from compulsory education in England on test scores and survey data on behavioural outcomes, such as the probability of absence from school or of a police warning. However, actions by parents and schools may confound the effect of meeting performance targets if, for instance, parents with a high education level make more effort than less educated parents in helping students to prepare for tests and in influencing their behaviour before and after tests are held.

Students whose ability appears not to be far from the mean value may concentrate in the test preparation on meeting the expected target that the Department for Education sets for the test, that is the focus in this paper and that is described in Section 2. For students with high or low ability, the Department sets proportional targets that are however "implicit" or of lower importance with respect to meeting the expected one. I identify the effect of meeting a performance target in test scores, with respect to missing it, on students' behaviour by exploiting discontinuities in test scores that the expected and implicit targets in test scores offer. The target is absolute rather than relative as scores are not normalised and targets are set before the distribution of students' test scores in one year is known. Thanks to this research design, I can tease out the effect of confounders that influence students' test scores and behaviour, such as parents' or teachers' effort, as they can only imperfectly influence students' scores in the neighbourhood of a target threshold. However, parents can influence the behaviour of students also after tests by, for example, inducing students to study harder

[1] See Stiglitz (2000) for a review of contributions in information economics to overcome informational failure in such settings as employer-employee contracts as well as in welfare-to-work programs for unemployed workers.

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

if they missed a test or, similarly, inducing them to carry out extra activities beyond school. In addition, parents with different characteristics, e.g. education level, may influence the behaviour of similar students differently. I will verify whether these or similar influences by parents are relevant to interpret the results in the empirical analysis.

Recent studies show evidence of the determinants of a variety of behavioural outcomes of young individuals that are similar to the ones that I use in this paper. Foliano *et al.* (2010) find that an increase in a proxy for value added by schools increases the disengagement in school by students in compulsory education in the UK. In related research Gibbons *et al.* (2010) find that neighbours' characteristics, such as the socio-economic composition and labour market opportunities, have a positive but insignificant effect on test scores, while the sign of the effect on behavioural outcomes such as general attitudes towards schooling and substance use is mixed. In the USA Reback (2010) finds that school counsellors decrease the probability of behavioural problems in elementary schools. Gaviria and Raphael (2001) show evidence of peer effects in alcohol and drugs use in high schools. Dee (2004) finds that additional schooling has a positive effect on proxies for civic behaviour such as voting and reading newspapers.[3]

Recent studies also show evidence of the effect of effort and achievement on achievement in the future. De Fraja *et al.* (2010) model as a game the simultaneous effort by students in studying and by parents and schools in supervising them to maximise students' achievement. They obtain a positive correlation between effort and achievement at school in equilibrium and find empirical support for it by using household data in the UK. Bandiera *et al.* (2009) instead study the effect of disclosing information about performance in tests on future performance in master degrees in the UK by exploiting variation in rules on performance feedback by university. They find a positive effect of feedback on performance in the empirical analysis. Similarly, Azmat and Iriberri (2009) study the relative performance feedback at school and find a positive effect on test scores of giving feedback to students on the distance of their scores from the average score in their class by using a natural experiment in a high school in Spain. Similarly, in this paper I use variation in performance in tests at school that a randomisation of test scores in the neighbourhood of discontinuities in scores.

The literature in economics claims that the effect of incentives on performance is positive, while the literature in psychology suggests the opposite as the two strands in the literature make different assumptions about the determinants of individuals' motivation. Benabou and Tirole (2002, 2003) reconcile the mixed evidence by studying the effect of incentives on motivation with a principal-agent model. Its predictions are that the effect is negative if the agent cares about the principal's beliefs and it is positive otherwise, because the agent may think that the principal has a bad opinion of her if a principal asks for high effort. This intuition offers an interpretation of how motivation and beliefs may influence students' behaviour in this paper.

The recent increase in interest by policy-makers in the role of education in influencing students' wider behaviour in society confirms the pressing need for additional knowledge on the impact of performance targets on behaviour. In 2001 the Department for Education in the USA funded "No Child Left Behind" (NCLB). This is a multi-billion dollar policy initiative to study the determinants of test score gaps among students of schooling age and prevent adverse effects in adulthood for those left behind at school. It helps to address such policy issues as low employment and wage profiles over time as well as health problems that lagging

[3]Grossman (2006) reviews the literature on the positive non-market returns to education in the long-term by focusing on such outcomes in adulthood as consumption patterns, health, fertility, child quality or well-being. Similarly, Oreopoulos and Salvanes (2009) find positive effects of education on such measures of well-being as health, marriage, parenting, trust and social ties, and a negative effect on risky behaviour in the USA.

behind at school may lead to in adulthood.[4] Since 2003 the Department for Education in the UK has funded "Every child matters", which is a similar policy initiative to NCLB but it puts additional emphasis on well-being and fostering positive behaviour in children. Hence, evidence in this paper on the impact of targets in tests for students on their behaviour may be of interest to policy-makers who deal with education and public policies for children. The reason is that most policy papers have focused on such determinants of children's education outcomes as parents' education and income.[5] In contrast, little is known about the effect of targets in education on behaviour.

In the empirical analysis I find that in simple Probit regressions an increase in the average test score by one unit is associated with a decrease in the probability of an unauthorised absence from school by 3 percentage points or 22% with respect to the mean probability of this outcome. Estimates also show that a unit increase in test scores is associated with a reduction in the probability of being bullied by 9 percentage points or 21%. Similarly, estimates of differences in the mean probability of these behavioural outcomes for subsamples of students who met a target versus students who did not meet it are negative, although not statistically significant.

Regression discontinuity estimates show instead that the effect of meeting the expected performance target set by the Department for Education on behavioural outcomes tends not to be significant. However, for low and high ability students, the effect of meeting the other (implicit) targets versus not meeting them appears to decrease the probability of being bullied by 15 and 4 percentage points respectively, or 34% and 9%. I discuss these results in section 5 to reconcile the mixed signs and significance. The effect of meeting targets for high ability students also varies by gender, e.g. it decreases the probability of being bullied by 7 percentage points or 14% for females. Finally, the effect varies by parents' education, e.g. it decreases the probability of being bullied by 7 percentage point or 16%, only for high ability students whose parents have a low qualification or none while.

This paper offers a novel contribution to the literature on the effects of achievement in test scores by testing whether achieving a target in test scores in compulsory education, versus missing the target, has an effect on students' subsequent behaviour. A similar test on the link between targets in test scores and behaviour can be repeated in the future to inform policy decisions. This can be done by using similar linked administrative and survey data in the UK, as well as in other countries whose governments collect rich data on students' achievement at school and on their behaviour, such as the National Longitudinal Survey of Youth in the USA.

The structure of the rest of the paper is as follows. Section 2 describes the institutional setting and the data of students in compulsory education in England. This sets the ground for the research design in section 3 and the empirical analysis in section 4. Finally, section 5 discusses the results and concludes.

2 Institutional setting and data

In this section I describe the source of exogenous variation in test scores that identifies the effect of meeting performance targets in tests in primary school on students' behaviour in secondary

[4] See Hastings and Weinstein (2007) for the evaluation of experimental policies in the No Child Left Behind program.

[5] Currie and Moretti (2003) and Acemoglu and Pischke (2001) show evidence of the effect of parents' socio-economic background on children's education in the USA while Chevalier and Lanot (2002) show similar evidence in the UK.

school and estimate this effect by using data on students in state schools in England.[6] In this country there are 11 years of compulsory education, which is divided into the Foundation Stage Profiles plus 4 Key Stages, as Table 1 shows. It starts at age 3-4 with the Foundation Stage Profile. Primary school starts at age 5-6 with Key Stage 1 and it is followed by Key Stage 2, as columns (1)-(3) show.[7] 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 interpret a student's progress throughout compulsory education.[8]

I use two linked datasets in the empirical analysis.[9] The first is the National Pupil Database (NPD) which is an administrative dataset with information on test scores of all students in England. It also contains information from the Pupil Level Annual School Census about the ethnicity of students, whether they are eligible for Free School Meals (FSM), the English as Additional Language (EAL) program or the Special Educational Needs (SEN) program. EAL and SEN provide additional support at school to students who meet the eligibility criteria.[10] The second dataset is the Longitudinal Study of Young People in England (LSYPE). This is a longitudinal survey of young people in England who are drawn from the cohort of students taking Key Stage 2 tests in 2001. The students in the survey are born between September 1989 and August 1990 and they are representative of the cohort of test-takers in 2001, whose test scores are recorded in the NPD dataset. The first wave of the survey was held in 2004 when students were 13-14 years old, in Year 9, and attending Key Stage 3 in secondary school. The survey contains information on the education and behaviour of students as well as the education, employment status and work experience of their parents. The sample size of the first wave is 15,770 students.[11]

Table 2 shows summary statistics of variables that measure students' behaviour and test scores for the full sample and for subsamples by gender. The sample size of the dataset used in the empirical analysis is equal to approximately 13,500. It is smaller than that of the full sample in the survey because variables from the survey that I use as outcome variables to measure behaviour suffer from item non-response. The second panel in Table 2 shows the incidence of missingness on these variables, which varies from 5% to 10% and I discuss this further in section 2.2.

2.1 Key Stage 2 tests

Students sit compulsory tests in English, Maths and Science when they are 10 or 11 years old, in year 6, which is the last one in Key Stage 2. The tests are set by the Qualifications

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

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

[8] 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".

[9] The datasets are linked by using the identification number of students, thus leading to a negligible loss in observations due to the linkage.

[10] The government determines eligibility for FSM status based on multiple criteria about receipt of social benefits by parents. Instead teachers and psychologists determine eligibility for EAL and SEN status based on criteria that the Department for Education set out and that allow for discretionary assignment as some of the criteria are subjective.

[11] Additional information about the survey design is available in NatCen (2010).

and Curriculum Development Agency (QCDA), which is an independent authority from the Department for Education.[12] Students are also assessed by their teachers before results in the Key Stage 2 tests are known.

2.1.1 Tests and scores

External examiners rather than teachers 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. For example, in the year in question, 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.[13] The Department for Education turns raw test scores into "fine grade" test scores. These are decimal numbers in the range 2-6 and 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 I use the *fine grade* point score rather than the raw test score as it is advantageous to obtain precise inference.[14]

Examiners know the thresholds for each achievement level when they mark tests. However, the institutional setting of compulsory education in England ensures imprecise manipulation of test scores in each subject and hence of the average test score that is computed over tests in English, Maths and Science at Key Stage 2, henceforth the "fine grade average test score". First, the QCDA is independent from schools and administers the training of examiners who mark test scripts and who do not know students and viceversa. This rules out perfect manipulation of test scores as examiners have no information about students and their behaviour. Second, one examiner gets all test scripts in one type of test, e.g. English, in a school. Third, 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 of the average test score. Finally, external grading of tests and no monetary incentives for teachers' performance do not make efforts by teachers a serious concern for the research strategy.[15]

2.1.2 Targets and disclosure of test scores

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. Critically for the research design in this paper, the underlying test scores are not disclosed, as the results sheet that schools use to communicate test results to students

[12]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.

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

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

[15]Wilson (2004) shows some evidence of responses by teachers to incentives as an increase in test scores by students in a school with respect to their past achievement may increase future enrolment in the school. Average school performance to inform school choice in compulsory education in England has been disclosed since 1992 using value added models (Ray (2010)). In the USA instead Eberts *et al.* (2002) and Ladd and Walsh (2002) show evidence of the effect of monetary incentives to teachers on test scores by students.

and parents in Figure 2 shows.[16] For example, two students whose fine grade score in the Maths test is 3.03 and 3.97 get level 3 in Maths. 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 very unlikely achieve the expected target at Key Stage 2, level 4, or viceversa so high that they will score considerably above that target. This leads to three sharp discontinuities in the "treatment" (meeting a performance target) in tests in English, Maths or Science because the probability that a student meets a target in, for example, Maths, goes sharply from zero if he scored less than 4, e.g. 3.97, to one if he scored 4 or above, e.g. 4.05.

I will also exploit discontinuities in the fine grade average test score, which is the average of test scores over the three subjects. It is the best available proxy to measure the overall performance by a student in all tests at Key Stage 2 and is also not disclosed to students or parents.[17] Table 3 shows in each cell percentages of students with a certain achievement level in tests in English, Maths and Science at Key Stage 2. For example, the cell in the top row and in column (6) shows that 4.42% of students has scored level 3 in English and level 4 in both Maths and Science. These students have marginally failed the expected achievement level 4 in only one out of three tests and their average fine grade point score would be equal to 3.9 if their test score in English were only a few points below the expected target. Similarly, the cell in the second row and in column (6) shows that 23.66% of students achieved level 4 in all three tests. These students have met the expected achievement level in all tests and their average fine grade point score is equal to 4. In addition, the first row in the table shows that most of the students who score level 3 in English, thus failing to meet the target in this subject, meet at most the expected target, level 4, in Maths and/or in Science. Only 0.04% instead score above the target, i.e. they achieve level 5 in these other two subjects, as column (10) shows. This offers empirical support to the assumption that students focus in the test preparation on meeting the expected target, level 4, in all tests, rather than on scoring above the expected target, level 5, in two tests and put less effort in the third test, in which they may score at level 3. This leads to a fuzzy discontinuity because students, who met the performance target at level 4 on average, may have failed to obtain a triplet of 4s in the tests at Key Stage 2, although these students are very few as Table 3 shows.[18]

Figure 1 shows the empirical density of the fine grade average test score and it also shows the targets as vertical dashed lines. The support of this variable is censored because 1.6% students in the linked dataset get 2.5 as total score as Figure 1 shows. This is the minimum score that a student gets regardless of how poor is the performance in the tests. I exclude students whose fine grade average test score is equal to 2.5 from the sample that I use in the empirical analysis because such censoring is of negligible relevance when focusing on the test scores of students in a small neighbourhood of thresholds. In addition, the density and hence the number of observations are lower around the implicit target for low ability students than around the other targets.

[16] 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.

[17] The Department for Education uses the average fine grade point score as an input to compute value added in schools. See Ray (2010) for additional information about value added in compulsory education in England.

[18] See Lee and Lemieux (2010) for additional information about sharp and fuzzy regression discontinuity designs.

2.2 Survey evidence of students' behaviour

A section in the questionnaire of the LSYPE survey asks questions about a child's behaviour at school and beyond to the main parent. This is defined as "the parent most involved in the young person's education" (NatCen (2009)). In the empirical analysis I use as outcomes the following proxies for behaviour: absence, suspension and expulsion from school, being bullied, and police warnings to children's parents. Each outcome variable is binary: it is equal to one if a student's main parent answered "yes" to a question on the behaviour of her child and zero otherwise. Table 4 lists in column (1) the outcome variables. Column (2) shows the wording of the questions underlying each variable. Column (3) shows the number of years prior to the survey interview date in which students' behaviour may have occurred. I also use an outcome variable on self-reported unauthorised absence by students, that is also known as truancy. These data offer proxies for behaviour that is representative of a cohort of young people in England. What allows me to assess the effect of meeting a performance target on behaviour is that answers to the questionnaire in wave 1 refer to events that occur after the disclosure of the results in tests at Key Stage 2 in July 2001.

Figure 3 describes the timing of the events from Key Stage 2 tests onwards. Students sat tests at age 10-11 in May 2001 and got tests results in July 2001. In September 2001 they started secondary school with Key Stage 3. March to October 2004 is the time period in which the survey data was collected via face to face interviews with main parents and students. The questions to the main parent on students' absence and whether they were bullied are about behaviour that occurred up to a year earlier, i.e. between March and October 2003. The questions on suspension and expulsion from school and on a police warning are instead about behaviour that occurred up to three years before the survey and up to three months before the disclosure of tests results in July 2001. I will perform robustness checks in the empirical analysis in section 4.3 to deal with potential reverse causality.

The top panel in Table 2 shows the full list of outcome variables and their summary statistics. 14% of students self-report to have been truant, i.e. absent at least one day from school without authorisation. 43% of main parents report that students were bullied, 10% that students were suspended, and 7 % that the police warned them about their children's actions.[19] These variables are proxies for behaviour by students which may capture their learning process about positive behaviour, as well as about potential threats to their education and health beyond what they learn at school and from their parents.

Variables in the survey suffer from non-response which can be due to a number of reasons including refusal to answer, inability to self-complete the questionnaire and ignorance about the answer. The number of observations in the sample that I use in the empirical analysis varies by outcome variable under the assumption of no selective missingness in the outcome variable on either side of a threshold in test score. This is an identifying assumption of the research design. I discuss it in section 4.3 and robustness checks do not reject missingness at random in the outcome variable, as the percentages of missing observations to the left and to the right of a threshold are not significantly different. Answers by main parents about students' behaviour offer the advantage of a smaller recall error with respect to surveying students. The reason is that students are 13-14 years old at the time of the survey and answer questions are about events up to three years earlier than the interview date, when their recollection of past events may be imprecise. A dummy that is equal to one if a student is bullied and zero otherwise is a proxy for a latent behaviour rather than a decision by a student. One may

[19]Parents also report information about expulsions from a school. However, the low variation in suspension with a mean probability equal to 0.006 makes this behavioural outcome little informative to learn about the determinants of students' behaviour.

expect no effect of just meeting a performance target in test scores on the probability of being bullied with respect to a similar student who has just missed the target. A non-zero effect may instead suggest statistical discrimination. This is because certain students may react to meeting a performance at school in a variety of ways which may lead other students to bully them for their reactions rather than for their achievement in tests.[20] Parents can influence these measures of behaviour by, for example, emphasising in various ways the importance of going to school regularly and being friendly to other students. This emphasises that parents' influence on students' behaviour can be very high on such measures of behaviour as absence while it may be lower on others such as being bullied. Evidence in the summary statistics that nearly one in two students are bullied suggests that parents may over-report whether their child is bullied. This may occur if parents mistakenly conclude that difficulties in socialising that children may experience at some point are in fact episodes in which their children are bullied.

Figure 4 and Table 5 show means of the outcome variables by achievement level. The probability of different measures of behaviour for students achieving levels 2 and 3 is the greatest among all achievement levels at Key Stage 2. It then decreases with the achievement level. However, certain measures of behaviour as, for example, unauthorised absence and being bullied show low variation between achievement levels. This emphasises the importance of assessing whether these correlations between performance targets and behaviour also have a causal interpretation.

3 Research design

I assess the effect of a meeting the performance target at Key Stage 2 on measures of students' behaviour after the disclosure of tests results. I use as outcome variable a dummy B to measure students' behaviour. B is equal to one if two to three years after Key Stage 2 tests a student, for example, played truant and zero otherwise. The outcome variable can be interpreted as a binary indicator $B = I\{B^* < \bar{B^*}\}$ to describe a student's behaviour that is driven by the unobservable motivation or beliefs B^*, for example, about the importance of school if one considers the outcome absence.

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

Consider in equation (1) a Probit regression of B on a continuous measure of test score T, the fine grade score of each test at Key Stage 2 or the average test score over all tests, 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, truant due to a unit increase in the fine grade test score. 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 behaviour by students with different performances in tests. However, unobservable ability of children, parental care and school arrangements that the error term in equation (1) contains may lead to a spurious correlation between test scores and behaviour if unobservables correlate with test scores.

$$P \quad = \quad I\{T \geq \bar{T}\} \tag{2}$$

$$B \quad = \quad f(T) + \beta_{RD} P + U \tag{3}$$

[20] Anderson *et al.* (2006) survey different sources of statistical discrimination in laboratory experiments which are conducted in classrooms and eye colour, subjective preferences for colours or redistribution of pennies determine group divisions.

Using instead information on students who marginally met an achievement target, e.g. obtain a test score greater or equal to 4, with respect to students who marginally failed to meet it, e.g. obtain a score smaller than 4, helps to identify the effect of meeting a performance target with respect to not meeting it on behaviour by using a a regression discontinuity (RD) research design.[21] This is the interpretation of the parameter β_{RD} in equation (3). P is equal to one if a student's test score T is greater or equal than a performance target $\bar{T}$, and zero otherwise as equation (2) shows. β_{RD} is, for example, negative if students who met a performance target ($P = 1$) are also less likely to play truant than those who did not meet it ($P = 0$), as meeting the target may have increased their motivation to go to school. Three thresholds $\bar{T}$ equal to 3, 4 or 5 in test scores determine whether a student meets a performance target P in equation (2). The expected performance target that the Department for Education sets for all students at Key Stage 2 is 4. By exploiting this target 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 the disclosure of the tests 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. For example, students may prepare for tests by focusing on test topics that will be in a test with high probability. Second, obtaining a test score to the left of a threshold or target (control group) or to the right of it (treatment group) can arguably be seen 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 I use test scores by subject in English, Maths and Science and also the average test score over all tests as running variables. 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.[22] 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 4.3 offer evidence that this assumption holds.

The discontinuities in tests score in English, Maths and Science are sharp as the probability of meeting the target in each test jumps from zero to one if a students scores just to the right with respect to just to the left of a threshold. This allows one to estimate the effect of meeting a target on behaviour by fitting in equation (3) smooth polynomials $f(T)$ in test scores T of a binary variable B equal to one, for example, if a student plays truant and zero otherwise. I estimate two smooth polynomials non-parametrically and separately for subsamples of students whose test score is smaller than the threshold $\bar{T}$ and for those whose score is greater than it. β_{RD} is estimated as the difference in the level of the polynomials at the threshold $\bar{T}$.

Conversely, the discontinuity in the average fine grade test score is fuzzy. Students who do not meet the performance target in at least one of the tests in English, Maths and Science, and would have a probability of receiving the treatment equal to zero in a sharp design, may with a probability between zero and one meet the expected target on average. For example, they may

[21] 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.

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

meet the target for high ability students, level 5, in English and Maths and fail to meet the expected target, level 4, in Science but still score 4 on average. One can recover the treatment effect by dividing the difference in behaviour between students to the right and to the left of a threshold by the share of students who meet the target at either side of the threshold. This is the Wald formulation of a treatment effect in an instrumental variable strategy which can be estimated by using Two Stage Least Squares (2SLS).[23] I estimate the specification in equations (2)-(3) using 2SLS for the fuzzy RD design and $f(T)$ is estimated by using polynomials in the distance of the test score T from the threshold $\bar{T}$ of up to the fourth order. The performance targets 3, 4 and 5 in test scores are thresholds at different percentiles of the distribution of the test score. I estimate the effect on behaviour of meeting a target on behaviour by using a window that is centered at $\bar{T}$ and contains observations 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 whose test score is equal to 3 or 4. 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. I choose the bandwidth of the 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 I also add as covariates in the polynomials pre-determined characteristics of students that are listed in Table 2.

4 Results

In section 4.1 I show three sets of estimates, one for each test, of the effect of meeting a performance target in a test on behaviour. In section 4.2 I show similar evidence but by using the fine grade average test score as measure of performance in tests.

4.1 Test scores by subject: English, Maths and Science

Table 6 shows estimated marginal effects in Probit regressions of a dummy for a behavioural outcome in column (1) on the fine grade test score in English, Maths and Science. In each column I estimate a different specification, starting with a baseline and ending with my preferred one. Details of each specification are in the bottom panel in the table and Table 2 lists the covariates in the regressions. Excluded categories of multiple-valued discrete covariates in the regressions are: teacher assessment levels equal to 5 by test subject, white ethnicity, no special educational needs status and enrolment in Community schools at Key Stage 2. Negative and statistically significant coefficients suggest that test scores and positive behaviour by students are complements as increasing the fine grade average test score by one point is associated with a decrease, for example, in the probability of being bullied by 2-5 percentage points or 25-50% with respect to the mean probability of unauthorised absence. Table 7 shows Probit and RD estimates that are obtained by using the English test score as a measure of performance. Column (2) repeats the Probit estimates that are shown in column (5) in Table 6. An increase in the test score by a unit appears in general to have a negative and significant effect on several behavioural outcomes for the full sample in column (2) and the males subsample in column (6) while the effect is not significant for the females subsample in column (10). RD estimates show instead that meeting the expected target with respect to not meeting it has a significant and negative effect only on the probability of expulsion for the full sample in column (4) and

[23] Hahn *et al.* (2001) show the connection between the Wald estimator and the fuzzy RD design, as well as discussing its estimation.

results by gender subsamples are not statistically significant. Table 8 shows estimates that are obtained by using as the running variable scores in the Maths test. Neither Probit nor RD estimates tend to be significant. However, RD estimates show that the effect of meeting the expected target, level 4, increases the probability of a police warning for the full sample in column (4) and for the males subsample in column (12). Similarly, meeting the expected target increases the probability of being bullied for males. Finally, Table 9 shows estimates that are obtained by using scores in the Science test and neither Probit nor RD estimates tend to be significant. The positive effect of meeting the expected target, level 4, on the probability of unauthorised absence for females in column (8) is one of the few significant RD estimates. In addition, meeting the implicit target for high ability students has a positive and significant impact on the probability of being bullied for males in column (13).

The estimates offer evidence of the effect of meeting a performance target in one test at Key Stage 2 on behaviour. However, they may be biased as only one out of the three measures of achievement in tests is considered in the RD regression in equation (3). The remaining two measures of achievement are not controlled for and may differ for students to the left or to the right of a threshold, e.g. in English, thus potentially leading to a spurious estimate of the effect of meeting a target. For example, the estimated effect of meeting the expected achievement level in the Maths test on the probability of a police warning for males equals 10% in column (12) in Table 8. Table 5 shows that this is equal to the mean probability of a police warning among males, thus casting doubts on the econometric specification that includes only one out of three measures of ability. In addition, evidence in Table 13 does not exclude that the research design exploiting thresholds in the tests by subject is invalid due to suspicious jumps in the density of test scores at the threshold. This is an empirical challenge as the test scores by subject offer three proxies for ability, students receive achievement levels separately by each test and they are intended to meet the expected target in each test.

4.2 Average test score

The fine grade average test score over all tests allows one to consider all three proxies for ability in the data. This is a more suitable proxy to study the overall effect of meeting targets on behaviour. Estimates of differences in the mean probability of, for example absence, between students who score at the expected performance target 4 and students who score at the target 3 are not statistically different from zero as Table 10 shows. Such estimates are equal to the difference in the relevant histogram bars for the same variable in Figure 4. For example, estimates in column (3) in the table say that a student who scores at level 4 in Key Stage 2 tests is on average 7 percentage points less likely to be truant in the following one to two years with respect to a counterfactual student who scores at level 3. This is what one also obtains by eyeballing the difference in height of the histogram bar for unauthorised absence between achievement levels 4 and 3 in Figure 4. Such a difference is not statistically significant as the high p-value shows. Estimates for subsamples by gender are instead not statistically significant either, as columns (5)-(10) in Table 10 show.

However, Probit estimates of the effect on behaviour of an increase by one unit in test score may be spurious due to such confounders as the support that students get to prepare for tests by parents and teachers, as well as to develop a positive behaviour. The RD design instead allows one to disentangle at the discontinuity the direct effect of incentives that the expected targets induce on behaviour from the effect of confounders. The thresholds in the average fine grade test score, which is constructed by the Department for Education as an average of scores in tests in English, Maths and Science, are fuzzy because a student who fails to meet targets in at least one test can still meet the target on average. However, evidence in section

2.1 shows that this is empirically relevant for only about 1% of students. The challenge lies in interpreting the effect by considering the possible contributions to the observed students' behaviour by students themselves, teachers and parents.

4.2.1 Full sample

Table 11 shows estimates of the effect of meeting performance targets equal to 3, 4 and 5 on behaviour for the full sample of students and for subsamples by gender. The bottom row in each panel in the table shows the number of observations in windows of size two around a target that I use to estimate the polynomials. For example, column (3) shows estimates using a threshold equal to 3 in the fine grade average score. If students score below or above it, they achieve level 2 or 3 in the fine grade average test score at Key Stage 2. Meeting this achievement level decreases the probability of being bullied by 15 percentage points or 34% with respect to the mean probability of being bullied that Table 2 shows, but the effect is only significant at 10% level. However, meeting the target has no effect on other measures of behaviour. Instead column (4) in the table shows that meeting the target 4 which the Department for Education expects students to meet at Key Stage 2 has mixed effects on different behavioural outcomes but it is not statistically significant. Finally, meeting the target 5 for high ability students has a negative effect equal to 4 percentage points or 10% on the probability of being bullied but is only significant at the 10% level.

Overall, estimates in columns (2)-(4) in Table 11 show that few RD estimates are statistically significant, thus suggesting that the impact of meeting targets in tests on behaviour is low.[24] Insignificant but different from zero estimates may be economically significant but also imprecise due to high standard errors. Standard significance tests may be unfit to do inference in a design with multiple discontinuities. For example, the precision of estimates for the threshold 3 along the running variable may be influenced by the observations that are close to the threshold equal to 4 and no guidance from theory has been so far provided to the best of my knowledge.[25] In addition, the apparent effects of meeting performance targets differ by ability group as, for example, the negative effect on the probability of being bullied is greater in absolute value at threshold 3 than at 5. Finally, Probit estimates are smaller than RD ones at low ability and at high ability threshold and estimates of the differences in the mean probability are greater than statistically significant RD estimates. This suggests that Probit estimates tend to underestimate the effect of performance targets on behaviour for students with different ability while difference in the mean probability of behaviour tend to overestimate the effect of significant estimates.

4.2.2 Subsample by gender

Columns (6)-(9) in Table 11 show estimates of the effect of meeting a performance target for the subsample of females. The effect is negative and significant on the probability of being bullied at target 5 for high ability students and it is equal to 7 percentage points or 16%. Conversely, Probit estimates understate the effect for high ability students. Estimates of differences in the mean probability of being bullied instead overstate these effects. Finally, meeting targets has either an insignificant effect on the probability of other behavioural outcomes or

[24]Estimates from parametric regressions have the same sign and similar precision to the non-parametric ones in Table 11. Additional estimates on students' self-reported behavioural outcomes about alcohol, smoking cannabis, vandalism and fights are available upon request.

[25]See Lee and Card (2008) for corrections in inference in a RD design with an integer-valued running variable.

the point estimates exceed sample means in Table 2, thus suggesting a potential problem with the econometric specification or the sample size in the neighbourhood of a threshold. Columns (10)-(13) show estimates for the subsample of males. Probit estimates are negative and statistically significant. RD estimates instead show that meeting the expected performance target by the Department for Education for males has an insignificant effect on the probability of all behavioural outcomes, except that of a police warning which increases by 6 percentage points or 60% at the 10% significance level. Moreover, meeting the target 5 for high ability students has a negative and significant effect on the probability of unauthorised absence which is equal to 6 percentage points or 43%. In contrast, estimates of Probit regressions and differences in the mean probability of behavioural outcomes overstate the effect of meeting targets in test scores. These estimates give relevant information on the difference in the behavioural response to performance targets by gender that estimates obtained by using the full sample in columns (2)-(4) would hide. The probability of being bullied is responsive to meeting a performance target in tests only for females. Instead the probability of unauthorised absence and of a police warning, although significant at 10% level but not at 5%, are responsive to meeting the expected target only for males.

4.3 Robustness checks

I 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 a student's behaviour with three robustness checks. The first one tests the identifying assumption that the design is as if it was locally randomised at the thresholds. The second one tests whether estimates are sensitive to the definition of the variables in the empirical analysis. The third one assesses potential actions by students, parents and schools after the test scores are disclosed as this may be helpful to interpret the estimates.

4.3.1 Pre-determined characteristics and gaming around a threshold

A RD design to study the effect of performance targets is similar to the assignment of students to either side in the neighbourhood of targets 3, 4 and 5 in a coin-flip experiment. The research design is valid if two assumptions hold empirically as Lee and Lemieux (2010) suggest.

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. This is because such variables 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 males who score just to the left of a threshold is not different from the share of males scoring just to the right of it. Otherwise, the effect on behaviour that one attributes to meeting a performance target may be confounded by the correlation between gender and performance in tests, which invalidates the randomised design around a threshold. I estimate in Table 12 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 characteristics that are determined before tests at Key Stage 2 on the running variable at Key Stage 2 lead to no rejection of the null hypothesis of no significant difference in the value of pre-determined characteristics of students, thus supporting the validity of the design.

The second assumption is the inability of students, parents, test markers or schools to perfectly manipulate test scores. This holds if before test scores are disclosed students, parents or

schools can at best imperfectly guess whether they have scored to the left or right of a target in any subject or manipulate it. Marking of test scripts by external examiners rather than by students' teachers ensures local randomisation of test scores in the neighbourhood of a threshold.[26] I test empirically the absence of manipulation in test scores by plotting an undersmoothed histogram of the fine grade average test score in Figure 1. I use a binwidth equal to 0.025 to obtain histogram bins such that they contain an arbitrarily small number of students separately to the left and right of a threshold and no bin contains the threshold value. Visual inspection of such bins suggests no suspicious jumps in the empirical density around thresholds. Moreover, McCrary (2008) developed a formal test of the null hypothesis of no manipulation. This 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 13 shows 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 that mark tests in 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.

4.4 Definition of the running variable and the thresholds

The research design is valid if the estimates are not sensitive to changes in the characteristics that contribute to define the fine grade average test score, which is the running variable in the research design. Similarly, meeting falsified performance targets, rather than the ones that the Department for Education sets, should have no effect on behaviour.[27]

Firstly, I estimate the effect on behaviour of meeting a threshold $\bar{T}$ by using a window that is centered at $\bar{T}$ and contains observations in a different interval than $[\bar{T}-1, \bar{T}+1]$ in the preferred specification that I use to estimate the smooth polynomials. Table 14 shows that estimates which I obtain for each target by altering the size of the window of observations, e.g. a window of size 2.05-3.95 in column (3) versus one of size 2-4 in the preferred specification in column (2), are similar to those in the preferred specification in columns (2), (6) and (10).

Secondly, I estimate the effect of meeting a performance target on the probability of suspension from school and on police warning by using subsamples of observations which vary by the date in which the survey data were collected. This is to test the sensitivity of estimates to potential reverse causality because the time window for survey questions on police warnings and suspension from school is up to three years before the interview date. It is the period April 2001 to October 2004 while tests scores are disclosed in July 2001 and a police visit, for example, in May 2001 occurs before Key Stage 2 tests. I create the subsamples by only

[26] 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.

[27] I show estimates that I obtain by using the full sample of observations but similar ones hold for estimates that are obtained using subsamples by gender that are available upon request.

considering survey dates from a certain month, e.g. July, onwards. Estimates in Table 15 do not reverse the sign or significance level with respect to those in the preferred specification at the top of each panel in the table. This evidence supports the research design.

Finally, I test with a falsification exercise whether thresholds 3, 4 and 5 in the fine grade average test score are the only relevant ones to students, parents or schools and no other threshold such as 3.4 or 3.5 is. Table 16 shows in column (14) that, for example, the effect of meeting the target 5 is significant on the probability of being bullied in the preferred specification, while it is not for falsified targets such as 4.5 or 5.1 in the neighbourhood of 5. Very few significant estimates, and at 10% significance level for decimal thresholds greater than thresholds 4 and 5, support the validity of the research design.[28]

4.5 Actions by students, parents and schools after the disclosure of tests results

Behavioural outcomes by students are observed one to three years after the disclosure of results in tests at Key Stage 2. In this section I test two assumptions about the behaviour of parents and schools after tests results are disclosed.

The first assumption is that missingness in the answers by main parents to survey questions is at random in a small neighbourhood around a threshold. It holds, for example, if parents in poor households feel as comfortable to disclose certain information as those in rich households do. To test this, I estimate polynomials in test scores of binary variables equal to one in the event of item non-response by a main parent and zero otherwise. I obtain estimates separately for students who are to the left and right of a threshold and I assess the significance of the difference in the value of the polynomials at the threshold. Similarly, the assumption holds if the choice by students and parents of a certain type of secondary school, e.g. Voluntary Aided, is arguably random in a small neighbourhood around a threshold. Table 17 shows insignificant estimates of regression discontinuity regressions of dummies for non-response,in the top panel in the table, and for secondary school types, in the bottom panel, on the running variable at Key Stage 2, thus leading to no rejection of the null hypothesis of no difference around a threshold. This excludes selective non-response about students' behaviour and differential responses by type of school as possible interpretations of the estimates.

The second assumption is that the behaviour of students after the test results are disclosed does not vary with characteristics of parents and schools that, however, were not different at either side of the threshold in test scores, as section 4.3.1 shows. Potential intervention by parents with different characteristics, such as their education level, is informative to interpret the estimates of the effect of meeting performance targets on behaviour. This may occur, for example, if students in a small neighbourhood to the left of the threshold are persuaded by parents and teachers that they missed a target due to bad luck and may, as a consequence, behave more similarly to students who met the target than they would, had they not learnt this. I test whether parents who completed compulsory education or have a higher education level may react more or less about their children's potential behavioural response to test scores than parents who did not complete compulsory education. Table 18 shows estimates of the effect of meeting performance targets on behaviour for the full sample of students in columns (2)-(4) and separately for subsamples of students whose main parent did or did not complete at least compulsory education in columns (5)-(7) and columns (8)-(10). The magnitude of

[28]Certain estimates are significantly different from zero for decimal thresholds smaller than 4 although their magnitude is greater than means in summary statistics. This suggests future work to study whether the intuition of the choice rule to compute the optimal bandwidth of the polynomial in Imbens and Kalyanaraman (2009) holds in a design with multiple thresholds along the running variable.

the effects on all outcomes tends to be greater for students whose parents did not complete compulsory education than for students whose parents did. For example, the estimates of the effect of meeting any performance target on the probability of being bullied are twice to ten times greater for students whose parents did not complete compulsory education.

5 Discussion

In this paper I assess whether meeting absolute performance targets in tests at school has an effect on subsequent behaviour by students. In the empirical analysis, proxies for the behaviour of students in compulsory education in state schools in England include unauthorised absence from school and police warnings to parents about a student's behaviour. Simple Probit regressions show that the apparent effect of an increase in test scores decreases the probability of absence from school or of a police warning. However, if such regressions are misspecified due to unobservables that are correlated with test scores, the estimates are biased and may lead to inaccurate policy decisions. By teasing out the effect of confounders, such as parents' education, with a regression discontinuity design that thresholds in test scores offer, I find instead that meeting the performance target that the Department for Education expects students to meet at age 11 has an insignificant effect on their behavioural outcomes. However, I also find a negative and significant effect of meeting lower and higher implicit performance targets on the probability of being bullied for low and high ability students.

The effect of meeting the expected target on any proxy for behaviour is negative but not significant for subsamples by gender, while meeting targets for low and high ability females decreases the probability of being bullied, and meeting the target for high ability males also decreases the probability of unauthorised absence. Finally, the effect varies by parents' education, with significant estimates for high ability students whose parents have a low qualification or none. Probit regressions tend to underestimate the effect of implicit targets for low and high ability students as a Probit specification overlooks the institutional setting of targets in test scores. Similarly, estimates of differences in the mean probability of behaviour, separately for students to the left and right of a target, tend to overestimate the effect as they cannot tease out the effect on outcomes of unobservable characteristics that correlate with test scores.

Several reason may help to explain why the effect of the expected target on students' behaviour is not statistically insignificant. Firstly, if the Department for Education sets targets for students and schools that are not considered relevant, this leads to no behavioural response by students who may ignore their performance with respect to the target and by schools that may focus on other measures of students' achievement. Secondly, parents and teachers may persuade students who have just missed a target that they are similar to those who met it, thus attenuating the behavioural response by students. This is because students' behaviour is observed after information on whether students met performance targets is disclosed to students and parents. Evidence about discontinuities at the expected achievement target does not rule out empirically these as possible mechanisms linking achievement in test scores and behaviour. Finally, the number of observations in a small neighbourhood of each of the three performance targets in the empirical analysis varies from approximately 400 to 3,000, which may influence the precision of the RD estimates.

Estimates that are obtained by using subsamples by gender and parents' education level show negative and significant effects, at 10% and 5% level, of meeting performance targets on the probability of being bullied and of police warning. However, the lack of empirical relevance of meeting targets in achievement on such outcomes as suspension and expulsion from school as well as a police warning is reassuring for policy-makers as explicit and implicit performance

targets for students and teachers suggest no major behavioural implication that would require an intervention by schools or the government. What it may suggest instead is caution in the use of certain survey information. Parents may have over-reported about whether their children were bullied in the light of the high incidence of bullying at 46% in the summary statistics. This variable may in fact capture a latent attitude for socialising by students, as well as parents' expectations about the behaviour of the children. In addition, a positive but marginally significant effect of the expected target on a police warning only for males may help parents and schools to study its determinants inside the school and outside, which they may have not learnt about this from estimates using the full sample. Heterogeneous effects by main parent's education level with lower behavioural responses to performance targets by children of more literate parents offer some empirical support for a tradeoff between nature and nurture in students' development as Lizzeri and Siniscalchi (2008) and the literature on economics, genetics and sociology also suggest. Moreover, effects vary by the type of control which a parent has over a student's behavioural outcome. Parents have more control over such activities as absence from school than over bullying, which offers additional support to this tradeoff.

The regression discontinuity design also suggests that using a categorical measurement of performance to assess whether students meet performance targets in tests offers policy makers a test to empirically assess behavioural effects of targets in test score in the future and, hence, to inform policy. Reassuringly, the empirical evidence shows that performance targets tend to have an insignificant effect on behaviour, thus suggesting no unintended effects on students. Additional knowledge on the impact of incentives on over-motivation or under-motivation in individuals would be useful to inform future policy in education and beyond, and also reconcile the contrasting results on incentives and motivation in Economics and Psychology that Benabou and Tirole (2002, 2003) survey. Future research will focus on the role of performance targets in education by assessing whether they have an effect on students' achievement and behaviour in secondary school and beyond.

References

Acemoglu, D. and Pischke, J. S. (2001). Changes in the wage structure, family income, and children's education. *European Economic Review*, **45** (4-6), 890–904.

Anderson, L., Fryer, R. and Holt, C. (2006). Discrimination: experimental evidence from psychology and economics. In W. Rogers (ed.), *Handbook on Economics of Discrimination*, *4*, Edward Elgar, pp. 97–118.

Azmat, G. and Iriberri, N. (2009). *The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students*. CEP Discussion Papers dp0915, Centre for Economic Performance, LSE.

Bandiera, O., Larcinese, V. and Rasul, I. (2009). *Blissful Ignorance? Evidence From a Natural Experiment on The Effect of Individual Feedback on Performance*. Policy Research Working Paper Series 4122, University College London.

Benabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, **117** (3), 871–915.

— and — (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, **70** (3), 489–520.

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.

Chevalier, A. and Lanot, G. (2002). The relative effect of family characteristics and financial situation on educational achievement. *Education Economics*, **10** (2), 165–181.

Currie, J. and Moretti, E. (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *Quarterly Journal of Economics*, **118** (4), 1495–1532.

De Fraja, G., Oliveira, T. and Zanchi, L. (2010). Must try harder: Evaluating the role of effort in educational attainment. *The Review of Economics and Statistics*, **92** (3), 577–597.

Dee, T. S. (2004). Are there civic returns to education? *Journal of Public Economics*, **88** (9-10), 1697–1720.

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

Eberts, R., Hollenbeck, K. and Stone, J. (2002). Teacher performance incentives and student outcomes. *Journal of Human Resources*, **37** (4), 913–927.

Foliano, F., Meschi, E. and Vignoles, A. (2010). *Why do children become disengaged from school?* DoQSS Working Papers 10-06, Department of Quantitative Social Science - Institute of Education, University of London.

Gaviria, A. and Raphael, S. (2001). School-based peer effects and juvenile behavior. *The Review of Economics and Statistics*, **83** (2), 257–268.

Gibbons, S., Silva, O. and Weinhardt, F. (2010). *Do Neighbours Affect Teenage Outcomes? Evidence from Neighbourhood Changes in England*. SERC Discussion Papers 0063, Spatial Economics Research Centre, LSE.

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).

Grossman, M. (2006). Education and nonmarket outcomes. *Handbook of the Economics of Education*, vol. 1, *10*, Elsevier, pp. 577–633.

Hahn, J., Todd, P. and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, **69** (1), 201–09.

Hastings, J. S. and Weinstein, J. M. (2007). *No Child Left Behind: Estimating the Impact on Choices and Student Outcomes*. Working Paper 13009, National Bureau of Economic Research.

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

Imbens, G. W. 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.

Ladd, H. F. and Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review*, **21** (1), 1–17.

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.

Lizzeri, A. and Siniscalchi, M. (2008). Parental guidance and supervised learning. *The Quarterly Journal of Economics*, **123** (3), 1161–1195.

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

NatCen (2009). Longitudinal study of young people in england: Wave one documentation. Study Number 5545, UK Data Archive, http://www.esds.ac.uk/.

— (2010). Longitudinal study of young people in england: User guide to the datasets: Wave one to wave six. Study Number 5545, UK Data Archive, http://www.esds.ac.uk/.

Oreopoulos, P. and Salvanes, K. G. (2009). *How large are returns to schooling? Hint: Money isn't everything*. Working Paper 15339, National Bureau of Economic Research.

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. http://curriculum.qcda.gov.uk/.

Ray, A. (2010). *School Value Added Measures in England: A Paper for the OECD Project on the Development of Value-Added Models in Education Systems*. Tech. rep., UK Department for Education.

Reback, R. (2010). Schools' mental health services and young children's emotions, behavior, and learning. *Journal of Policy Analysis and Management*, **29** (4), 698–725.

Stiglitz, J. E. (2000). The contributions of the economics of information to twentieth century economics. *The Quarterly Journal of Economics*, **115** (4), 1441–1478.

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.

Wilson, D. (2004). Which ranking? the impact of a 'value-added' measure of secondary school performance. *Public Money & Management*, **24** (1), 37–45.

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: The table illustrates the stages into 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. The compulsory school leaving exam is the General Certificate of Secondary Education (GCSE), which most students sit in year 11, when they are 15-16 years old. DirectGov (2010) offers additional information about the national school curriculum in England.

Table 2: Summary statistics by gender

Variable Names	All	Females	Males
Outcome variables (LSYPE)			
Unauthorised absence	0.14	0.14	0.14
Being bullied	0.43	0.45	0.39
Suspension	0.10	0.06	0.14
Expulsion	0.01	0.01	0.01
Police warning	0.07	0.05	0.10
Missingness in outcome variables			
Missing: unauthorised absence	0.06	0.06	0.06
Missing: being bullied	0.10	0.09	0.11
Missing: suspension	0.05	0.05	0.05
Missing: expulsion	0.05	0.05	0.05
Missing: police warning	0.06	0.05	0.06
Covariates: Key Stage 2 test scores (NPD)			
Fine grade average test score	4.57	4.58	4.55
S.d.	0.62	0.60	0.64
English fine grade test score	4.44	4.55	4.33
S.d.	0.73	0.68	0.75
Maths fine grade test score	4.45	4.41	4.49
S.d.	0.81	0.78	0.83
Science fine grade test score	4.74	4.73	4.75
S.d.	0.61	0.59	0.62
English teacher assessment level 2	0.04	0.02	0.06
English teacher assessment level 3	0.17	0.13	0.22
English teacher assessment level 4	0.43	0.39	0.48
English teacher assessment level 5	0.21	0.22	0.20
Maths teacher assessment level 2	0.03	0.02	0.04
Maths teacher assessment level 3	0.17	0.16	0.19
Maths teacher assessment level 4	0.42	0.39	0.44
Maths teacher assessment level 5	0.23	0.19	0.28
Science teacher assessment level 2	0.02	0.01	0.02
Science teacher assessment level 3	0.11	0.10	0.13
Science teacher assessment level 4	0.46	0.43	0.50
Science teacher assessment level 5	0.26	0.22	0.31
Covariates: Key Stage 2 school type			
Community school	0.60	0.54	0.67
Voluntary aided school	0.16	0.15	0.18
Voluntary controlled school	0.09	0.08	0.10
Foundation school	0.03	0.03	0.03

Continued on next page...

Continued from previous page...

Variable Names	All	Females	Males
Male	0.45	0.00	1.00
Covariates: ethnicity			
Asian	0.06	0.07	0.07
Black	0.03	0.03	0.02
Other	0.05	0.05	0.05
White	0.86	0.85	0.86
Covariates: socio-economic background			
Free school meals	0.13	0.11	0.14
English additional language	0.18	0.27	0.08
No Special Education Needs (SEN)	0.82	0.87	0.77
SEN statement	0.04	0.02	0.05
SEN non-statemented	0.14	0.11	0.18
Main parent has a degree	0.13	0.15	0.11
Main parent higher education	0.13	0.13	0.13
Main parent GCSE	0.45	0.44	0.46
Main parent other qualification	0.11	0.10	0.11
Main parent no qualification	0.18	0.18	0.18
Main parent's father has a degree	0.07	0.08	0.05
Total n. observations	15770	7727	8043

Notes:
i) 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. I use final survey weights in the LSYPE survey to obtain summary statistics that are representative of the cohort of students in the NPD administrative dataset. The weights correct for sampling at the student and school levels, as well as for non-response. See NatCen (2010) for additional information about the LSYPE survey design. The top two panels show summary statistics of the outcome variables in the LSYPE survey data and non-response. The remaining panels show summary statistics of students' characteristics in the NPD administrative dataset. In the last panel in the table, Free School Meals 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 is a dummy equal to one if a student who is not British native gets support in English. SEN is a dummy equal to one if a student obtains additional support by teachers at school. The last panel also shows shares of students' parents by education level.
ii) Outcome variables are in the LSYPE dataset and they are binary variables equal to one if a student's main parent answers yes to a question on the student's behaviour up to one to three years before the interview date and zero otherwise, with the exception of unauthorised absence which is self-reported by students. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension* is equal to one if a student has been suspended from school up to three years before the interview date. *Expulsion* is equal to one if a student has been suspended at school up to three before years before the interview date. *Police warning* is equal to one if the police contacted parents due to their child's behaviour up to three years before the interview date.
iii) The total number of observations in the last row in the table refers to the full LSYPE sample which is linked to the NPD administrative dataset. The number of observations in the empirical analysis is obtained by multiplying the total number of observations by the share of missing observations in an outcome variable due to non-response. For example, the number of observations for the binary variable that is equal to one if a student was bullied is 15770*(1-0.10)=14193. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.

Table 3: Percentages of students by achievement level in tests in English, Maths and Science at Key Stage 2

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	KS2 Maths level 3			KS2 Maths level 4			KS2 Maths level 5		
	KS2 Science level 3	KS2 Science level 4	KS2 Science level 5	KS2 Science level 3	KS2 Science level 4	KS2 Science level 5	KS2 Science level 3	KS2 Science level 4	KS2 Science level 5
KS2 English level 3	3.39	6.15	0.16	0.46	4.42	0.46	0.00	0.12	0.04
KS2 English level 4	1.07	8.50	0.33	0.48	23.66	7.23	0.01	3.32	5.71
KS2 English level 5	0.00	0.42	0.10	0.00	5.36	5.82	0.00	2.17	14.72

Notes: The table shows in each cell percentages of students with a certain achievement level in tests in English, Maths and Science at Key Stage 2 (KS2). For example, the cell in the top row and in column (6) shows that 4.42% of students has scored 3 in English and 4 in Maths and Science. These students have marginally failed the expected achievement level 4 in one test and their average fine grade point score is equal to 3.9 if their test score in English is only a few points below the expected target. Similarly, the cell in the second row and in column (6) shows that 23.66% of students achieved level 4 in English, Maths and Science tests. These students have just met the expected achievement level 4 in all tests test and their average fine grade point score is equal to 4. In addition, the first row in the table shows that the greatest percentage of students who score level 3 in English, thus failing to meet the target in this subject meets at most the expected targets in Maths and/or in Science in columns (2)-(6), while only 0.04% scores above the expected target, level 5, in Maths and Science. The table contains 98.6% of observations in the linked dataset while 1.4% of students who scored level 2, the lowest possible, in at least one test are excluded. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.

Table 4: Questions to main parent in the LSYPE questionnaire

(1) Variable Name	(2) Question	(3) No. of years before the survey date and in which behaviour may be observed
Unauthorised absence	In the last 12 months, have you ever played truant, that is missed school without permission, even if it was only for a half day or a single lesson?	Up to 1 year
Being bullied	The next question is about any bullying or other bad behaviour from other pupils at (his/her) school that you know have happened to (name of sample member) in the last 12 months. As far as you know, have any of these things happened to (name of sample member) at (his/her) school in the last 12 months? 1. Called names by other pupils at his/her school 2. Sent offensive or hurtful text messages or emails 3. Shut out from groups of other pupils or from joining in things 4. Made to give other pupils his or her money or belongings 5. Threatened by other pupils with being hit or kicked or with other violence 6. Actually being hit or kicked or attacked in any other way by other pupils 7. Any other sort of bullying 8. No, none of these things have happened in the last 12 months	Up to 1 year
Suspension	Has (name of sample member) been temporarily excluded, that is suspended, from a school for a time, in the past 3 years?	Up to 3 years
Expulsion	Has (name of sample member) been permanently excluded, that is expelled from school for good, in the past 3 years?	Up to 3 years
Police warning	Have the police got in touch with you (or your husband/ or your wife/ or your partner) about (name of sample member) because of something he/she had done in the last 3 years? 1. Yes , in last 3 years 2. No 3. Not in the last three years	Up to 3 years

Notes: The table lists in column (1) the names of the variables in the LSYPE dataset which I use as outcome variables in the empirical analysis. Column (2) shows the wording of the questions in the questionnaire of the LSYPE in Wave 1. Column (3) shows the number of years before the survey date and in which behaviour may be observed. All questions are answered by the student's main parent except the one on unauthorised which is answered by the student. Main parent is defined as "the parent most involved in the young persons education" in NatCen (2010), that also offers additional information about the survey design and the questionnaire. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.

Table 5: Summary statistics of outcome variables measuring students' behaviour by average achievement level in tests at Key Stage 2 and by gender

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	All sample				Females				Males			
	2	3	4	5	2	3	4	5	2	3	4	5
Unauthorised absence	0.19	0.22	0.15	0.08	0.14	0.19	0.15	0.09	0.23	0.25	0.14	0.07
Being bullied	0.71	0.52	0.44	0.34	0.87	0.57	0.48	0.39	0.57	0.49	0.39	0.30
Suspension	0.21	0.19	0.09	0.04	0.09	0.12	0.05	0.02	0.32	0.25	0.13	0.05
Expulsion	0.06	0.01	0.00	0.00	0.03	0.00	0.00	0.00	0.08	0.02	0.00	0.00
Police warning	0.12	0.13	0.08	0.03	0.06	0.09	0.05	0.02	0.17	0.15	0.10	0.04
N. observations	176	1894	6229	2855	79	880	3165	1395	97	1014	3064	1460

Notes:
i) The table shows in each cell the share of students by behavioural outcome along rows in column (1) and by average achievement level in tests at Key Stage 2 along columns. The shares in columns (2)-(5) are obtained by using all observations in the linked data. For example column (2) shows that 19% of students who achieved level 2 on average in all tests played truant. Instead columns (6)-(9) and columns (10)-(13) are obtained by using subsamples of females and males respectively. I use final survey weights in the LSYPE survey to obtain summary statistics that are representative of the cohort of students in the NPD administrative dataset. The weights correct for sampling at the student and school levels, as well as for non-response. See NatCen (2010) for additional information about the LSYPE survey design.
ii) The outcome variables are equal to one if a student's main parent answers yes to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension* is equal to one if a student has been suspended from school up to three before years before the interview date. *Expulsion* is equal to one if a student has been suspended at school up to three before years before the interview date. *Police warning* is equal to one if the police contacted parents due to their child's behaviour up to three years before the interview date. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.

Table 6: Marginal effects in a Probit regression of a binary variable equal to one in the event of a behavioural outcome and zero otherwise on the fine grade scores in tests in English, Maths and Science

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	English				Maths				Science			
Unauthorised absence	-.05	-.01	-.007	-.01	-.04	-.02	-.009	-.02	-.05	-.01	-.007	-.01
	(.004)***	(.008)	(.009)	(.008)	(.004)***	(.007)**	(.008)	(.007)**	(.005)***	(.008)	(.009)	(.008)
Obs.	10463	9909	9909	9909	10473	9916	9916	9916	10454	9899	9899	9899
Being bullied	-.08	-.05	-.05	-.05	-.09	-.07	-.06	-.07	-.08	-.02	-.02	-.02
	(.007)***	(.01)***	(.01)***	(.01)***	(.006)***	(.01)***	(.01)***	(.01)***	(.008)***	(.01)*	(.01)	(.01)*
Obs.	9592	9159	9159	9159	9596	9162	9162	9162	9586	9151	9151	9151
Suspension	-.06	-.02	-.02	-.02	-.05	-.01	-.004	-.01	-.06	-.02	-.01	-.02
	(.003)***	(.006)***	(.006)***	(.006)***	(.003)***	(.006)*	(.006)	(.005)**	(.004)***	(.006)***	(.006)*	(.006)***
Obs.	10166	9701	9701	9701	10169	9703	9703	9703	10157	9691	9691	9691
Expulsion	-.003	-.0007	-.0006	-.0007	-.003	.0002	.0000765	.0002	-.003	-.0002	-.0003	-.0002
	(.0005)***	(.0003)**	(.0004)	(.0003)**	(.0006)***	(.0005)	(.0005)	(.0005)	(.0007)***	(.0004)	(.0004)	(.0004)
Obs.	10183	9718	9718	9718	10187	9721	9721	9721	10174	9708	9708	9708
Police warning	-.04	-.01	-.02	-.01	-.03	-.01	-.01	-.01	-.03	-.01	-.01	-.01
	(.003)***	(.005)***	(.006)***	(.005)***	(.003)***	(.005)**	(.005)*	(.005)**	(.003)***	(.005)**	(.006)**	(.005)**
Obs.	10119	9660	9660	9660	10122	9663	9663	9663	10109	9650	9650	9650
Covariates		X	X	X		X	X	X		X	X	X
Survey weights			X	X			X	X			X	X
Clustered S.e.				X				X				X

Notes:
i) The table shows estimates of marginal effects in Probit regressions of dummies in column (1) that are equal to one in the event of a behavioural outcome and zero otherwise on the fine grade scores in tests in English, Maths and Science. In Column (2) the outcome variable is regressed on the fine grade test score in the English test. In Column (3) covariates from the NPD dataset are added. They include test scores, dummies for the type of school a student goes to and dummies for ethnicity and for socio-economic background such as main parent's education. In column (4) I estimate the regression by using survey weights. In column (5) I also cluster standard errors at the school level. Columns (6)-(9) and columns (10)-(13) show estimates of the same regressions but using test scores in Maths and Science respectively. Excluded categories of multiple-valued discrete covariates are: teacher assessment levels equal to 5 by test subject, white ethnicity, no special educational needs (SEN) status and enrolment in Community schools at Key Stage 2. The marginal effects are computed at the mean value of test scores. 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's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports absence with no authorisation from school at least once. *Being bullied* is equal to one if the student was bullied. *Suspension*, *Expulsion* and *Police warning* equal one if a student up to three before years before the interview date has been respectively suspended, expelled from school or if the police contacted parents due to their child's behaviour. 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 linked dataset.

Table 7: Estimates of the effect of meeting a performance target on the probability of behavioural outcomes by exploiting a jump from the left to the right of thresholds in the fine grade test score in the English test (Sharp Regression Discontinuity)

(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
Unauthorised absence	-.01	-.05	.03	-.02	-.0000409	-.06	.02	-.02	-.02	-.06	.03	-.03
	(.008)	(.06)	(.03)	(.02)	(.01)	(.10)	(.05)	(.03)	(.01)**	(.07)	(.03)	(.03)
Obs.	9909	2049	6546	7852	4893	776	3047	4110	5016	1273	3499	3742
Being bullied	-.05	.04	.03	.04	-.05	-.05	.09	.06	-.06	.14	-.02	.009
	(.01)***	(.16)	(.04)	(.04)	(.02)**	(.09)	(.06)*	(.06)	(.02)***	(.20)	(.05)	(.04)
Obs.	9159	1895	6005	7255	4537	709	2784	3820	4622	1186	3221	3435
Suspension	-.02	.04	.02	-.009	-.01	.28	.04	-.03	-.03	.05	.01	.01
	(.006)***	(.11)	(.02)	(.01)	(.007)*	(.29)	(.03)	(.02)*	(.01)***	(.15)	(.04)	(.03)
Obs.	9701	1991	6368	7700	4760	742	2931	4010	4941	1249	3437	3690
Expulsion	-.0007	.03	-.01	.002	-7.11e-07	.32	-.01	.002	-.001	.002	-.02	.001
	(.0003)**	(.03)	(.007)**	(.001)*	(8.44e-07)	(.21)	(.01)	(.001)*	(.0006)*	(.008)	(.009)*	(.001)
Obs.	9718	1998	6380	7710	2997	745	2932	4011	4954	1253	3448	3699
Police warning	-.01	.02	.006	-.003	-.004	.47	.04	-.02	-.03	-.03	-.02	.01
	(.005)***	(.04)	(.02)	(.02)	(.006)	(.19)**	(.03)	(.02)	(.009)***	(.12)	(.03)	(.03)
Obs.	9660	1977	6335	7674	4751	743	2922	4001	4909	1234	3413	3673

Notes:
i) Estimates in the table are equal to the difference in the probability of behavioural outcomes in 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 fine grade test score in English. I 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. I 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. I 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 using survey weights and clustering standard errors at the school level. In all regressions I 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's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension*, *Expulsion* and *Police warning* are equal to one if a student up to three years before the interview date has been respectively suspended, expelled from school or if the police contacted parents due to their child's behaviour. 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 linked dataset.

Table 8: Estimates of the effect of meeting a performance target on the probability of behavioural outcomes by exploiting a jump from the left to the right of thresholds in the fine grade test score in the Maths test (Sharp Regression Discontinuity)

(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
Unauthorised absence	-.02	.05	.005	-.02	-.02	-.30	-.02	-.007	-.01	.69	.03	-.03
	(.007)**	(.06)	(.02)	(.02)	(.01)*	(.23)	(.03)	(.03)	(.01)	(.17)***	(.04)	(.02)
Obs.	9916	2517	6845	7359	4896	1287	3515	3595	5020	1230	3330	3764
Being bullied	-.07	.11	.01	-.0007	-.07	-.06	-.05	.04	-.07	-.23	.14	-.03
	(.01)***	(.14)	(.03)	(.03)	(.02)***	(.12)	(.05)	(.04)	(.02)***	(.19)	(.06)**	(.04)
Obs.	9162	2304	6300	6823	4539	1168	3248	3358	4623	1136	3052	3465
Suspension	-.01	-.12	.02	.001	-.01	-.08	-.004	-.03	-.01	-.31	.05	.03
	(.006)*	(.06)**	(.02)	(.01)	(.006)*	(.13)	(.02)	(.02)	(.01)	(.17)*	(.04)	(.02)
Obs.	9703	2443	6661	7222	4762	1231	3416	3517	4941	1212	3245	3705
Expulsion	.0002	-.003	-.007	-.001	-1.16e-06	-.02	-.005	.0003	.0006	.02	-.004	-.004
	(.0005)	(.02)	(.006)	(.002)	(8.38e-07)	(.02)	(.005)	(.003)	(.0009)	(.01)*	(.006)	(.004)
Obs.	9721	2449	6673	7234	3005	1235	3418	3518	4954	1214	3255	3716
Police warning	-.01	-.14	.03	.003	-.01	-.51	-.03	-.001	-.01	-.07	.10	.01
	(.005)**	(.11)	(.02)*	(.01)	(.005)**	(.18)***	(.02)	(.02)	(.008)	(.07)	(.03)***	(.02)
Obs.	9663	2431	6631	7194	4754	1231	3410	3509	4909	1200	3221	3685

Notes:
i) Estimates in the table are equal to the difference in the probability of behavioural outcomes in 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 fine grade test score in Maths. I 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. I estimate the probability using by smooth polynomials in test scores and separately for students to the left and right of a threshold. I 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. I 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 using survey weights and clustering standard errors at the school level. In all regressions I 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's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension*, *Expulsion* and *Police warning* are equal to one if a student up to three years before the interview date has been respectively suspended, expelled from school or if the police contacted parents due to their child's behaviour. 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 linked dataset.

Table 9: Estimates of the effect of meeting a performance target on the probability of behavioural outcomes by exploiting a jump from the left to the right of thresholds in the fine grade test score in the Science test (Sharp Regression Discontinuity)

(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
Unauthorised absence	-.01	.24	.02	-.01	-.01	.18	.11	-.04	-.02	-.09	-.005	.007
	(.008)	(.09)***	(.03)	(.02)	(.01)	(.09)**	(.05)**	(.03)	(.01)	(.13)	(.05)	(.02)
Obs.	9899	936	6245	8931	4886	452	3113	4421	5013	484	3132	4510
Being bullied	-.02	-.50	-.02	.05	-.03	.27	.02	.002	-.01	.27	-.11	.10
	(.01)*	(.17)***	(.05)	(.03)	(.02)*	(.11)**	(.06)	(.03)	(.02)	(.12)**	(.08)	(.05)**
Obs.	9151	827	5724	8295	4533	396	2872	4125	4618	431	2852	4170
Suspension	-.02	.76	.02	.006	-.01	-.22	.05	.007	-.02	.05	-.008	.004
	(.006)***	(.56)	(.04)	(.01)	(.006)**	(.05)***	(.04)	(.02)	(.01)**	(.12)	(.06)	(.03)
Obs.	9691	875	6053	8783	4756	414	3008	4328	4935	461	3045	4455
Expulsion	-.0008	.10	.01	.0004	-.005	.005	-.005	.003	-.007	.14	.008	-.002
	(.001)	(.18)	(.006)	(.003)	(.002)**	(.05)	(.007)	(.003)	(.003)**	(.17)	(.01)	(.004)
Obs.	10057	862	5989	8701	4943	434	2979	4494	5114	450	3143	4623
Police warning	-.01	-.13	-.06	.03	-.01	.03	.01	.05	-.01	-.13	-.13	.02
	(.005)**	(.16)	(.03)**	(.02)**	(.005)**	(.04)	(.04)	(.02)***	(.009)	(.05)***	(.05)***	(.02)
Obs.	9650	868	6021	8749	4747	416	3002	4317	4903	452	3019	4432

Notes:
i) Estimates in the table are equal to the difference in the probability of behavioural outcomes in 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 fine grade test score in Science. I estimate the probability using by smooth polynomials in test scores and separately for students to the left and right of a threshold in the running variable. I estimate the probability using by smooth polynomials in test scores and separately for students to the left and right of a threshold. I 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. I 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 using survey weights and clustering standard errors at the school level. In all regressions I 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's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension*, *Expulsion* and *Police warning* are equal to one if a student up to three years before the interview date has been respectively suspended, expelled from school or if the police contacted parents due to their child's behaviour. 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 linked dataset.

Table 10: Difference in the probability of behavioural outcomes by students' average achievement level in tests at Key Stage 2 for the full sample of students and separately by gender

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	All sample			Females			Males		
	2-3	3-4	4-5	2-3	3-4	4-5	2-3	3-4	4-5
Unauthorised absence	-0.02	-0.07	-0.07	0.04	-0.06	-0.06	-0.06	-0.09	-0.08
P-value	0.69	1.00	1.00	0.16	1.00	1.00	0.92	1.00	1.00
Being bullied	-0.16	-0.09	-0.09	-0.21	-0.10	-0.10	-0.11	-0.09	-0.09
P-value	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
School suspension	-0.05	-0.11	-0.06	0.01	-0.07	-0.03	-0.10	-0.13	-0.10
P-value	0.93	1.00	1.00	0.37	1.00	1.00	0.98	1.00	1.00
School expulsion	-0.04	-0.01	-0.00	-0.02	-0.00	-0.00	-0.05	-0.01	-0.00
P-value	0.99	1.00	0.93	0.89	0.93	0.62	0.99	1.00	0.95
Police warning	-0.02	-0.05	-0.05	-0.01	-0.05	-0.03	-0.03	-0.05	-0.07
P-value	0.75	1.00	1.00	0.57	1.00	1.00	0.78	1.00	1.00

Notes:
i) The table shows in each cell the difference in the estimated probability of a student's behavioural outcome in column (1) between students who obtained an average achievement level to the left and to the right of a threshold. For example column (2) shows that students who overall achieved level 3 in tests are 2 percentage points less likely to be be absent than students achieving level 2. The difference in shares in columns (2)-(4) are obtained by using all observations in the linked data while columns (5)-(7) and columns (8)-(10) are obtained by using subsamples by gender. I use final survey weights in the LSYPE survey to obtain summary statistics that are representative of the cohort of students in the NPD administrative dataset. The weights correct for sampling at the student and school levels, as well as for non-response. See NatCen (2010) for additional information about the LSYPE survey design. *ii*) The outcome variables are equal to one if a student's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension* is equal to one if a student has been suspended from school up to three before years before the interview date. *Expulsion* is equal to one if a student has been suspended at school up to three before years before the interview date. *Police warning* is equal to one if the police contacted parents due to their child's behaviour up to three years before the interview date. 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 linked dataset.

Table 11: Estimates of the effect of meeting a performance target on the probability of behavioural outcomes by exploiting a jump from the left to the right of thresholds in the fine grade average test score (Fuzzy Regression Discontinuity)

(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
Unauthorised absence	-.07	.08	-.05	-.01	-.05	-.05	-.01	.03	-.09	.13	-.07	-.06
	(.007)***	(.07)	(.04)	(.02)	(.01)***	(.12)	(.04)	(.03)	(.01)***	(.09)	(.05)	(.02)***
Obs.	9858	1729	7119	8128	5154	853	3742	4300	5255	979	3784	4276
Being bullied	-.11	-.15	.04	-.04	-.12	-.19	.04	-.05	-.10	-.02	-.01	-.05
	(.009)***	(.08)*	(.03)	(.02)*	(.01)***	(.11)*	(.05)	(.04)	(.01)***	(.12)	(.05)	(.05)
Obs.	9550	1586	6552	7532	4737	758	3411	3978	4813	905	3446	3908
Suspension	-.10	.006	.03	-.02	-.07	.32	-.003	.004	-.13	-.02	.09	-.04
	(.007)***	(.06)	(.02)	(.02)	(.009)***	(.13)**	(.02)	(.01)	(.01)***	(.12)	(.04)**	(.03)
Obs.	9657	1676	6936	7980	4972	798	3590	4173	5148	960	3672	4188
Expulsion	-.008	-.004	-.0009	.004	-.006	.003	.005	.004	-.01	.04	-.01	.006
	(.002)***	(.04)	(.005)	(.002)*	(.002)**	(.02)	(.004)	(.004)	(.003)***	(.03)	(.01)	(.003)*
Obs.	9674	1679	6948	7994	4978	802	3593	4175	5159	959	3682	4200
Police warning	-.06	.10	.02	-.008	-.04	.15	-.04	.002	-.07	.08	.05	-.01
	(.006)***	(.06)	(.02)	(.01)	(.007)***	(.09)	(.03)	(.01)	(.009)***	(.07)	(.03)*	(.02)
Obs.	9615	1666	6898	7948	4962	798	3579	4163	5109	945	3639	4164

Notes:
i) The table shows estimates of the difference in the probability that a student's main parent answers "yes" to a question on the student's behaviour in column (1) for students to the left and right of a test score target that the Department for Education set at Key Stage 2. Regression discontinuity estimates in the table are the coefficient of a dummy equal to one if a test score T is greater or equal than a threshold $\bar{T}$ and zero otherwise in a Two Stage Least Squares regression in equations (2)-(3) of a dummy B to proxy a behavioural outcome on test score T. $f(T)$ in equation (3) is estimated by using polynomials in the distance of the test score T from the threshold $\bar{T}$ of up to the fourth order. The performance targets in test scores $\bar{T}$ at Key Stage 2 are set by the Department for Education and they are equal to 3, 4 and 5. The running variable is the fine grade average test score over the scores in English, Maths and Science. I obtain each estimate by using 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. I 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 using survey weights and clustering standard errors at the school level. In all regressions I 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's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension*, *Expulsion* and *Police warning* are equal to one if a student up to three years before the interview date has been respectively suspended, expelled from school or if the police contacted parents due to their child's behaviour. 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 linked dataset.

Table 12: Test of the difference in the value of characteristics that are determined before Key Stage 2 test scores are disclosed for students whose test scores are just to the left or to the right of targets 3, 4 or 5

(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
Male	.10	.02	.06	.	.	.	.	.	.
	(.08)	(.03)	(.03)*						
Test scores (Teacher Assessment)									
English level 3	-.05	.04	.003	-.06	.07	-.002	-.03	.0002	-.0000325
	(.08)	(.04)	(.007)	(.14)	(.06)	(.004)	(.10)	(.05)	(.01)
English level 4	-.02	-.03	.02	-.05	-.06	-.05	.	.007	.09
	(.02)	(.04)	(.03)	(.04)	(.06)	(.05)		(.05)	(.05)*
Maths level 3	-.18	-.06	-.003	-.39	-.05	-.005	-.08	-.07	-.001
	(.08)**	(.04)	(.003)	(.15)***	(.06)	(.005)	(.11)	(.05)	(.003)
Maths level 4	.01	.06	.07	.	.06	.10	.03	.06	.04
	(.02)	(.04)	(.04)*		(.05)	(.06)*	(.03)	(.05)	(.06)
Science level 3	.06	.01	-.001	-.04	.06	-.004	.11	-.03	.
	(.09)	(.04)	(.002)	(.13)	(.06)	(.003)	(.12)	(.05)	
Science level 4	.04	-.02	.02	.02	-.03	.008	.05	-.003	.03
	(.03)	(.04)	(.04)	(.04)	(.06)	(.05)	(.04)	(.05)	(.05)
School type at Key Stage 2									
Voluntary aided school	.05	-.01	.03	.10	.007	.05	.0005	-.02	.009
	(.05)	(.02)	(.02)	(.09)	(.03)	(.03)*	(.07)	(.03)	(.04)
Voluntary controlled school	-.02	-.003	.007	-.003	-.02	.02	-.03	.02	-.01
	(.04)	(.02)	(.02)	(.05)	(.02)	(.02)	(.06)	(.03)	(.02)
Foundation school	.007	-.002	.006	.02	.002	.03	.007	-.005	-.01
	(.007)	(.009)	(.009)	(.02)	(.01)	(.02)*	(.009)	(.01)	(.01)
Ethnicity									
Black	.07	-.01	.0005	.12	-.03	-.008	.04	.0002	.02
	(.05)	(.02)	(.02)	(.08)	(.03)	(.02)	(.05)	(.02)	(.02)
Asian	-.08	.03	-.0009	-.21	.02	.01	.001	.04	-.02
	(.08)	(.03)	(.02)	(.12)*	(.04)	(.03)	(.11)	(.03)	(.03)
Other	.005	.009	.01	-.04	.004	-.0006	.04	.01	.02
	(.04)	(.01)	(.01)	(.02)*	(.02)	(.02)	(.05)	(.02)	(.02)

Continued on next page...

Continued from previous page...

(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
Government interventions in schools									
SEN statement	.06	-.01	.002	.06	.007	-.002	.09	-.03	.01
	(.06)	(.01)	(.003)	(.14)	(.007)	(.001)	(.09)	(.02)	(.007)
SEN non-statemented	-.11	.07	.01	-.09	.10	-.002	-.18	.02	.03
	(.08)	(.04)*	(.01)	(.16)	(.04)**	(.01)	(.12)	(.05)	(.02)*
Free school meals	.05	.01	.008	.13	.03	.02	-.02	.001	.002
	(.08)	(.03)	(.02)	(.11)	(.04)	(.02)	(.11)	(.04)	(.03)
English additional language	-.08	.03	.004	-.09	.02	.008	-.06	.05	-.0003
	(.08)	(.03)	(.02)	(.12)	(.04)	(.03)	(.11)	(.04)	(.03)
Main parent (MP)									
MP with a degree	.008	.02	.06	.04	.03	.06	-.0001	.008	.04
	(.02)	(.02)	(.02)**	(.02)**	(.02)	(.03)*	(.03)	(.02)	(.03)
MP higher education	-.09	.03	-.02	-.01	.03	-.06	-.16	.02	.03
	(.05)*	(.02)	(.02)	(.05)	(.03)	(.04)	(.07)**	(.02)	(.03)
MP compulsory education (GCSE)	.06	-.06	-.03	.05	-.09	-.01	.08	-.05	-.11
	(.07)	(.04)*	(.03)	(.14)	(.06)	(.04)	(.09)	(.04)	(.05)**
MP other qualification	.04	-.02	.03	.05	-.05	.02	.03	.01	.03
	(.07)	(.02)	(.02)	(.10)	(.03)	(.02)	(.09)	(.03)	(.02)
MP's father with a degree	-.01	-.01	.02	.04	-.03	.04	-.05	-.0000109	-.01
	(.03)	(.02)	(.02)	(.04)	(.03)	(.02)**	(.03)	(.02)	(.02)
Obs.	2041	8030	8987	946	4001	4514	1095	4029	4473

Notes:
i) 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. 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. I 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 fine grade average test score over the scores in English, Maths and Science. I 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 $\bar{T}$. I 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 linked dataset. Section 4 offers additional information on the empirical analysis.
iii) The top panel in the tables shows estimates by achievement level in tests in English, Maths and Science that their teachers assess. The second panel shows estimates by type of school at Key Stage 2. The third panel shows estimates by ethnicity. The fourth panel from estimates by eligibility for government interventions in schools. Free School Meals 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 is a dummy equal to one if a student who is not British native gets support in English. SEN is a dummy equal to one if a student obtains additional support by teachers at school. The last panel shows shares of students' parents by education level.

Table 13: 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			Females			Males		
	3	4	5	3	4	5	3	4	5
English	5.32	14.40	17.81	5.32	14.40	17.81	5.32	14.40	17.81
Maths	5.27	10.97	7.42	5.27	10.97	7.42	5.27	10.97	7.42
Science	2.71	10.87	16.77	2.71	10.87	16.77	2.71	10.87	16.77
Average	1.26	0.54	0.09	1.26	0.54	0.09	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 behaviour. 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 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. The last row shows t-statistics of tests of the fine grade average test score.

Table 14: Sensitivity to changes in the sample size that is determined by the window size around a threshold of regression discontinuity estimates of the effect of meeting a performance target on the probability of a behavioural outcome by exploiting discontinuities in the fine grade average test score

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Test score threshold 3				Test score threshold 4				Test score threshold 5			
	2-4	2.05-3.95	2.15-3.85	2.25-3.75	3-5	3.1-4.9	3.3-4.7	3.5-4.5	4-6	4.1-5.9	4.3-5.7	4.5-5.5
Unauthorised absence	.08	.09	.08	.08	-.05	-.05	-.04	-.007	-.01	-.005	-.005	-.008
	(.07)	(.08)	(.08)	(.08)	(.04)	(.03)	(.04)	(.05)	(.02)	(.02)	(.02)	(.02)
Obs.	1729	1577	1330	1073	7119	6436	5006	3492	8128	7772	6802	5417
Being bullied	-.15	-.14	-.15	-.15	.04	.04	.03	.03	-.04	-.04	-.06	-.04
	(.08)*	(.09)*	(.08)*	(.08)*	(.03)	(.03)	(.04)	(.04)	(.02)*	(.02)*	(.03)*	(.04)
Obs.	1586	1442	1226	998	6552	5921	4594	3181	7532	7196	6325	5044
Suspension	.006	-.02	-.01	.009	.03	.03	.05	.05	-.02	-.02	-.02	-.02
	(.06)	(.07)	(.06)	(.08)	(.02)	(.02)	(.03)*	(.03)*	(.02)	(.02)	(.01)	(.01)
Obs.	1676	1523	1293	1051	6936	6265	4856	3357	7980	7622	6694	5357
Expulsion	-.004	-.006	-.004	-.004	-.0009	-.002	-.001	.0006	.004	.003	.004	.003
	(.04)	(.04)	(.04)	(.04)	(.005)	(.005)	(.006)	(.006)	(.002)*	(.003)	(.003)	(.003)
Obs.	1679	1526	1296	1054	6948	6279	4869	3367	7994	7636	6704	5363
Police warnings	.10	.13	.13	.10	.02	.02	.02	.03	-.008	-.007	-.007	-.006
	(.06)	(.07)*	(.07)*	(.06)	(.02)	(.02)	(.02)	(.03)	(.01)	(.01)	(.01)	(.02)
Obs.	1666	1513	1284	1043	6898	6234	4832	3346	7948	7591	6670	5337

Notes:
i) The table shows regression discontinuity estimates which I obtain by altering the size of the window of observations in the estimation with respect to the size of the window in the preferred specification in columns (2), (6) and (10) for test score targets $\bar{T} = 3, 4, 5$. I obtain the estimates by modifying the size of the window $[\bar{T} - 1, \bar{T} + 1]$ around a target $\bar{T}$ in each column in the table, $\bar{T} - 1$ is the target to the left of $\bar{T}$ and $\bar{T} + 1$ is the target to the right of it. The research design is valid if the estimates in the preferred specification and the ones that are obtained by using a modified window are similar. Regression discontinuity estimates in the table are equal to the difference in the probability that a student's main parent answers "yes" to a question on the student's behaviour in column (1) for students to the left and right of a test score target that the Department for Education set at Key Stage 2. They are obtained by estimating the coefficient of a dummy equal to one if a test score T is greater or equal than a threshold $\bar{T}$ and zero otherwise in a Two Stage Least Squares regression in equations (2)-(3) of a dummy B to proxy a behavioural outcome on test score T. $f(T)$ in equation (3) is estimated by using polynomials in the distance of the test score T from the threshold $\bar{T}$ of up to the fourth order. The performance targets in test scores $\bar{T}$ at Key Stage 2 are set by the Department for Education and they are equal 3, 4 and 5. The running variable is the fine grade average test score over the scores in English, Maths and Science. I obtain the bandwidth to estimate the polynomials by using the choice rule in Imbens and Kalyanaraman (2009). In all regressions I 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. Significance levels are as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Section offers 3 offers additional information on the research design and section 4 on the results in the empirical analysis.
ii) The outcome variables are equal to one if a student's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension*, *Expulsion* and *Police warning* are equal to one if a student up to three years before the interview date has been respectively suspended, expelled from school or if the police contacted parents due to their child's behaviour. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.

Table 15: Sensitivity to the dates in which the survey data on behaviour were collected with respect to the disclosure of test scores of regression discontinuity estimates of the effect of meeting a performance target on the probability of a behavioural outcome

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Suspension				Police warning			
	OLS	RD 2-3	RD 3-4	RD 4-5	OLS	RD 2-3	RD 3-4	RD 4-5
In/after April	-.04	.03	.03	-.02	-.03	.10	.02	-.007
	(.01)***	(.06)	(.02)	(.02)	(.008)***	(.06)	(.02)	(.01)
Obs.	9619	1667	6907	7951	9576	1656	6868	7919
In/after May	-.04	-.08	.01	-.01	-.03	.13	.03	.002
	(.01)***	(.07)	(.02)	(.02)	(.009)***	(.08)*	(.03)	(.02)
Obs.	7061	1222	5073	5839	7030	1215	5047	5815
In/after June	-.04	-.05	.04	.004	-.03	.13	.06	-.02
	(.02)***	(.09)	(.03)	(.02)	(.01)**	(.08)*	(.03)*	(.02)
Obs.	4473	772	3196	3701	4458	768	3182	3690
In/after July	-.05	-.14	.08	.03	-.03	.31	.11	.01
	(.02)**	(.11)	(.05)	(.04)	(.02)*	(.09)***	(.04)***	(.03)
Obs.	2134	372	1531	1762	2124	372	1523	1752
In/after August	-.09	.002	.12	-.04	-.04	.08	.17	.0006
	(.04)**	(.07)	(.08)	(.05)	(.03)	(2.83e-15)***	(.10)*	(.04)
Obs.	757	130	543	627	749	129	536	620

Notes: *i*) The table shows regression discontinuity estimates that I obtain by using subsamples of observations which differ by the period in the year in which the survey data on students' behavioural outcomes were collected with respect to estimates obtained from the full sample. The research design is valid if an overlap between the time period in which test scores are disclosed in July 2001 and the May 2001-October 2003 time window in which the survey data was collected does not lead to reverse causality. This may occur only for the the behavioural outcomes *Suspension* and *Police warning* and is not relevant empirically if estimates from the full samples and those from subsamples by survey month are similar. Regression discontinuity estimates in the table are equal to the difference in the probability that a student's main parent answers "yes" to a question on the student's behaviour in column (1) between students to the left and right of a test score target that the Department for Education set at Key Stage 2. They are obtained by estimating the coefficient of a dummy equal to one if a test score T is greater or equal than a threshold $\bar{T}$ and zero otherwise in a Two Stage Least Squares regression in equations (2)-(3) of a dummy B to proxy a behavioural outcome on test score T. $f(T)$ in equation (3) is estimated by using polynomials in the distance of the test score T from the threshold $\bar{T}$ of up to the fourth order. The performance targets in test scores $\bar{T}$ at Key Stage 2 are set by the Department for Education and they are equal 3, 4 and 5. The running variable is the fine grade average test score over the scores in English, Maths and Science. I obtain each estimate by using 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. I obtain the bandwidth to estimate the polynomials by using the choice rule in Imbens and Kalyanaraman (2009). In all regressions I 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. Significance levels are as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Section offers 3 offers additional information on the research design and section 4 on the results in the empirical analysis.
ii) Outcome variables are equal to one if a student's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Suspension* and *Police warning* are equal to one if a student up to three years before the interview date has been respectively suspended from school or if the police contacted parents due to their child's behaviour. 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 linked dataset.

Table 16: Sensitivity to the value of the performance target of regression discontinuity estimates of the effect of meeting a policy-relevant with respect to a falsified performance target in test scores on the probability of behavioural outcomes by exploiting discontinuities in the fine grade average test score

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
	Test score threshold 3						Test score threshold 4						Test score threshold 5					
	3	3.1	3.2	3.3	3.4	3.5	4	4.1	4.2	4.3	4.4	4.5	5	5.1	5.2	5.3	5.4	5.5
Unauthorised absence	.08	-.11	-.006	.05	-.0008	.006	-.05	.03	.04	-.11	-.008	-.007	-.01	.001	.01	-.001	.002	-.02
	(.07)	(.07)	(.06)	(.06)	(.08)	(.08)	(.04)	(.03)	(.03)	(.03)***	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.03)	(.02)
Obs.	1739	1739	1739	1739	1739	1739	7134	7134	7134	7134	7134	7134	8129	8129	8129	8129	8129	8129
Being bullied	-.15	.12	-.14	.14	.03	.006	.04	.07	.03	.003	.02	.006	-.04	.02	.03	-.02	-.01	-.09
	(.08)*	(.09)	(.07)*	(.07)*	(.07)	(.08)	(.03)	(.03)**	(.03)	(.03)	(.03)	(.04)	(.02)*	(.03)	(.04)	(.04)	(.04)	(.05)*
Obs.	1592	1592	1592	1592	1592	1592	6564	6564	6564	6564	6564	6564	7533	7533	7533	7533	7533	7533
Suspension	.001	.009	.01	.08	-.01	-.002	.03	.03	-.03	-.005	.02	.004	-.02	-.006	.02	.005	-.05	-.005
	(.06)	(.07)	(.07)	(.06)	(.06)	(.06)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.01)	(.02)	(.01)	(.02)***	(.02)
Obs.	1682	1682	1682	1682	1682	1682	6949	6949	6949	6949	6949	6949	7981	7981	7981	7981	7981	7981
Expulsion	-.004	.03	.01	.0009	.03	.01	-.0009	-.006	-.0001	-.003	.001	.004	.005	-.0008	.	.	.	
	(.04)	(.04)	(.02)	(.02)	(.01)*	(.01)	(.005)	(.005)	(.004)	(.005)	(.004)	(.002)*	(.004)	(.003)				
Obs.	1685	1685	1685	1685	1685	1685	6961	6961	6961	6961	6961	7995	7995	7995				
Police warning	.10	-.09	.08	.06	.03	-.06	.02	-.02	-.007	.003	-.009	.008	-.008	.002	.009	.02	-.04	-.007
	(.06)	(.07)	(.04)*	(.04)	(.05)	(.04)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.01)	(.01)	(.01)	(.02)	(.01)***	(.008)
Obs.	1672	1672	1672	1672	1672	1672	6911	6911	6911	6911	6911	6911	7949	7949	7949	7949	7949	7949

Notes:
i) The table shows estimates which I obtain by using policy-relevant performance targets such as 4 and falsified ones such as 4.1 to test the relevance of performance targets 3, 4 and 5 that the Department for Education set at Key Stage 2. The research design is valid if the effect of the falsified targets on the probability of behavioural outcomes is not significant. Regression discontinuity estimates in the table are equal to the difference in the probability that a student's main parent answers "yes" to a question on the student's behaviour in column (1) between students to the left and right of a test score target that the Department for Education set at Key Stage 2. They are obtained by estimating the coefficient of a dummy equal to one if a test score T is greater or equal than a threshold $\bar{T}$ and zero otherwise in a Two Stage Least Squares regression in equations (2)-(3) of a dummy B to proxy a behavioural outcome on test score T. $f(T)$ in equation (3) is estimated by using polynomials in the distance of the test score T from the threshold $\bar{T}$ of up to the fourth order. The performance targets in test scores $\bar{T}$ at Key Stage 2 are set by the Department for Education and they are equal 3, 4 and 5. The running variable is the fine grade average test score over the scores in English, Maths and Science. I obtain each estimate by using 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. I obtain the bandwidth to estimate the polynomials by using the choice rule in Imbens and Kalyanaraman (2009). In all regressions I 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. Significance levels are as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Section offers 3 offers additional information on the research design and section 4 on the results in the empirical analysis.
ii) Outcome variables are equal to one if a student's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension* is equal to one if a student has been suspended from school up to three years before the interview date. *Expulsion* is equal to one if a student has been expelled from school up to three years before the interview date. *Police warning* is equal to one if the police contacted parents due to their child's behaviour up to three years before the interview date. Summary statistics of the variables are shown in Table 2.

Table 17: Test of the difference in the value of characteristics that are determined after Key Stage 2 test scores are disclosed for students whose test scores are just to the left or to the right of targets 3, 4 or 5

(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
Missingness in outcome variables									
Unauthorised absence	-.06	.04	-.01	-.06	.05	-.01	-.06	.02	-.02
	(.07)	(.02)*	(.01)	(.09)	(.03)	(.02)	(.09)	(.02)	(.02)
Being bullied	-.02	.01	-.0008	.003	.04	.008	-.04	.002	-.02
	(.09)	(.02)	(.02)	(.12)	(.04)	(.02)	(.12)	(.03)	(.03)
Suspension	.01	.02	-.006	-.07	.02	.01	.04	.02	-.03
	(.08)	(.02)	(.01)	(.10)	(.03)	(.02)	(.10)	(.02)	(.02)
Expulsion	.04	.01	-.006	-.05	.02	.009	.07	.02	-.03
	(.09)	(.02)	(.01)	(.12)	(.03)	(.02)	(.10)	(.02)	(.02)
Police warning	.07	.02	-.006	-.05	.02	.01	.11	.01	-.03
	(.09)	(.02)	(.01)	(.12)	(.03)	(.02)	(.11)	(.02)	(.02)
Key Stage 3 school type									
Voluntary aided school	.003	-.03	-.006	.001	-.04	.003	.004	-.03	-.007
	(.03)	(.02)	(.02)	(.05)	(.03)	(.03)	(.04)	(.02)	(.03)
Voluntary controlled school	.02	-.003	.007	.02	-.001	-.001	.03	-.004	.02
	(.009)**	(.007)	(.01)	(.02)	(.008)	(.01)	(.02)	(.01)	(.01)
Foundation school	-.03	.01	-.01	-.05	.06	.001	-.03	-.03	-.03
	(.06)	(.02)	(.02)	(.08)	(.03)**	(.03)	(.09)	(.03)	(.03)
Community Special school	-.03	.	-.0006	-.06	.	.	.005	.	-.002
	(.03)		(.0008)	(.06)			(.04)		(.002)
Obs.	2041	7988	8987	946	3981	4492	1095	4007	4473

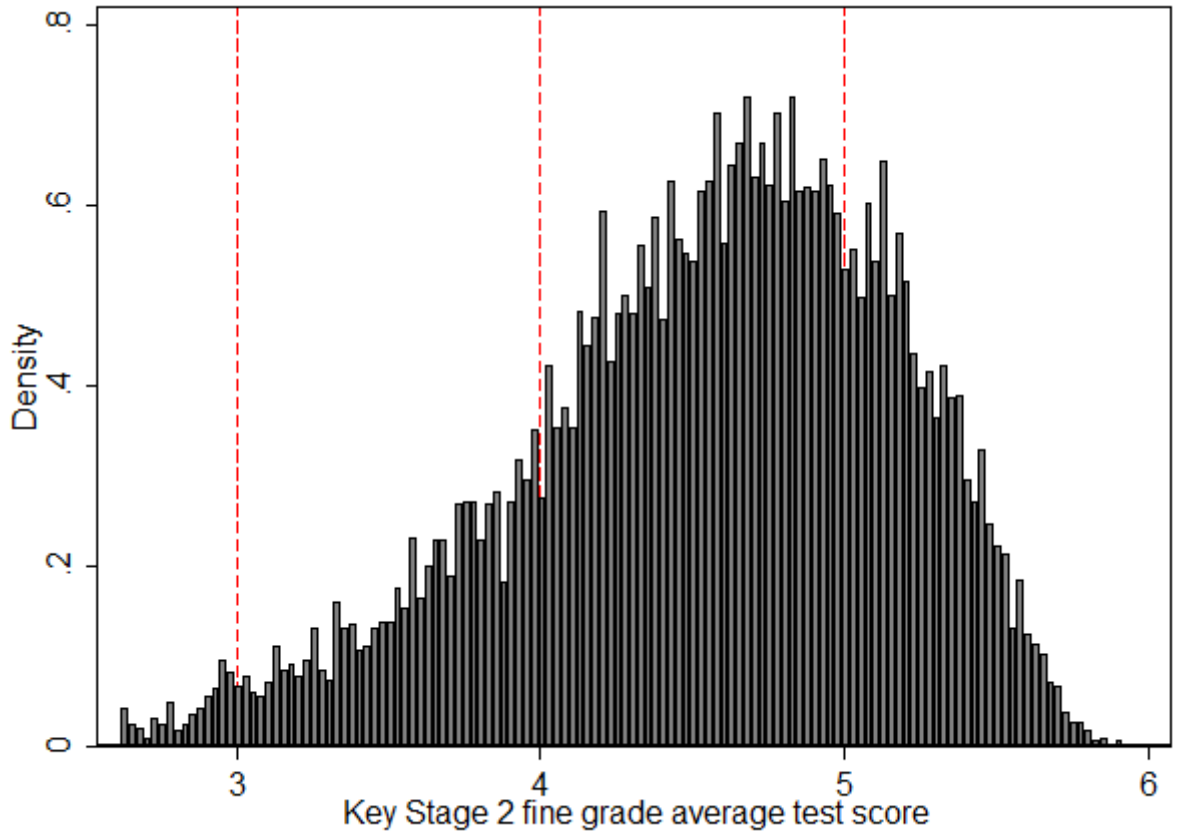
Note: The table shows estimates of the difference between students in the left neighbourhood of a threshold $\bar{T}$ in test score T and those in the right neighbourhood in probability of having a certain characteristic in column (1) that is determined after the disclosure date of test scores at Key Stage 2. 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. I 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 fine grade average test score over the scores in English, Maths and Science. I use a window that is centered at $\bar{T}_c$ 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 $\bar{T}$. I 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 linked dataset. Section 4 offers additional information on the empirical analysis. The top panel in the tables shows estimates by missingness in the outcome variables on behavioural outcomes. The second panel shows estimates by type of secondary school that students attend at Key Stage 3.

Table 18: Sensitivity to parents' education level of regression discontinuity estimates of the effect of meeting a performance target on the probability of a behavioural outcome by exploiting discontinuities in the fine grade average test score

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	All sample			GCSE or higher education			Low or no qualification		
	3	4	5	3	4	5	3	4	5
Unauthorised absence	.08	-.05	-.01	.07	-.08	.002	.01	-.03	-.01
	(.07)	(.04)	(.02)	(.08)	(.05)	(.04)	(.12)	(.04)	(.02)
Obs.	1729	7119	8128	952	2875	2547	809	4422	5820
Being bullied	-.15	.04	-.04	-.11	.02	-.009	-.33	.05	-.07
	(.08)*	(.03)	(.02)*	(.10)	(.05)	(.05)	(.10)***	(.05)	(.04)*
Obs.	1586	6552	7532	802	2414	2147	820	4317	5619
Suspension	.006	.03	-.02	.03	.002	-.01	-.08	.05	-.03
	(.06)	(.02)	(.02)	(.10)	(.04)	(.02)	(.08)	(.03)*	(.02)
Obs.	1676	6936	7980	852	2564	2278	862	4563	5952
Expulsion	-.004	-.0009	.004	.06	-.005	.01	-.04	.005	.002
	(.04)	(.005)	(.002)*	(.03)**	(.009)	(.01)	(.05)	(.003)	(.002)
Obs.	1679	6948	7994	853	2571	2287	864	4569	5958
Police warning	.10	.02	-.008	.07	-.03	.01	.10	.05	-.007
	(.06)	(.02)	(.01)	(.06)	(.03)	(.03)	(.08)	(.03)	(.02)
Obs.	1666	6898	7948	844	2549	2271	858	4538	5926

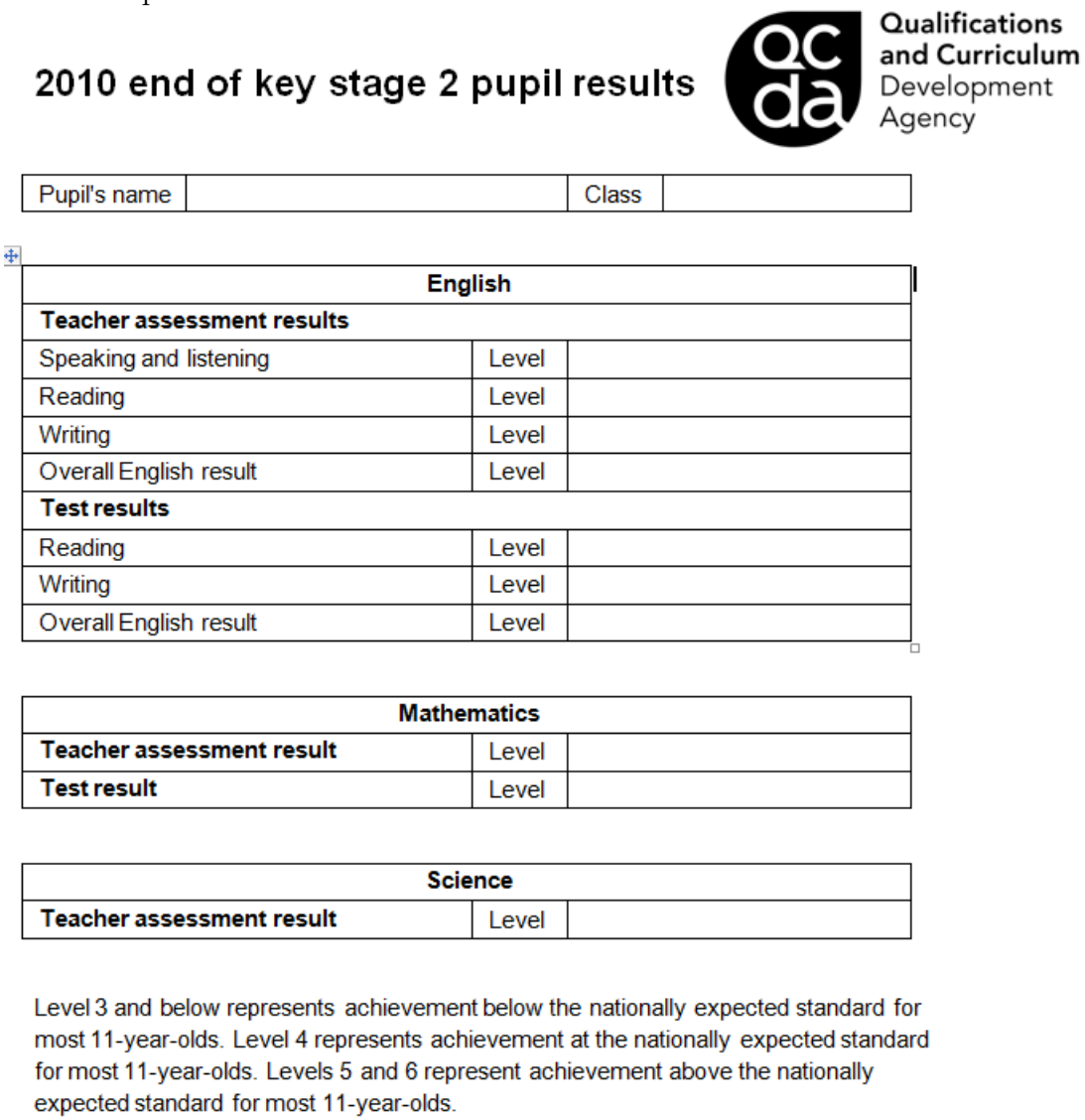
Notes:
i) The table shows regression discontinuity estimates that I obtain by using the full sample of observations and subsamples by whether students' main parents have a GCSE or higher education or otherwise. Regression discontinuity estimates in the table are equal to the difference in the probability that a student's main parent answers "yes" to a question on the student's behaviour in column (1) for students to the left and right of a test score target that the Department for Education set at Key Stage 2. They are obtained by estimating the coefficient of a dummy equal to one if a test score T is greater or equal than a threshold $\bar{T}$ and zero otherwise in a Two Stage Least Squares regression in equations (2)-(3) of a dummy B to proxy a behavioural outcome on test score T. $f(T)$ in equation (3) is estimated by using polynomials in the distance of the test score T from the threshold $\bar{T}$ of up to the fourth order. The performance targets in test scores $\bar{T}$ at Key Stage 2 are set by the Department for Education and they are equal 3, 4 and 5. The running variable is the fine grade average test score over the scores in English, Maths and Science. I obtain each estimate by using 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. I obtain the bandwidth to estimate the polynomials by using the choice rule in Imbens and Kalyanaraman (2009). In all regressions I 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. Significance levels are as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Section offers 3 offers additional information on the research design and section 4 on the results in the empirical analysis.
ii) The outcome variables are equal to one if a student's main parent answers "yes" to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension*, *Expulsion* and *Police warning* are equal to one if a student up to three years before the interview date has been respectively suspended, expelled from school or if the police contacted parents due to their child's behaviour. 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 linked dataset.

Figure 1: 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 4.3 offers additional information about robustness checks to assess the validity of the regression discontinuity design.

Figure 2: Report that schools use to disclose students' achievement levels in Key Stage 2 tests to students and parents



2010 end of key stage 2 pupil results

Qualifications and Curriculum Development Agency

Pupil's name		Class	

English		
Teacher assessment results		
Speaking and listening	Level	
Reading	Level	
Writing	Level	
Overall English result	Level	
Test results		
Reading	Level	
Writing	Level	
Overall English result	Level	

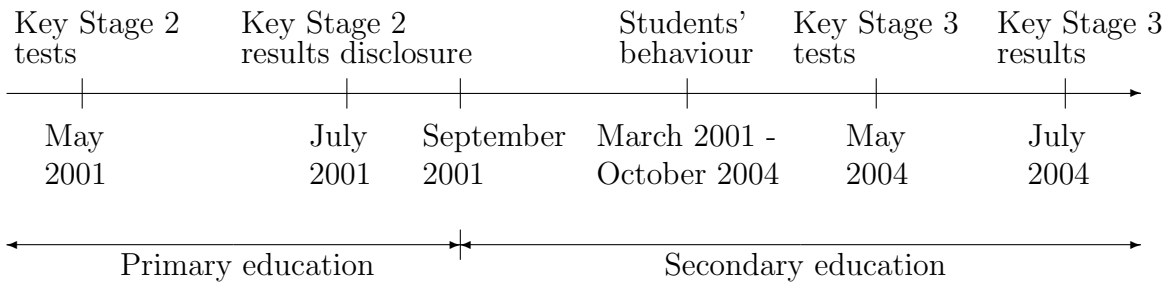
Mathematics		
Teacher assessment result	Level	
Test result	Level	

Science		
Teacher assessment result	Level	

Level 3 and below represents achievement below the nationally expected standard for most 11-year-olds. Level 4 represents achievement at the nationally expected standard for most 11-year-olds. Levels 5 and 6 represent achievement above the nationally expected standard for most 11-year-olds.

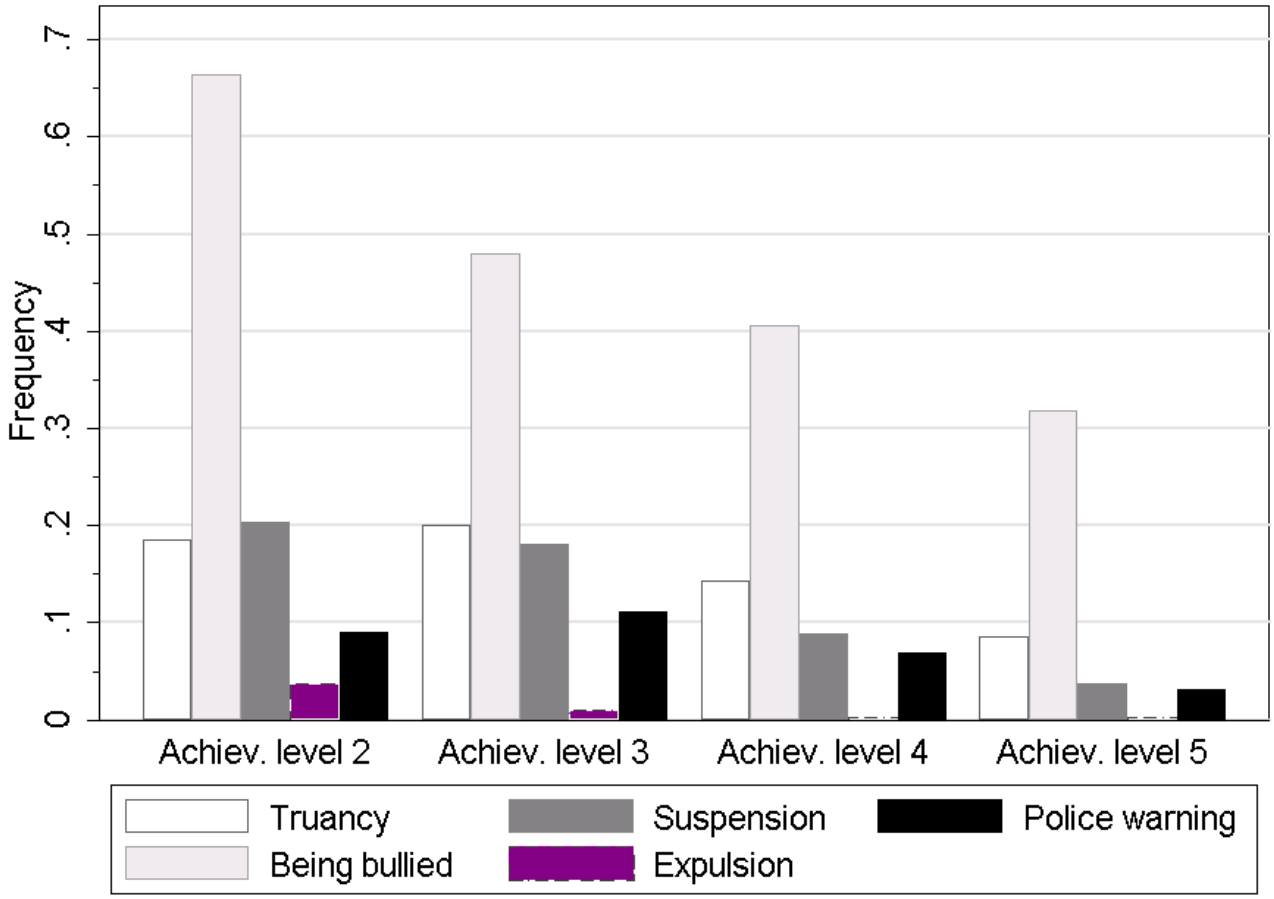
Note: The figure shows the template of the report that schools use to disclose achievement level in tests in English, Maths and Science at Key Stage 2 to students and parents in 2010. The template is similar to the one that schools used in 2001 when tests in Science were externally marked, while they are not from 2010 onwards. The report discloses achievement levels that are a categorical measure of achievement and is equal to 2, 3, the expected level 4 by the Department of Education, or 5. The paragraph at the bottom of the template reports details of the performance target that the Department for Education sets for students in tests at Key Stage 2, level 4, as well as implicit targets for low ability students, level 3, and for high ability ones, level 5. The scores by subject which students obtain in tests are measured on a continuous scale and they are not disclosed. This offers a research design to identify the effect of meeting an achievement level or target on students' behaviour. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.

Figure 3: Illustration of the timeline of tests at Key Stage 2 and behavioural outcomes of students 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. Their behaviour in the period March 2001 to October 2004 is surveyed and recorded in wave 1 in the LSYPE survey dataset. 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 linked dataset.

Figure 4: Mean probability of students' behavioural outcomes by average achievement level in tests at Key Stage 2



Note: The figure shows the mean probability of a behavioural outcome 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. For example, about 20% of students whose average achievement is at level 3 are truant as the second set of bar charts from the left in the figure shows. *Unauthorised absence* is equal to one if a student self-reports an unauthorised absence from school. *Being bullied* is equal to one if the student was bullied. *Suspension* is equal to one if a student has been suspended from school up to three years before the interview date. *Expulsion* is equal to one if a student has been expelled from school up to three years before the interview date. *Police warning* is equal to one if the police contacted parents due to their child's behaviour up to three years before the interview date. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.