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

Marcello Sartarelli†

April 19, 2012

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 study the effect of meeting performance targets in school tests on the probability that students subsequently engage in risky behaviour. This is helpful to assess whether behavioural channels such as motivation and effort by students, parents and teachers, that meeting a target may lead to, play a role in influencing risky behaviour. I address potentially spurious correlations between test scores and students' behaviour by exploiting a regression discontinuity design in test scores and a linked dataset with information on students in compulsory education in England. I find that meeting targets that the government set for students at age 11 has a negative but not significant effect on proxies for risky behaviour, such as the probability of unauthorised absence from school or of a police warning to students' parents. In addition, the effect varies by students' ability, gender, parents' education level and type of risky behaviour. In contrast, linear probability model estimates of the effect are negative and significant. The empirical evidence informs policy decisions about education and support to young people by suggesting *i*) no major behavioural implications of performance targets in tests and *ii*) that naive estimates not correcting for unobservables are spurious and may lead to inaccurate policy.

JEL Classification: C21, I20, I21, I28
Keywords: absence, bullying, education, performance targets, police warning, regression discontinuity, risky behaviour, suspension, test scores

*Thanks for helpful comments to Peter Backus, Lorraine Dearden, Francesca Foliano, Bo Honoré, John Micklewright, Alfonso Miranda, Olmo Silva, Jeff Smith, Anna Vignoles, seminar participants at the European University Institute Micro Group, Institute of Education, IMT Lucca, Lancaster University, Policy Society Institute, Rutgers University, Università di Bari, Università di Cagliari, the 2009 Applied Economics Workshop in Petralia Soprana, the 2009 Australasian Meeting of the Econometric Society, the 2009 Brucchi Luchino Conference, the 2nd Meeting of the Danish Microeconometric Society, the 2009 North American Summer Meeting of the Econometric Society and to the Department for Education for access to the data. I gratefully acknowledge financial support by the ADMIN Node of the National Centre for Research Methods (ESRC grant RES-576-25-0014). All errors are mine.

†Institute of Education, University of London, 20 Bedford Way, London WC1H0AL, UK and University of Alicante, Economics Department, Apartado de Correos 99, 03080 Alicante, Spain, Email: marcello.sartarelli@ua.es

1 Introduction

Performance targets are important in helping individuals to build human capital or signal 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 vice versa for those missing the target. In addition, high ability individuals may exert little effort as the payoff is not proportional to their performance. 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. If low effort or performance by students persist over time, parents, teachers or policy-makers may intervene to remediate this. Similarly, shareholders or policy-makers may intervene in a firm or sector if the performance by the management or employees is poor.

In this paper I study whether meeting performance targets in tests deters students from engaging in risky behaviour in the future, or it may conversely induce such behaviour. I set out to answer this empirically by using linked administrative data on test scores in compulsory education in England and survey data on students' risky behaviour, such as the probability of unauthorised absence from school or of a police warning to parents due to a student's behaviour. If a student fails to meet a performance target in test scores, this may decrease the motivation for studying and hence make a student more prone to engage in risky behaviour, and vice versa if a student meets a target. A competing mechanism suggests instead that failing to meet a target increases a student's motivation to make up for this in the future, hence decreasing the likelihood of risky behaviour. Self-confidence and beliefs about ability are additional channels through which performance feedback can foster positive behaviour, e.g. leading a healthy and safe life in youth and adulthood. For example, an unauthorised absence from school may lead to no consequence in the future. Alternatively, the student may engage in risky activities while absent from school and, perhaps as a result, get a warning from the police. Performance targets in employment contracts have been widely explored, while little is known about performance targets as incentives for students.[1] Hence, understanding whether meeting performance targets in education has an impact on students' behaviour is

[1]See Prendergast (1999) for a review of incentives and performance targets in employer-employee contracts and Stiglitz (2000) for a review of contributions in information economics to overcome informational failure in such settings as employer-employee contracts, or in welfare-to-work programs.

helpful to inform policy decisions about education, such as the design of school curricula, and about support programs for young people and their families.

Unobserved actions by parents and schools may confound the effect of meeting performance targets. For instance, teachers may spend more time helping high ability students than others, or parents with a high education level may make more effort than less educated parents in helping students to prepare for tests, and also in influencing their behaviour before and after tests are held. Hence, I identify the effect of meeting a performance target in test scores, with respect to missing the target, on the probability of risky behaviour by exploiting discontinuities in test scores and I estimate it by using targets that the Department for Education in England set in tests for students.[2] Thanks to this research design, I can tease out the effect of confounders that influence students' test scores, such as parents' or teachers' effort, as they can only imperfectly influence students' scores in a small neighbourhood of a cutoff in test scores.

In the empirical analysis regression discontinuity design estimates show that the effect of meeting a performance target on the probability of engaging in risky behaviour tends to be negative although not statistically significant. On the contrary, potentially spurious correlations that are estimated by using a linear probability model show that the effect is greater in absolute value and significant. The results are similar if the estimates are obtained by subsamples of students with different baseline characteristics, although the effect of meeting certain targets increases the probability of some measures of risky behaviour, and in addition the effect tends to be more significant for males, for the non-white ethnic group, for students with more educated parents and for students who were assigned to government support programs in school at Key Stage 2.

In the recent literature that studies the determinants of young individuals' behaviour Foliano *et al.* (2010) find that an increase in value added by schools in compulsory education in the UK increases a measure of students' disengagement that is commonly used in the psychometric literature. Similarly, Gibbons *et al.* (2010) find that neighbours' characteristics, such as the socio-economic composition and labour market opportunities, have a positive but non-significant 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 related research Reback (2010) finds that school counselors decrease the probability of students' behavioural problems

[2]The targets are 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.

in elementary schools in the USA. Similarly, Imberman (2011) finds that attending a charter school has a positive impact on non-cognitive skills, such as school attendance, by exploiting repeated observations of students in charter and non-charter schools. In contrast, Gaviria and Raphael (2001) find evidence of peer effects in alcohol and drugs use in high schools in the USA by exploiting variation in social interactions at the school and neighbourhood level.[3]

Among recent studies on the effect of achievement in school tests Bandiera *et al.* (2012) find a positive effect of disclosing information about performance in tests on future performance in master degrees in a university in the UK, by exploiting variation in the rules on performance feedback between the departments offering the degrees. Similarly, Azmat and Iriberri (2009) find a positive effect on test scores of giving relative performance feedback to students about the distance of their scores from the average score in their class, by using a natural experiment in a high school in Spain. In related research Hemelt (2011) finds a positive effect of meeting achievement targets by schools on students' future achievement, by exploiting a regression discontinuity design that the school accountability system in the No Child Left Behind program in the USA offers.

Studying the effect of incentives on performance is a research of high interest in economics and in psychology among other disciplines. The consensus in economics is that the effect is positive (Lazear (2000) and Prendergast (1999)), while the literature in psychology suggests the opposite (Deci *et al.* (1999) and Flink *et al.* (1990)), as the two disciplines make different assumptions about the determinants of individuals' motivation. Benabou and Tirole (2002, 2003) reconcile these contrasting results by studying the effect of incentives on motivation in a principal-agent model whose predictions are that the effect is negative if the agent cares about the principal's beliefs and vice versa, since the agent infers that the higher the effort a principal asks for, the worse the principal's assessment of the agent's skills.

The recent increase in interest by policy-makers in the role of education in influencing students' behaviour 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 a multi-billion dollar policy initiative, "No Child Left Behind", that rewards schools that increase students' performances in tests. Experimental designs in the policy are helpful to study the determi-

[3]Grossman (2006) reviews the literature on the positive non-market returns to education in the long-term by focusing on the following outcomes in adulthood: 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.

nants of test score gaps among students, that may lead to adverse effects in adulthood for those left behind at school.[4] In the UK the Department for Education funded a similar policy initiative since 2003, "Every child matters", although its main focus is fostering well-being and positive behaviour in children. Hence, evidence in this paper on the impact of targets in school tests on students' behaviour is of interest to policy-makers who deal with education and public policies for young people. While there is consensus in the literature on the relevance of parents' education for children's education and its monetary and non-monetary benefits for the children, in contrast, little is known about the effect of performance targets in tests, or similar characteristics of the institutional setting in education, on students' behaviour.[5]

This paper offers a novel contribution to the literature that studies the effect of achievement in education on individuals' behaviour by *i*) describing the competing mechanisms that explain why students meeting performance targets in tests may modify their behaviour and *ii*) estimating the causal effect of meeting targets on behaviour by exploiting exogenous variation that cutoffs in test scores offer. Finally, the research design in the paper can be exploited to inform policy decisions in the the future by periodically estimating the effect of meeting targets on behaviour, or other policy variables of interest, thanks to linked administrative and longitudinal survey data in the UK, as well as in other countries whose governments collect rich data on young people, e.g. 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 on 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 characteristics of the institutional setting in compulsory education in England that generate the exogenous variation to identify the effect of meeting performance targets in tests on students' risky behaviour, as well as the data that I use to estimate the effect. Compulsory education is divided into the Foundation Stage plus 4 Key Stages, summing up to 11 years as Table 1 shows. It starts at age 3-4 with the Foundation Stage. Primary school starts at age 5-6 with Key Stage 1 and it is followed by Key Stage 2, as columns (1)-(3)

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

[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) and Dearden *et al.* (1997) show similar evidence in the UK.

in the table show.[6] 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.[7]

I use two linked datasets in the empirical analysis.[8] The first is an administrative dataset with information on test scores of all students in state schools in England (National Pupil Database).[9] 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 contains information from a survey of young people in England (Longitudinal Study of Young People in England). They are students who are representative of the cohort of test-takers in 2001 and who were born between September 1989 and August 1990. The first wave of the survey was held in 2004 when students were 13-14 years old, during Key Stage 3 in secondary school. The survey contains proxies for students' risky behaviour as well as information about their education, and about their parent's education, employment status and work experience.[11] Table 2 shows summary statistics of variables that measure students' risky behaviour, performance in test scores, as well as a rich set of information on gender, ethnicity and school-characteristics for the full sample of students, and also for subsamples by gender and by whether students met the expected performance target in Key Stage 2 tests on average. The sample size of the survey is 15,770 students. However, the dataset in the empirical analysis is smaller since the variables measuring risky behaviour in the survey suffer from item non-response, which varies from 5% to 10%, and I discuss this further in section

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

[7] DirectGov (2010) is a government-maintained website to inform citizens about the characteristics of services in the public sector in the UK and 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".

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

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

[10] The government determines whether a student is eligibile for FSM status based on multiple criteria about receipt of social benefits by parents. In contrast, teachers and psychologists determine eligibility for EAL and SEN status partly based on subjective criteria that the Department for Education set out, which may lead to a discretionary assignment of student to the programs.

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

2.2.

2.1 Key Stage 2 tests

Students sit compulsory tests in English, Maths and Science in the last year in Key Stage 2, when they are 10 or 11 years old, and they are also assessed by their teachers before results in the tests are known.[12] Students' test scripts are graded by using numerical scores on an integer scale in the range 0-100, although it varies slightly by test. The scores are then grouped into four categorical achievement levels, the lowest of which is 2 while the highest is 5. For example, in the year in question, if a score in the Maths test is lower than the cutoff value 22, the achievement level in Maths is equal to 2, while if the score is in the interval 22-48 the level is equal to 3.[13] The Department for Education converts integer scores into decimal scores in the range 2-6 that are obtained by weighting test scores by the distance from the nearest cutoff to the right of the score. For example, the cutoff at 22 in the Maths test score in the earlier example corresponds to 3 in the decimal point score scale. Similarly, integer scores equal to 21 and 23 correspond to 2.96 and 3.04 in the decimal score. In the empirical analysis I use the decimal scores rather than integer ones to simplify the description of the results since cutoffs in the integer scale vary by test, although the change of scale does not affect the results.

External examiners mark students' tests, rather than their teachers. Examiners know the cutoff score for each achievement level in tests and mark all scripts in one subject from a school. In addition, they were instructed by the Department for Education to double-check during the marking process the scripts of those students whose initial marks were very close to a cutoff (Quinlan and Scharaschkin (1999)). Figure 1 shows that this practice leads to jumps in the height of bins in the histograms of test scores in English, Maths and Science at cutoff values 3, 4 and 5, since an examiner may assess that a test script whose provisional score was lower than the cutoff is instead worth a greater score.[14] The figure also shows in the bottom-right the histogram of the average score in all tests at Key Stage 2, whose frequency

[12]The tests are set by the Qualifications and Curriculum Development Agency (QCDA), which is an independent authority from the Department for Education. 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 cutoffs in all tests at Key Stage 2.

[14]Test score variables are censored because 1.6% students in the linked dataset scored 2.5 in one or more tests, which is the minimum score for a student regardless of how poor the performance in the tests. I exclude from the sample that I use in the empirical analysis students whose decimal score was equal to 2.5 because such censoring is of negligible relevance when focusing on the test scores of students in a small neighbourhood of cutoffs.

does not jump at cutoff values. After completing the description of the institutional setting I will address in this section the implications that the marking process has on the validity of the research design and I will show empirical evidence in favour of its validity in section 4.

Categorical achievement levels in externally marked tests in English, Maths and Science at Key Stage 2, together with teacher assessment in these subjects, are disclosed to students and parents. Critically for the research design in this paper, schools do not disclose the underlying test scores, as the results sheet in Figure 2 shows.[15] For example, two students whose 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. The results sheet in Figure 2 also contains a note to help students and parents interpreting tests results, as well as reminding them that the expected target at Key Stage 2 is level 4 in each test. Level 3 is an additional target that applies to low ability students whose score may be lower than the expected target 4, although greater or equal to 3. Similarly, level 5 applies to high ability students who may score considerably above the expected target 4. This leads to three treatments: meeting the low target 3, meeting the expected target 4 or meeting the high target 5. The effect of each of these treatments on students' risky behaviour is identified by exploiting sharp discontinuities in test score since 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.

In the empirical analysis I choose the average score over all tests as running variable to identify the effect of meeting a performance target on risky behaviour by using a regression discontinuity design since it is the best available overall measure of achievement in tests at Key Stage 2 and students are expected by the government to meet the target 4 in all three tests.[16] Studying the effect of meeting performance targets in one test versus a different one or versus more than one test on risky behaviour by using a multi-dimensional RDD (Papay *et al.* (2011)) would offer additional evidence on the determinants of students' risky behaviour. However, neither the variation in scores in the type and number of tests nor the sample size is great enough in the data to exploit a multi-dimensional RDD since approximately 90% of students who met the expected target 4 in one test also met the target in the average score, as the fourth panel in Table 2 shows.

Three characteristics of the rules regulating exam-marking in compulsory education in England

[15] 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.

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

ensure that any potential manipulation of the average test score that may arise in the marking process is imperfect, i.e. there is some randomness over whether students meet a performance target at Key Stage 2, thus offering the exogenous variation that identifies the effect of meeting a target on risky behaviour.[17] First, examiners who mark test scripts do not know students and vice versa, they are fully trained to mark exams in a consistent way and each examiner marks only one of three exams for the same student, as one examiner gets all test scripts in one type of test, e.g. English, in a school. This rules out perfect manipulation of the average test score as examiners have no information about students and their behaviour. Second, a student has his or her tests in English, Maths and Science each marked by a different examiner who only knows the score in one of the three tests by a student. This rules out manipulation both of the other two tests and of the average test score. Third, teachers' efforts are unlikely to be unbalanced towards teaching exclusively to prepare students to tests since there are no monetary incentives for teachers' performance, hence ruling out "teaching to the test" as a threat to the research design.[18]

2.2 Survey evidence of students' risky behaviour

A section in the survey questionnaire consists of questions about a child's behaviour to the child's main parent, who is defined as "the parent most involved in the young person's education" (NatCen (2009)). In the empirical analysis I use as outcomes a set of proxies for students' risky behaviour: unauthorised absence also known as truancy, suspension and expulsion from school, being bullied by other students and police warnings to parents due to a student's behaviour. Each outcome is a dummy equal to one if a student's main parent answered "yes" to a question in the survey and zero otherwise.[19] The survey data was collected from March to October 2004 via face to face interviews with main parents. All questions refer to behaviour that occurred up to a year earlier, i.e. between March and October 2003, except questions on suspension and expulsion from school and on a police warning that refer about behaviour that occurred up to three years before the survey. Since the latter questions may refer to

[17] See Lee and Lemieux (2010) for additional details about the discussion of a regression discontinuity design as a locally randomised experiment in which the running variable is imprecisely manipulated.

[18] 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)). Elwood and Murphy (2002) also summarise the institutional setting in education in the UK, adding insights from the literature in Sociology. 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.

[19] Table A-1 in the appendix shows the wording of the questions underlying each variable in the survey.

events that may have occurred up to April 2001, i.e. three months before the disclosure of tests results in July 2001, I will discuss how to deal with potential reverse causality in the empirical analysis in section 4.[20] The 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 and it varies by outcome from 5% to 10%.

Overall, the outcome variables on risky behaviour by students may capture information about students' choices beyond what they experience at school and with their parents. 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 choice by a student. One may 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 as 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 to tests results rather than for the tests results themselves.[21] The top panel in Table 2 shows the list of outcome variables and their summary statistics. By considering the full sample in the dataset in column (1) parents report that 40% of students were bullied, 14% were truant, i.e. absent at least one day from school without authorisation, 3% were absent from school for one month or longer, 9% were suspended, and that for 7 % the police warned parents about their children's actions. In addition, the table shows in columns (5) and (6) that the probability of risky behaviour is higher for the subsample of students who did not meet the expected performance target 4 on average in all tests at Key Stage 2 than for the students who met the target.[22] This highlights the importance of assessing whether the association between performance targets and risky behaviour also has a causal interpretation.

3 Research design

Let the outcome variable to measure students' risky behaviour be a dummy B that is equal to one if a student engages in risky behaviour, e.g. unauthorised absence from school, and zero otherwise. It can be interpreted as capturing either whether the student's unobserved attitude

[20] Figure A-1 in the appendix describes the timing of the events from Key Stage 2 tests to the collection of survey information.

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

[22] Similarly, Figure A-2 in the appendix shows that the probability of different measures of risky behaviour for students achieving levels 2 and 3 is the greatest among all achievement levels at Key Stage 2 and it then decreases with the achievement level.

to risky behaviour B^* is greater than a cutoff $\bar{B}^*$, i.e. $B = I\{B^* \geq \bar{B}^*\}$ or, similarly, the discount factor to weight the future consequences of certain choices. Also let T be the average score over all tests at Key Stage 2, $\bar{T}$ be a performance target in tests and $P = I\{T \geq \bar{T}\}$ be a dummy equal to 1 if a student meets the performance target and zero otherwise. Estimating a linear probability model of the effect of meeting a performance target in tests on the probability of risky behaviour may lead to spurious estimates as unobservables such as children's ability, parental care or school practices correlate with test scores.

$$B = \alpha + \beta I\{T \geq \bar{T}\} + f(T - \bar{T}) + U \tag{1}$$

Conversely, by using a regression discontinuity design (RDD) I identify the effect of meeting a performance target overall in tests on risky behaviour since the design compares the behaviour of students whose average test score is greater than the cutoff by a very small amount with those whose score is smaller than the cutoff by an equally small amount. Since obtaining a score barely to the left or to the right of a cutoff is arguably due to chance, students whose score is in a small neighbourhood on either side of a cutoff are not different in observables and unobservables[23] The parameter β in equation (1) is, for example, negative if students who met a performance target ($T \geq \bar{T}$) are also less likely to be bullied by other students than those who did not meet it ($T < \bar{T}$), as meeting the target may have increased their self-confidence. By exploiting three cutoffs $\bar{T} = 3, 4, 5$ in the average test score that are performance target for students in England I estimate the effect of the treatment "meeting a performance target in the test score" with respect to the control "not meeting a performance target in the test score" on the probability that a student engages in risky behaviour. The RDD holds under the identifying assumptions that students on the left of the cutoff $\bar{T}$ are similar to those on the right of it, for example, in their socio-economic background that parents' education proxies. In other words, obtaining a test score that is just lower than a cutoff value or target (control group) or just to the right of it (treatment group) can arguably be seen as a stochastic shock to the test score due to nature.

The discontinuities in the average test score are sharp as the probability of meeting a target jumps from zero to one if a student scores just to the right with respect to just to the left of a cutoff. Hence, I estimate the effect of meeting the target by fitting equation (1) with smooth polynomials $f(T - \bar{T})$ in the difference in test score T from the cutoff $\bar{T}$ and separately for

[23] See Thistlethwaite and Campbell (1960) and Trochim (1984) for the early development of the RDD. See instead Imbens and Lemieux (2008) and Lee and Lemieux (2010) for a survey of the most recent the advances in the theory as well the recent increase in the number of applications of RDD in economics.

subsamples of students whose test score is smaller than the cutoff $\bar{T}$ and for those whose score is greater than $\bar{T}$. In other words, $\hat{\beta}$ is an estimate of the distance at the cutoff in the level of the polynomials, or the average effect of meeting a performance target at the cutoff. I obtain the estimates by using as observations those in the interval $[\bar{T}-1, \bar{T}+1]$, that goes from the cutoff $\bar{T}-1$ that is to the left of $\bar{T}$ to the cutoff $\bar{T}+1$ to the right of it. A smaller window around the cutoff $\bar{T}$ would omit relevant observations of students whose test score is in a neighbourhood of the cutoff. In contrast, a larger window would include observations of those students whose average test score is in a neighbourhood of a lower or higher cutoff than the one that is considered in the estimation, hence leading to confounded estimates. 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. In the preferred specification I also add baseline characteristics of students and schools, such as gender, ethnicity, school type, assignment to support programs at school and school characteristics, so as to increase the precision of the RDD estimates.

4 Results

This section firstly discusses the validity of the research design and empirical evidence in support of it (section 4.1) and then presents the main results in the empirical analysis (section 4.2).

4.1 Validity of the research design

A research design to identify the effect of meeting a performance target on students' risky behaviour is valid if it mimics a controlled experiment in which the treated group consists of students who are randomly assigned a score greater or equal than the target score, and the control group consists of students who are randomly assigned a score lower than the target. Hence, I assess empirically whether, similarly to a controlled experiment, *i*) students' baseline characteristics, e.g. gender, are balanced at the cutoffs 3, 4 and 5 in the scores and *ii*) the distribution of the average test score is smooth at the cutoffs.[24]

The value of students' and schools baseline covariates is determined before the test scores are disclosed. Hence, if the average test score is "as if" locally randomised at a cutoff, it leads to balanced baseline covariates at cutoffs in the score both for individual covariates, i.e. the share

[24]See Lee and Lemieux (2010) for a discussion about the threats to the validity of a RDD and for the approaches that are available to test for this empirically.

of males should be the same on either side of a cutoff, and by considering all covariates jointly. Otherwise, the effect on risky behaviour of meeting a performance target is confounded by the correlation between, for example, gender and performance in tests, thus invalidating the randomised design around a cutoff in a test score. Table 3 shows estimates of the difference in the value of baseline covariates at a cutoff that is obtained by fitting smooth polynomials of a covariate in the average test score, separately for subsamples of students whose score is to the left or to the right of cutoffs 3, 4 and 5. Small and non-significant estimates of the difference in baseline covariates, such as gender, ethnicity, type of school, participation in government support programs or the number of teachers or students in a class, suggest that the baseline covariates are balanced at cutoffs. Table 4 shows instead t-statistics and p-values to test the null hypothesis that the joint difference in the value of all covariates at a cutoff is zero in a seemingly unrelated model with as many equations as are the covariates, and in each equation I regress a covariate on a dummy equal to one if the student met a performance target and a polynomial in the difference in the score relative to a target. P-values different from zero in the table suggest that the covariates are jointly balanced.

Similarly, a randomisation of students' scores to either side of a cutoff in the average test score leads to a smooth distribution of the score at a cutoff value. On the contrary, a jump in the distribution at a cutoff suggests potential manipulation of the score by students, teachers or the external examiners who grade the tests.[25] I assess whether the average test score is manipulated at cutoffs by plotting a histogram of the score with a bin width equal to 0.05 to obtain histogram bins that contain an arbitrarily small number of students separately to the left and right of a cutoff and no bin contains the cutoff value. Visual inspection of the histogram in the bottom right of Figure 1 suggests no suspicious jumps in the height of the bins in the histogram at cutoffs. As per the jumps in the histograms of test score in each subject in Figure 1, they are a mechanical consequence of the re-grading policy by the Department for Education that instructed examiners to double-check during the marking process the scripts of those students whose provisional score in a small neighbourhood of a cutoff (Quinlan and Scharaschkin (1999)).[26] What is essential for the validity of the research

[25]See as examples of potential manipulation of scores at a cutoff Jacob and Lefgren (2004) who study the effect of remediation courses on test scores in schools in the USA and Urquiola and Verhoogen (2009) who study the effect of class size on test scores in schools in Chile. In both examples manipulation is induced by dysfunctional responses by teachers to incentives that are embedded in the institutional setting.

[26]The manipulation of individual tests at cutoffs and of no manipulation in the average test score that the histograms in Figure 1 suggest is in line with tests in in Table A-2 of the null hypothesis of no manipulation, that is not rejected if the difference in the height of the histogram bins on either side of a cutoff is sufficiently small (McCrary (2008)).

design is that marking of test scripts is conducted by external examiners rather than by students' teachers. Since examiners do not know students, they cannot by design manipulate scores perfectly, e.g. give consistently higher marks to female or non-white students. This ensures imperfect manipulation of test scores in a small neighbourhood of a cutoff in the first round of marking and also in a second, restricted and less accurate round in which examiners revise marks for the subset of students in a neighbourhood of a cutoff.[27]

4.2 Results for the full sample

Table 5 shows estimates of the effect of meeting performance targets 3, 4 and 5 in the average test score on the probability that students engage in risky behaviour for the full sample of students. The left panel in the table shows RDD estimates while the right one shows linear probability model (LPM) estimates, as an example of a naive estimator of the effect of meeting a performance target. Column (1) shows that the effect of meeting the target 3 decreases the probability of, for example, being bullied by 9 percentage points or 22.5% with respect to the mean probability of being bullied that Table 2 shows. The estimate is different from zero and hence economically significant, although it is not statistically significant at conventional levels. Similarly, estimates of the effect of meeting the target 3 on other proxies for risky behaivour are different from zero, although they are not statistically significant. Column (2) in the table shows that meeting the target 4, which the Department for Education expects all students to meet at Key Stage 2, has a smaller effect on the probability of risky behaviour than the effect of meeting target 3 and it is not statistically significant. Finally, meeting target 5 tends to have a negative effect, thus decreasing the probability of engaging in risky behaviour. The effect on the probability of a police warning and of one month absence are not statistically significant, while the probability of being bullied and of being expelled are at the 10% and 5% level respectively.

Overall, estimates in columns (1)-(3) in Table 5 show that only very few RDD estimates are statistically significant, thus not rejecting the hypothesis of no impact of meeting targets in tests on risky behaviour.[28] In addition, the effects of meeting performance targets differ by target as, for example, the negative effect on the probability of being bullied is greater in

[27]Lee and Lemieux (2010) and McCrary (2008) discuss similar cases in which a jump in the distribution of the running variable jumps at a cutoff for mechanical reasons rather than because the running variable is genuinely manipulated.

[28]The RDD estimates are robust to dropping the observations of to the students who were interviewed after they had learnt about their performance in Key Stage 2 tests, as the robustness checks in Table A-3 in the appendix suggest.

absolute value at cutoff 3 than at 4 and 5. Finally, LPM estimates in columns (4)-(6) in the table tend to be statistically significant, have the same sign and they are greater in absolute value than RDD estimates, thus suggesting that not correcting for the correlation between unobservables and test scores may lead to overestimate the effect of meeting the performance target in tests on risky behaviour.

4.3 Results by sub-samples

I also assess whether the effect of meeting a performance target in tests on the probability of risky behaviour vary if it is estimated by using subsamples of students who differ in the following baseline characteristics: gender, ethnic group, type of school, whether the main parent completed at least secondary education (GCSE) and by students' assignment to support programs at school.

Table 6 shows that separate RDD estimates for sub-samples by various baseline characteristics tend to be in line with economically significant although not statistically significant estimates for the full sample in Table 5. The main differences are that the effect of meeting any target on the probability of risky behaviour tends to be greater and more precisely estimated for males over females, e.g. a significant increase at cutoff 4 in the probability of police warning for males by 5 percentage points and of suspension by 10 percentage points respectively at 10% and 5% level, while the effects are negative and not significant for females. In addition, the effect tends to be greater for non-white students over white ones and and similarly for students going to a different school type than a Community school, although in both cases the effect tends not to be statistically significant. Finally, the effect of meeting any target tends to be more precisely estimated for students whose parents' completed compulsory education or beyond than for students whose parents did not, e.g. a significant increase at cutoff 4 in the probability of police warning by 5 percentage points and of suspension by 6 percentage points at 10% level for students with more literate parents, while the effects are not significant for other students.

Table 7 shows separate RDD estimates by students' assignment to government support programs at school: whether a student has English as Additional Language (EAL), Special Education Needs (SEN), high level special needs (SEN statemented) or Free School Meal (FSM). Separate estimates for the different sub-samples tend to be in line with economically significant although not statistically significant estimates for the full sample in Table 5. The main differences are that the effect of meeting target 4 on the probability of being bullied tends to

be positive and more precisely estimated for students who are assigned EAL (15 percentage points at 5% level) and FSM (12 percentage points at 5% level) than for other students and. Similarly, meeting the target 4 increases the probability of suspension for students assigned to EAL (10 percentage points at 1% level), SEN (14 percentage points at 1% level) and FSM (8 percentage points at 1% level).

5 Discussion

In this paper I study the effect of meeting performance targets in school tests on the probability that students engage in risky behaviour, such as police warnings to parents due to their children's behaviour. I use a regression discontinuity design that cutoffs in test scores offer and data on students in England to identify and estimate the effect, thus teasing out confounders such as unobserved students' ability, parental efforts or school practices. I find that meeting performance targets tends to have a negative although not statistically significant effect on the probability of risky behaviour. In addition, the precision of the estimates varies with students' baseline covariates. Estimates of the effect of meeting the target that are obtained by using linear probability models tend to have the same sign although they are greater and statistically significant at conventional levels, thus suggesting that they may capture a spurious correlation between test scores and unobservables and hence lead to inaccurate policy decisions.

Not rejecting empirically a null effect of meeting performance targets in tests on the probability of suspension and expulsion from school, or of a police warning, is reassuring for policy-makers as it suggests that the targets have no major behavioural implication for students that was not anticipated when they were setup. In addition, heterogeneous effects by gender, main parent's education level or by the assignment to support programs at school offer some empirical support for a tradeoff in students' development between nature and nurture. Children learn over time whether to follow either their innate attitudes and experiment them through their behaviour or conversely the rules and norms that they can learn from parents and teachers, or a combination of both attitudes and inputs from parents and teachers, as Lizzeri and Siniscalchi (2008) and the literature on economics, genetics and sociology also suggest. The size and significance of the effect also vary by the type of control which parents or teachers have over a student's risky behaviour, i.e. parents have a higher control over long-term absence from school than over bullying, which offers additional support to this tradeoff. Hence, similarly to students, parents and teachers can either let children fully follow their attitudes or fully direct

their actions, or a combination of both approaches.

Categorical measures of performance that act as targets for students in tests and rich data offer a useful empirical toolkit that policy-makers can use to assess behavioural effects of targets in test scores in the future. A negative although not significant effect of meeting targets on the probability of risky behaviour for certain groups of students, such as non-white students, suggests that targets may be valuable although weak signals to motivate these students. However, disclosing to students categorical information about their achievement in tests instead of richer information such as the percentile at which a student scored in the distribution of test scores leads to question whether there is a net gain or loss for students and teachers by disclosing richer information about achievement at school. However, answering this questions requires a structural model that can tease out competing behavioural channels and it is left for future research.

Additional empirical evidence on the impact of incentives on individuals' motivation and on future choices over education and in the labour market would be helpful to inform policy decisions, and also help to reconcile the contrasting results on incentives and motivation in economics and psychology that Benabou and Tirole (2002, 2003) survey. Hence in future research I will study on the role of performance targets as incentives in education by assessing whether they have an effect on students' achievement and risky behaviour in secondary school and beyond (Micklewright and Sartarelli (2011)) In addition, I will study whether meeting performance targets in tests has an impact on government financial support to schools (Sartarelli and Tampieri (2011)).

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. (2012). *Blissful Ignorance? Evidence From a Natural Experiment on The Effect of Individual Feedback on Performance*. mimeo, 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.

Dearden, L., Machin, S. and Reed, H. (1997). Intergenerational mobility in britain. *Economic Journal*, **107** (440), 47–66.

Deci, E. L., Koestner, R. and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, **125** (6), 627 – 668.

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.

Elwood, J. and Murphy, P. (2002). Tests, tiers and achievement: gender and performance at 16 and 14 in england. *European Journal of Education*, **37** (4), 395–416.

Flink, C., Boggiano, A. K. and Barrett, M. (1990). Controlling teaching strategies: Undermining children's self-determination and performance. *Journal of Personality and Social Psychology*, **59** (5), 916–924.

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.

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.

Hemelt, S. W. (2011). Performance effects of failure to make adequate yearly progress (ayp): Evidence from a regression discontinuity framework. *Economics of Education Review*, **30** (4), 702–723.

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

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

Imberman, S. A. (2011). The effect of charter schools on achievement and behavior of public school students. *Journal of Public Economics*, **95** (7-8), 850 – 863.

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.

Lazear, E. P. (2000). Performance pay and productivity. *The American Economic Review*, **90** (5), pp. 1346–1361.

Lee, D. S. 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.

Micklewright, J. and Sartarelli, M. (2011). *Incentives and Rules in Marking Tests: Evidence from Discontinuities in Scores in Compulsory Education in England*. mimeo, Institute of Education, University of London.

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.

Papay, J. P., Willett, J. B. and Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, **161** (2), 203–207.

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

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

Quinlan, M. and Scharaschkin, A. (1999). National curriculum testing: problems and practicalities. The British Educational Research Association, Annual Conference Brighton.

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.

Sartarelli, M. and Tampieri, A. (2011). *Do Performance Targets Affect Financial Support to Schools? Evidence from Discontinuities in Test Scores*. mimeo, Institute of Education.

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: 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			
	4-5		Reception	Tests	6-9/13 elements
Primary School	5-6	Key Stage 1	1		
	6-7		2	Teacher assessment in English, Maths and Science (EMS)	2
	7-10	Key Stage 2	3-5		
	10-11		6	National and teacher assessment in EMS	4
Secondary School	11-13	Key Stage 3	7	Teacher assessment	
	13-14		9	Teacher assessment 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 and column (3) the names of the different stages. Column (4) lists as a count the 11 years of compulsory education. 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 a 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

	All sample	Gender		Meeting target 4 on average in Key Stage 2 tests	
		Female	Male	No	Yes
Risky behaviour (dummy=0 No, =1 Yes; survey to student's main parent)					
Being bullied	0.40	0.44	0.36	0.49	0.38
Unauthorised absence	0.14	0.14	0.14	0.20	0.13
1 month absence	0.03	0.03	0.03	0.06	0.03
Suspension	0.09	0.05	0.13	0.18	0.07
Expulsion	0.00	0.00	0.01	0.01	0.00
Police warning	0.07	0.04	0.09	0.11	0.06
Gender and ethnicity (dummy=0 No, =1 Yes)					
Male	0.51	0.00	1.00	0.54	0.50
White	0.67	0.65	0.68	0.59	0.68
Black	0.06	0.07	0.06	0.08	0.06
Asian	0.19	0.19	0.18	0.25	0.17
Other	0.06	0.06	0.06	0.06	0.06
School type at Key Stage 2 (dummy=0 No, =1 Yes)					
Community	0.70	0.70	0.70	0.76	0.68
Voluntary aided	0.17	0.18	0.17	0.14	0.18
Voluntary controlled	0.09	0.09	0.09	0.07	0.09
Foundation	0.03	0.03	0.03	0.02	0.03
Other	0.01	0.01	0.01	0.00	0.01
Male	0.51	0.00	1.00	0.54	0.50
Meeting target 4 in Key Stage 2 tests (dummy=0 No, =1 Yes)					
English	0.77	0.82	0.72	0.17	0.90
Maths	0.72	0.71	0.73	0.07	0.87
Science	0.88	0.89	0.88	0.52	0.96
Teacher assessment in English	0.75	0.80	0.71	0.19	0.88
Teacher assessment in Maths	0.76	0.76	0.76	0.17	0.89
Teacher assessment in Science	0.84	0.84	0.84	0.37	0.94
Main parent's education and grandfather's (dummy=0 No, =1 Yes)					
Degree	0.11	0.11	0.11	0.03	0.13
Higher education	0.12	0.12	0.13	0.08	0.13
GCSE	0.42	0.42	0.42	0.34	0.44
Other qualification	0.11	0.11	0.11	0.15	0.10
Grandfather has a degree	0.06	0.07	0.06	0.04	0.07
School-level characteristics: students in support programs (dummy=0 No, =1 Yes)					
Special education needs (SEN)	0.21	0.20	0.21	0.23	0.20
Special education needs statemented	0.02	0.02	0.02	0.02	0.02
English additional language	0.16	0.17	0.15	0.21	0.15
Free school meal	0.17	0.17	0.17	0.22	0.16

Continued on next page...

Continued from previous page...

	All sample	Gender		Meeting target 4 on average in Key Stage 2 tests	
		Female	Male	No	Yes
School-level characteristics: staff and teachers (full-time equivalent)					
N. qualified teachers	16.11	16.14	16.08	16.47	16.03
S.d.	7.14	7.12	7.16	6.62	7.24
N. SEN support staff	1.53	1.54	1.52	1.74	1.49
S.d.	1.87	1.86	1.88	1.97	1.84
N. minority ethnic support staff	0.23	0.24	0.22	0.32	0.21
S.d.	0.76	0.77	0.74	0.84	0.74
N. other education support staff	1.30	1.35	1.26	1.41	1.28
S.d.	1.89	1.90	1.88	2.00	1.87
N. non teaching staff	4.86	4.93	4.78	5.51	4.71
S.d.	3.40	3.43	3.36	3.47	3.37
N. secretaries	1.04	1.03	1.04	1.05	1.03
S.d.	0.58	0.59	0.57	0.57	0.58
N. teachers and non teachers	20.81	20.95	20.68	21.99	20.55
S.d.	9.22	9.22	9.22	8.87	9.28
Student-teacher ratio	23.24	23.24	23.24	22.88	23.32
S.d.	3.14	3.08	3.20	3.12	3.14
School-level characteristics at Key Stage 1: classes and students					
N. classes with 1 teacher	3.09	3.12	3.07	3.10	3.09
S.d.	2.66	2.66	2.66	2.64	2.67
Average size of 1 teacher classes	18.15	18.38	17.93	17.89	18.21
S.d.	12.37	12.33	12.40	12.34	12.37
N. classes with 2+ teachers	0.24	0.25	0.23	0.32	0.23
S.d.	0.65	0.68	0.63	0.78	0.62
N. students in a 2+ teacher class	7.03	7.31	6.76	8.98	6.60
S.d.	19.19	19.88	18.49	21.78	18.54
School-level characteristics at Key Stage 2: classes and students					
N. classes with 1 teacher	6.20	6.14	6.26	6.26	6.18
S.d.	3.91	3.85	3.96	3.91	3.91
Average size of 1 teacher class	26.06	26.05	26.06	25.67	26.14
S.d.	8.72	8.70	8.73	8.70	8.72
N. classes with 2+ teachers	0.51	0.52	0.50	0.59	0.49
S.d.	1.06	1.09	1.03	1.10	1.05
Average size of 2+ teacher classes	7.99	7.92	8.05	8.74	7.82
S.d.	13.35	13.27	13.43	13.50	13.31
Observations	11307	5584	5723	2041	9266

Notes: The table shows summary statistics of outcome variables and covariates that I use in the empirical analysis in section 4. The table shows across columns from left to right summary statistics for the full sample and for sub-samples by gender and by whether a student met on average in tests the performance target 4 that the governments expects from students at Key Stage 2. The top two panels show summary statistics of the outcome variables in the survey data, with non-response varying from 1% to 15%. Outcome variables are dummies equal to one if a student or the main parent answered yes to a question on the student's behaviour up to one to three years before the interview date and zero otherwise. The remaining panels show summary statistics of students' and schools characteristics in the administrative dataset. The total number of observations in the last row refers to the linked survey-administrative dataset.

Table 3: Test of the balance of students' and schools baseline characteristics at cutoffs in the average test score

	Cutoff 3	Cutoff 4	Cutoff 5
Gender and ethnicity (dummy=0 No, =1 Yes)			
Male	.10	.02	.06
	(.08)	(.03)	(.03)*
White	.04	-.04	.005
	(.10)	(.03)	(.03)
Black	.07	-.01	.0005
	(.05)	(.02)	(.02)
Asian	-.08	.03	-.0009
	(.08)	(.03)	(.02)
Other ethnicity	.005	.009	.01
	(.04)	(.01)	(.01)
School type at Key Stage 2 (dummy=0 No, =1 Yes)			
Community	-.05	.01	-.05
	(.06)	(.03)	(.03)*
Voluntary aided	.05	-.01	.03
	(.05)	(.02)	(.02)
Voluntary controlled	-.02	-.003	.007
	(.04)	(.02)	(.02)
Foundation	.007	-.002	.006
	(.007)	(.009)	(.009)
Other type	.005	-.002	.004
	(.02)	(.003)	(.005)
Meeting targets in Key Stage 2 tests (dummy=0 No, =1 Yes)			
English	.06	.02	-.003
	(.03)*	(.04)	(.002)*
Maths	.02	.05	-.001
	(.02)	(.04)	(.005)
Science	.20	-.009	-.001
	(.05)***	(.02)	(.002)
Teacher assessment in English	-.06	-.02	-.003
	(.04)	(.04)	(.007)
Teacher assessment in Maths	-.04	.07	.003
	(.03)	(.04)*	(.003)
Teacher assessment in Science	-.03	-.006	.001
	(.05)	(.04)	(.002)
Main parent's and grandfather's education (dummy=0 No, =1 Yes)			
Degree	.008	.02	.06
	(.02)	(.02)	(.02)**
Higher education	-.09	.03	-.02
	(.05)*	(.02)	(.02)
GCSE	.06	-.06	-.03
	(.07)	(.04)*	(.03)
Other qualification	.04	-.02	.03
	(.07)	(.02)	(.02)
Grandfather with a degree	-.01	-.01	.02
	(.03)	(.02)	(.02)

Continued on next page...

Continued from previous page...

	Cutoff 3	Cutoff 4	Cutoff 5
School-level characteristics: students in support programs (dummy=0 No, =1 Yes)			
Special education needs (SEN)	-0.01	0.01	-0.001
	(.02)	(.01)	(.01)
Special education needs statemented	0.01	-.001	.0002
	(.02)	(.002)	(.001)
English additional language	-0.07	0.03	0.01
	(.06)	(.02)	(.01)
Free school meal	-0.06	0.01	0.01
	(.03)**	(.01)	(.01)
School-level characteristics: staff and teachers (full-time equivalent)			
N. qualified teachers	1.07	.03	-.11
	(1.12)	(.44)	(.46)
N. SEN support staff	.28	.002	-.20
	(.35)	(.12)	(.10)**
N. minority ethnic support staff	-.14	.03	-.03
	(.15)	(.05)	(.04)
N. other education support staff	.15	.14	-.01
	(.33)	(.12)	(.11)
N. non teaching staff	-.70	.17	-.16
	(.78)	(.22)	(.20)
N. secretaries	.06	.01	-.03
	(.10)	(.04)	(.03)
N. teachers and non teachers	-.35	.27	-.07
	(1.74)	(.60)	(.59)
Student-teacher ratio	.30	-.11	-.08
	(.60)	(.15)	(.17)
School-level characteristics at Key Stage 1 and 2: classes and students			
N. classes with 1 teacher (Key Stage 1)	-.71	.008	.30
	(.59)	(.18)	(.17)*
Average size of a 1 teacher class (Key Stage 1)	-4.56	.11	.84
	(2.67)*	(.83)	(.62)
N. classes with 2+ teachers (Key Stage 1)	-.22	-.01	-.03
	(.24)	(.04)	(.03)
Average size of a 2+ teachers class (Key Stage 1)	-.70	-.31	-.44
	(2.54)	(.81)	(.63)
N. classes with 1 teacher (Key Stage 2)	-.25	-.20	-.38
	(.71)	(.23)	(.27)
Average size of a 1 teacher class (Key Stage 2)	-2.55	-.26	.41
	(1.73)	(.43)	(.44)
N. classes with 2+ teachers (Key Stage 2)	.11	.10	-.03
	(.18)	(.06)	(.06)
Average size of a 2+ teachers class (Key Stage 2)	-.42	1.28	-.40
	(2.49)	(.74)*	(.75)
Observations	1920	7656	8619

Note: The table shows regression discontinuity estimates to assess the difference in the value of a covariate whose value is determined before the disclosure date of Key Stage 2 tests between those students whose average score T is just to the left of a cutoff $\bar{T}$ and those whose score is just to the right. I regress the covariates on smooth polynomials in test scores separately for students to the left and right of each of score target $\bar{T}$: 3, 4 and 5, that the Department for Education in the UK set. The running variable is the average score over tests in English, Maths and Science. I use a window that contains observations in the interval $[\bar{T}-1, \bar{T}+1]$, i.e. the cutoffs to the left and to the right of $\bar{T}$. I use the choice rule in Imbens and Kalyanaraman (2009) to obtain the bandwidth. The significance levels are as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Section 2 offers additional information on the institutional setting and on the data, and Section 4 on the empirical analysis.

Table 4: Test of the balance of all students' and schools baseline characteristics jointly at cutoffs in the average test score

	Cutoff 3	Cutoff 4	Cutoff 5
	Bandwidth = 0.3		
χ^2 statistic	40.21	15.11	20.97
P-value	0.01	0.82	0.46
	Bandwidth = 0.4		
χ^2 statistic	28.62	19.21	26.68
P-value	0.12	0.57	0.18
	Bandwidth = 0.5		
χ^2 statistic	28.03	23.28	27.36
P-value	0.14	0.33	0.16
	Bandwidth = 0.6		
χ^2 statistic	21.45	23.71	23.80
P-value	0.43	0.31	0.30
	Bandwidth = 0.7		
χ^2 statistic	18.17	20.99	21.89
P-value	0.64	0.46	0.41
	Bandwidth = 0.8		
χ^2 statistic	16.51	22.51	21.84
P-value	0.74	0.37	0.41
Observations	1920	7656	8619

Note: The table shows chi-squared statistics and p-values of the joint test of the null hypothesis of no discontinuities in the value of any of the covariates, whose value is determined before tests, at cutoffs $\bar{T}$ that are performance targets in the average score T in Key Stage 2 tests. I regress the covariates on a dummy $I\{T \geq \bar{T}\}$ that is equal to one if a student met a target and zero otherwise, fourth order polynomials in the distance in the actual test score from the target $(\bar{T} - T)$ and the interaction between the dummy and the distances. I stack equations for each covariate in a seemingly unrelated system of regressions (SUR). The test statistics of the null hypothesis that the effect of the target is jointly zero on all covariates follows a chi-squared distribution (see Lee and Lemieux (2010) for additional details). The table shows test statistics and p-values that are obtained by using different bandwidths in estimating the SUR model. Section 2 offers additional information on the institutional setting and on the data, and Section 4 on the empirical analysis.

Table 5: Regression discontinuity design (RDD) and linear probability model (LPM) estimates of the effect of meeting performance targets in the average test score on the on the probability of risky behaviour

	(1)	(2)	(3)	(4)	(5)	(6)
	RDD			LPM		
	Cutoff 3	Cutoff 4	Cutoff 5	Cutoff 3	Cutoff 4	Cutoff 5
Being bullied	-.09	.008	-.04	-.20	-.09	-.09
	(.11)	(.04)	(.02)*	(.04)***	(.01)***	(.01)***
Unauthorised absence	.05	-.05	-.01	.02	-.05	-.05
	(.07)	(.03)	(.02)	(.04)	(.01)***	(.008)***
1 month absence	-.05	-.004	-.01	-.03	-.02	-.01
	(.05)	(.01)	(.009)	(.03)	(.007)***	(.004)***
Suspension	.05	.03	-.02	-.008	-.07	-.04
	(.08)	(.02)	(.02)	(.04)	(.01)***	(.006)***
Expulsion	.02	-.003	.005	-.03	-.006	-.0004
	(.02)	(.005)	(.002)**	(.02)	(.002)***	(.001)
Police warning	.08	.01	-.002	.03	-.03	-.03
	(.06)	(.02)	(.01)	(.03)	(.009)***	(.005)***
Observations	1920	7656	8619	1808	7233	8158
RDD bandwidth	.32	.42	.69			

Note: The table shows in columns (1)-(3) regression discontinuity design (RDD) estimates of the effect of meeting a performance target in the average score in Key Stage 2 tests on the probability of risky behaviour by students. The outcomes are dummies equal to 1 if a student's main parent answered "yes" to a question on the probability that students engage in risky behaviour and zero otherwise, and they are measured over a time window that spans between one year before the interview date in the survey (Being bullied and the measures of absence) and three years (suspension, expulsion and police warning). I regress an outcome variable on smooth polynomials in test scores separately for students to the left and right of each of score target $\bar{T}$: 3, 4 and 5, that the Department for Education in the UK set. The running variable is the average score over tests in English, Maths and Science. I use a window that contains observations in the interval $[\bar{T}-1, \bar{T}+1]$, i.e. the cutoffs to the left and to the right of $\bar{T}$. I use the choice rule in Imbens and Kalyanaraman (2009) to obtain the bandwidth of the polynomial. The table also shows in columns (4)-(6) linear probability model (LPM) estimates of the effect of meeting a performance target, that are obtained by also including in the regressions covariates on gender, ethnicity, type of schools, assignment to government support programs and school characteristics, whose summary statistics are in Table 2. The estimates are robust to dropping the observations of those students who were interviewed after they had learnt about their performance in Key Stage 2 tests, as the robustness checks in Table A-3 in the appendix suggest. The significance levels are as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Section 2 offers additional information on the institutional setting and on the data, section 3 on the research design and section 4 on the empirical analysis. Summary statistics of the outcome variables are in Table 2.

Table 6: Regression discontinuity design estimates of the effect of meeting performance targets in the average tests score on the probability of risky behaviour by students' baseline characteristics

	Gender						White						Community school						Main parent's GCSE+					
	Female			Male			No			Yes			No			Yes			No			Yes		
	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5
Being bullied	-.20	.05	-.05	.01	-.02	-.05	-.14	.10	-.34	-.01	-.02	.03	-.03	-.008	.04	-.09	.01	-.10	-.06	-.03	-.01	-.18	.03	-.07
	(.11)*	(.05)	(.04)	(.12)	(.05)	(.05)	(.20)	(.06)	(.07)***	(.10)	(.05)	(.04)	(.25)	(.07)	(.06)	(.11)	(.04)	(.04)**	(.13)	(.06)	(.06)	(.12)	(.05)	(.04)*
Unauthorised absence	-.04	-.02	.03	.12	-.07	-.06	.16	-.02	-.03	.003	-.06	-.006	-.13	.03	-.05	.11	-.08	.0002	.02	-.07	-.002	.09	-.02	-.02
	(.12)	(.04)	(.03)	(.10)	(.05)	(.02)**	(.12)	(.04)	(.04)	(.10)	(.04)	(.02)	(.14)	(.05)	(.04)	(.09)	(.04)*	(.02)	(.09)	(.05)	(.04)	(.11)	(.04)	(.02)
1 month absence	-.01	-.02	-.005	-.07	.02	-.02	-.10	-.04	-.002	-.02	.02	-.006	-.24	-.008	-.02	-.06	.004	-.002	-.04	-.009	.03	-.03	.005	-.02
	(.07)	(.02)	(.01)	(.07)	(.02)	(.01)	(.08)	(.03)	(.02)	(.07)	(.02)	(.008)	(.21)	(.02)	(.01)	(.05)	(.02)	(.01)	(.07)	(.03)	(.02)	(.09)	(.02)	(.009)**
Suspension	.32	-.006	.005	-.03	.10	-.04	.13	.12	-.03	.02	.02	-.001	.003	.03	-.03	.06	.02	-.01	.06	.02	-.01	.006	.06	-.02
	(.13)**	(.02)	(.01)	(.13)	(.04)**	(.03)	(.12)	(.05)**	(.03)	(.11)	(.03)	(.01)	(.22)	(.04)	(.02)	(.08)	(.02)	(.02)	(.10)	(.04)	(.03)	(.12)	(.03)*	(.02)
Expulsion	.003	.005	.004	.02	-.01	.006	.01	-.01	.009	.005	-.003	.004	-.03	-.0002	-.0002	.02	-.002	.008	.02	-.01	.01	.004	.004	.004
	(.02)	(.004)	(.004)	(.03)	(.01)	(.003)*	(.01)	(.02)	(.005)*	(.03)	(.003)	(.003)	(.09)	(.01)	(.01)	(.01)	(.006)	(.004)**	(.02)	(.01)	(.01)	(.05)	(.005)	(.002)*
Police warning	.15	-.03	.003	.07	.05	-.008	-.03	-.01	-.03	.10	.03	.006	.12	.04	-.06	.06	.004	.03	.04	-.03	.009	.13	.05	-.006
	(.10)	(.03)	(.01)	(.08)	(.03)*	(.02)	(.08)	(.03)	(.03)	(.08)	(.02)	(.02)	(.12)	(.04)	(.02)**	(.06)	(.02)	(.02)	(.07)	(.03)	(.03)	(.08)*	(.03)*	(.02)
Observations	890	3828	4344	1030	3828	4275	778	2672	2688	1142	4984	5931	461	2203	2766	1459	5453	5853	1074	3137	2731	846	4519	5888
Bandwidth	.29	.83	.28	.45	.7	.47	.31	.66	.37	.45	.61	.38	.31	.42	.58	.35	.66	.43	.41	.61	.41	.33	.63	.28

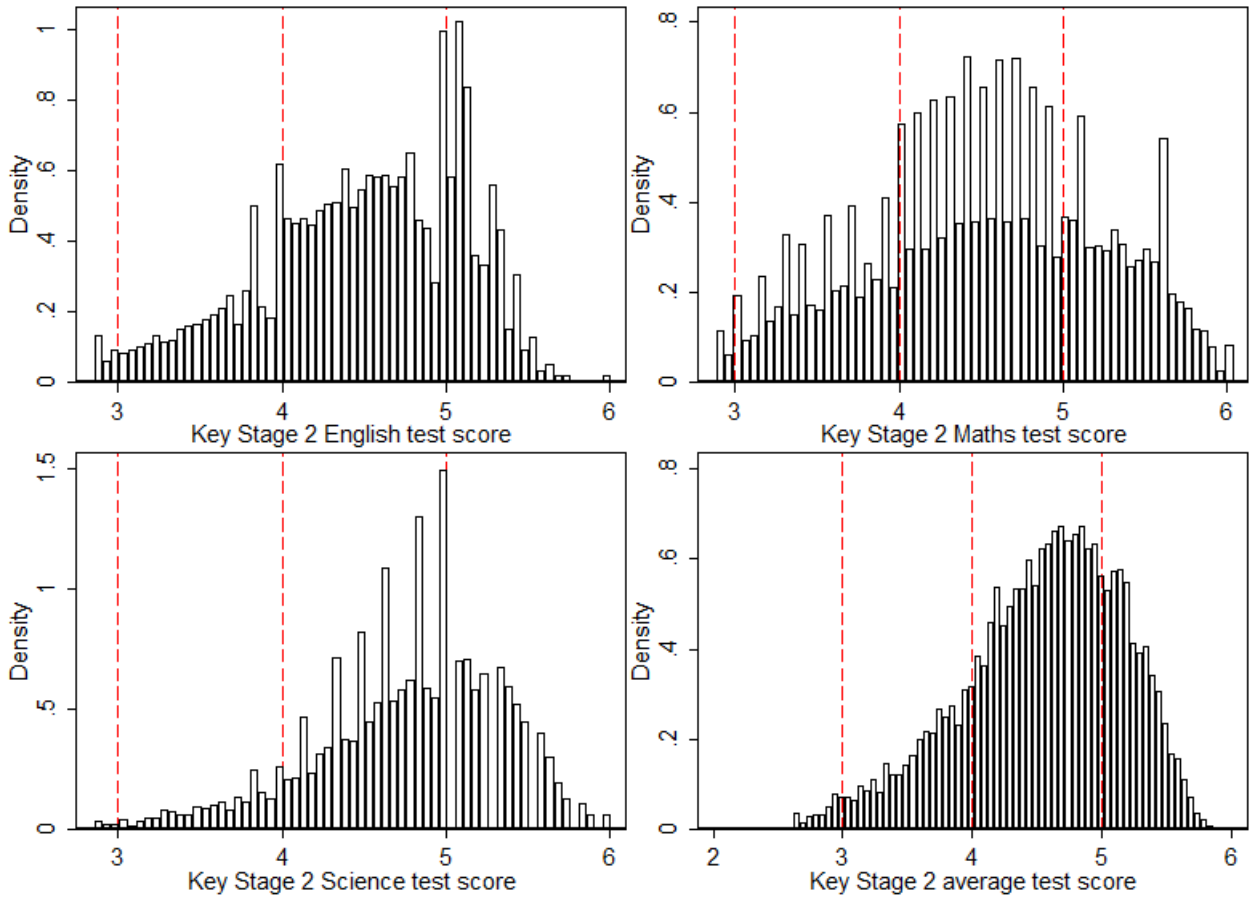
Notes: *i*) The table shows regression discontinuity design (RDD) estimates of the effect of meeting a performance target in the average score in Key Stage 2 tests on the probability that students engage in risky behaviour separately for sub-samples of students with different baseline covariates: whether a student is male, belongs to the white ethnic group, went to a community school at Key Stage 2 and the main parent completed at least secondary education (GCSE). The outcomes are dummies equal to 1 if a student's main parent answered "yes" to a question on the probability that students engage in risky behaviour and zero otherwise, and they are measured over a time window that spans between one year before the interview date in the survey (Being bullied and the measures of absence) and three years (suspension, expulsion and police warning). I regress an outcome variable on smooth polynomials in test scores separately for students to the left and right of each of score target $\bar{T}$: 3, 4 and 5, that the Department for Education in the UK set. The running variable is the average score over tests in English, Maths and Science. I use a window that contains observations in the interval $[\bar{T}-1, \bar{T}+1]$, i.e. the cutoffs to the left and to the right of $\bar{T}$. I use the choice rule in Imbens and Kalyanaraman (2009) to obtain the bandwidth of the polynomial. The estimates are robust to dropping the observations of those students who were interviewed after they had learnt about their performance in Key Stage 2 tests, as the robustness checks in Table A-3 in the appendix suggest. The significance levels are as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Section 2 offers additional information on the institutional setting and on the data, section 3 on the research design and section 4 on the empirical analysis. Summary statistics of the outcome variables are in Table 2.

Table 7: Regression discontinuity design estimates of the effect of meeting performance targets in the average test score on the probability of risky behaviour by students' assignment to support programs at Key Stage 2

	EAL						SEN						SEN statemented						FSM					
	No			Yes			No			Yes			No			Yes			No			Yes		
	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5
Being bullied	-.08	-.03	-.05	-.17	.15	-.09	.08	.03	-.07	-.23	-.03	-.05	-.06	-.02	-.05	-.14	.02	-.06	.18	-.08	-.04	-.28	.12	-.07
	(.10)	(.04)	(.03)	(.20)	(.07)**	(.07)	(.12)	(.05)	(.04)	(.14)	(.06)	(.05)	(.12)	(.04)	(.04)	(.15)	(.08)	(.06)	(.18)	(.05)*	(.03)	(.12)**	(.06)**	(.05)
Unauthorised absence	.03	-.05	-.02	.10	-.007	-.008	.08	-.05	-.02	.09	-.02	.01	.08	-.04	.004	-.005	-.02	-.03	.09	-.08	-.02	.09	-.01	-.002
	(.09)	(.03)	(.02)	(.13)	(.05)	(.05)	(.14)	(.04)	(.03)	(.08)	(.04)	(.03)	(.09)	(.04)	(.03)	(.12)	(.04)	(.03)	(.12)	(.04)*	(.02)	(.11)	(.04)	(.04)
1 month absence	-.07	-.002	-.01	-.009	.02	-.01	-.08	.01	-.02	-.03	-.02	-.007	-.05	.006	-.001	-.05	-.02	-.01	-.05	-.01	-.003	-.06	.01	-.02
	(.07)	(.02)	(.009)	(.05)	(.02)	(.02)	(.08)	(.02)	(.01)	(.06)	(.02)	(.01)	(.06)	(.02)	(.006)	(.08)	(.03)	(.02)	(.07)	(.02)	(.01)	(.07)	(.02)	(.02)
Suspension	.04	.02	-.01	-.09	.10	-.01	.03	-.04	-.01	.13	.14	-.02	-.07	.08	-.02	.31	.04	-.01	.06	-.005	-.007	.05	.08	-.03
	(.09)	(.02)	(.01)	(.20)	(.06)*	(.04)	(.13)	(.03)	(.02)	(.14)	(.04)***	(.03)	(.14)	(.03)**	(.01)	(.15)**	(.04)	(.04)	(.12)	(.02)	(.01)	(.12)	(.04)**	(.03)
Expulsion	.02	.002	.005	-.0002	-.02	.005	.03	-.005	.002	-.0004	-.0004	.006	-.02	-.003	.002	.05	-.002	.01	.02	-.002	.006	.003	-.005	-.005
	(.03)	(.004)	(.003)*	(.001)	(.02)	(.004)	(.03)	(.006)	(.001)	(.03)	(.008)	(.005)	(.03)	(.008)	(.001)**	(.03)*	(.005)	(.009)	(.02)	(.005)	(.003)*	(.03)	(.01)	(.01)
Police warning	.15	.007	.005	-.08	.05	-.03	-.05	.005	.003	.13	.01	-.008	.04	.004	.006	.18	.03	-.02	.10	.01	-.002	.10	.01	.001
	(.07)**	(.02)	(.02)	(.09)	(.04)	(.02)	(.09)	(.03)	(.02)	(.07)*	(.02)	(.02)	(.07)	(.02)	(.02)	(.12)	(.04)	(.03)	(.07)	(.02)	(.02)	(.10)	(.03)	(.02)
Observations	1269	5531	6570	450	1604	1667	789	3822	4804	930	3313	3433	1126	4955	5946	593	2180	2291	859	4350	5583	860	2785	2654
Bandwidth	.31	.81	.39	.43	.36	.32	.31	.73	.34	.48	1.04	.57	.39	.6	.51	.35	.56	.56	.47	.63	.29	.3	.3	.45

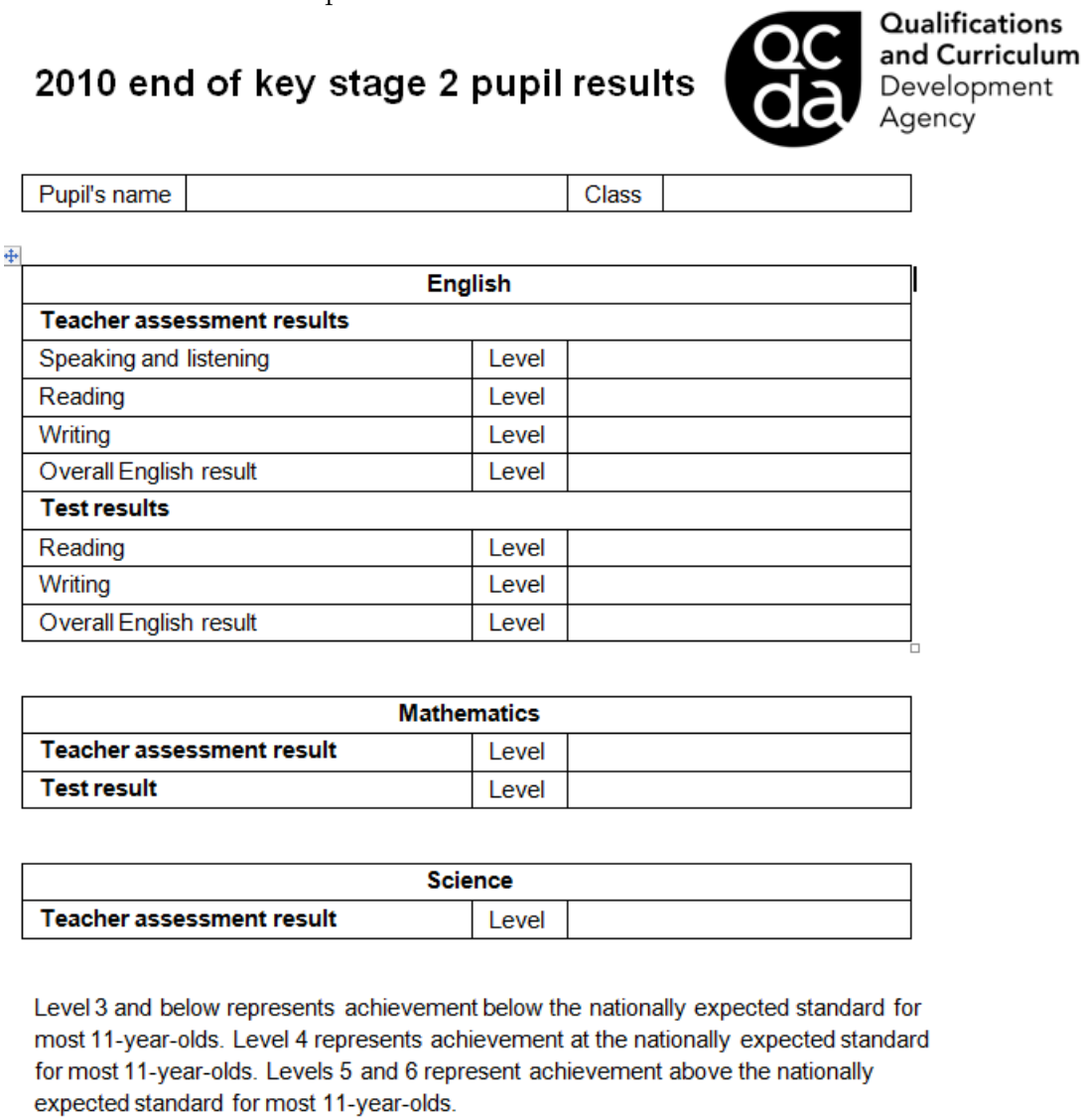
Notes: *i*) The table shows regression discontinuity design (RDD) estimates of the effect of meeting a performance target in the average score in Key Stage 2 tests on the probability that students engage in risky behaviour separately for sub-samples of students by assignment to government support programs in schools: whether a student has English as Additional Language (EAL), Special Education Needs (SEN), special needs of high level (SEN statemented) or Free School Meal (FSM). The outcomes are dummies equal to 1 if a student's main parent answered "yes" to a question on the probability that students engage in risky behaviour and zero otherwise, and they are measured over a time window that spans between one year before the interview date in the survey (Being bullied and the measures of absence) and three years (suspension, expulsion and police warning). I regress an outcome variable on smooth polynomials in test scores separately for students to the left and right of each of score target $\bar{T}$: 3, 4 and 5, that the Department for Education in the UK set. The running variable is the average score over tests in English, Maths and Science. I use a window that contains observations in the interval $[\bar{T}-1, \bar{T}+1]$, i.e. the cutoffs to the left and to the right of $\bar{T}$. I use the choice rule in Imbens and Kalyanaraman (2009) to obtain the bandwidth of the polynomial. The estimates are robust to dropping the observations of those students who were interviewed after they had learnt about their performance in Key Stage 2 tests, as the robustness checks in Table A-3 in the appendix suggest. The significance levels are as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Section 2 offers additional information on the institutional setting and on the data, section 3 on the research design and section 4 on the empirical analysis. Summary statistics of the outcome variables are in Table 2.

Figure 1: Histograms of test scores in English, Maths, Science tests and of the average test score at Key Stage 2



Notes: The figure shows histograms of scores in tests at Key Stage 2 with bin size equal to 0.05, as well as cutoffs in test scores as vertical and dashed lines at values 3, 4 and 5 on the horizontal axis. The scores can take values in the interval $[2.5, 6]$. The histograms are useful to detect if the distribution of test scores is smooth at cutoff values, hence supporting the validity of a regression discontinuity design to identify the effect of meeting a test score target on the probability of behaviour or vice versa. Visual inspection of the size of the bins of histograms at each cutoff suggests that tests in English, Maths and Science may be manipulated. However, the manipulation is mechanical rather than substantial since the external markers who are in charge of grading tests were instructed by the Department for Education to revise tests in a narrow neighbourhood of a cutoff during the marking process (Quinlan and Scharaschkin (1999)). The average test score is instead not manipulated since each of the three test that a student sat is marked by a different examiner. Section 2 offers additional information about the institutional setting and section 4 discusses the validity of the research design.

Figure 2: Tests results sheet 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 results sheet that schools currently use to disclose achievement levels in tests in English, Maths and Science at Key Stage 2 to students and parents. The template is similar to the one that schools used in 2001 except that Science tests have not been externally marked since 2010. The achievement level is a categorical measure of achievement that can take one of the following values: 2, 3, level 4 that the Department of Education expects all students to achieve, or 5. The paragraph at the bottom of the sheet reports details of the performance targets that the Department for Education set for students in tests at Key Stage 2. The underlying scores in each test are measured on a continuous scale and they are not disclosed, thus offering a research design to identify the effect of meeting an achievement level or target on students' risky behaviour. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.

6 Appendix

This section contains supplementary material about the description of the institutional setting in compulsory education in England and of the data in section 2 in the paper, and about the results in the empirical analysis in section 4.

Table A-1: Questions to main parent in the questionnaire

(1) Variable Name	(2) Question	(3) No. of years before the survey date and in which behaviour may be observed
Truancy	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 survey dataset which I use as outcome variables in the empirical analysis. Column (2) shows the wording of the questions in the survey questionnaire. Column (3) shows the number of years before the survey date in which students' behaviour may be observed. All questions are answered by a student's main parent except the one on unauthorised absence which is answered by the student. Main parent is defined as "the parent most involved in the young person's education" in the survey questionnaire. NatCen (2010) offers additional information about the survey. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.

Table A-2: T-statistics from tests of the null hypothesis of no manipulation in the average test score at a cutoff

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

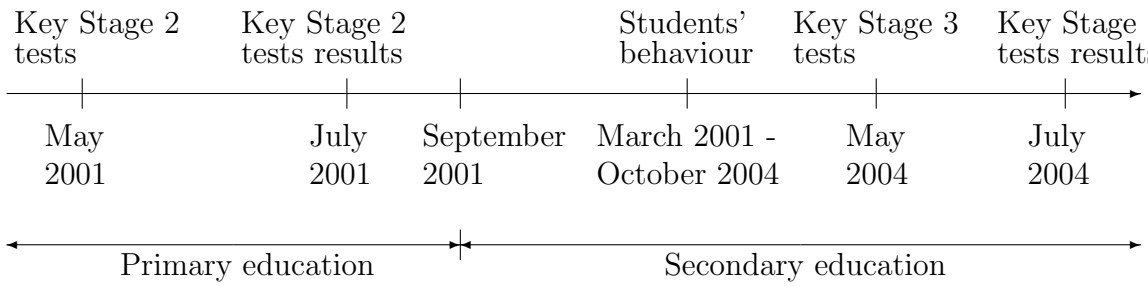
Notes: The table shows t-statistics of the null hypothesis of no manipulation of a test score at cutoffs, i.e. smoothnes in the distribution of the score at a cutoff, in a regression discontinuity design to identify the effect of meeting performance targets 3, 4 and 5 in test scores on the probability of risky behaviour (McCrary (2008)). The top three rows in the table show t-statistics for the null hypothesis of no manipulation of test scores in each test subject: English, Maths and Science. The last row shows t-statistics of tests of the average test score. The test does not reject the null hypothesis if the difference in the height of the histogram bins that are estimated separately for observations to the left and to the right of a cutoff is sufficiently small. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset and Section 4 discusses the validity of the research design.

Table A-3: Sensitivity of regression discontinuity design estimates of the effect of meeting a performance target on the probability of risky behaviour to the survey data collection dates relative to the disclosure of test scores

	Suspension			Police warning		
	Cutoff 3	Cutoff 4	Cutoff 5	Cutoff 3	Cutoff 4	Cutoff 5
In/after April	.03	.03	-.02	.10	.02	-.007
	(.06)	(.02)	(.02)	(.06)	(.02)	(.01)
Obs.	1667	6907	7951	1656	6868	7919
In/after May	-.08	.01	-.01	.13	.03	.002
	(.07)	(.02)	(.02)	(.08)**	(.03)	(.02)
Obs.	1222	5073	5839	1215	5047	5815
In/after June	-.05	.04	.004	.13	.06	-.02
	(.09)	(.03)	(.02)	(.08)**	(.03)**	(.02)
Obs.	772	3196	3701	768	3182	3690
In/after July	-.14	.08	.03	.31	.11	.01
	(.11)	(.05)	(.04)	(.09)***	(.04)***	(.03)
Obs.	372	1531	1762	372	1523	1752
In/after August	.002	.12	-.04	.08	.17	.0006
	(.07)	(.08)	(.05)	(2.83e-15)***	(.10)**	(.04)
Obs.	130	543	627	129	536	620

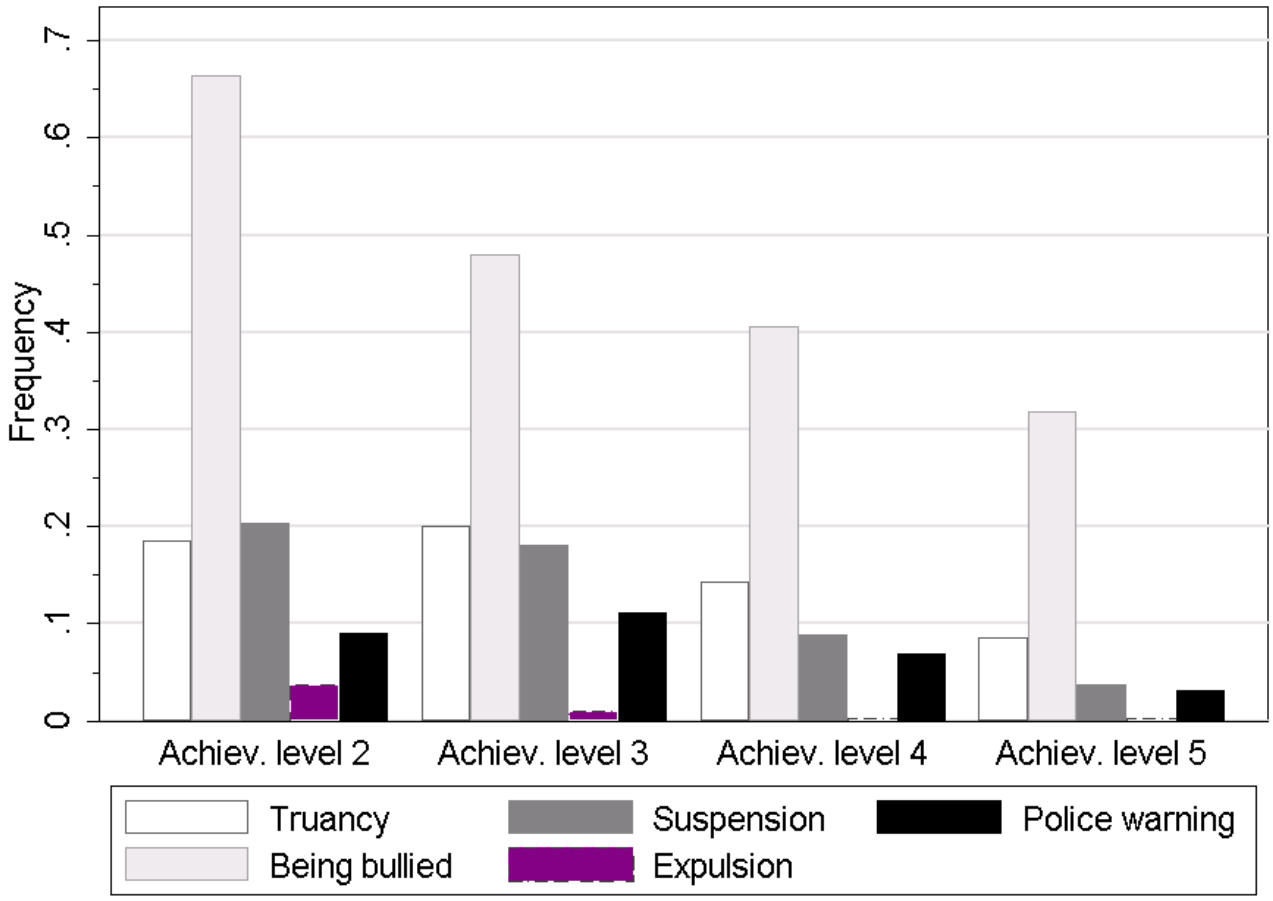
Notes: The table shows separate regression discontinuity design (RDD) estimates by using sub-samples of observations which differ by the period in the year in which the survey data on students' risky behaviour were collected with respect to estimates obtained from the full data sample in Table 5. The outcomes are dummies equal to 1 if a student's main parent answered "yes" to a question on the probability that students engage in risky behaviour and zero otherwise, and they are measured over a time window that spans between one year before the interview date in the survey (Being bullied and the measures of absence) and three years (suspension, expulsion and police warning). The research design is valid if the overlap between the time period in which test scores are disclosed in July 2001 and the May 2001-October 2003 time window to which the events in the survey data refer to is small and hence estimates do not suffer from reverse causality. This may occur only for the *Suspension* and *Police warning* and is not relevant empirically if estimates from the full samples and those from sub-samples by survey month are similar. RDD estimates in the table are equal to the difference in the probability of enaging in risky behaviour between those students whose average test score is to the right of a cutoff value, that the Department for Education set at Key Stage 2, and those students whose score is to the left of the cutoff. I regress an outcome variable on smooth polynomials in test scores separately for students to the left and right of each of score target $\bar{T}$: 3, 4 and 5, that the Department for Education in the UK set. The running variable is the average score over tests in English, Maths and Science. I use a window that contains observations in the interval $[\bar{T}-1, \bar{T}+1]$, i.e. the cutoffs to the left and to the right of $\bar{T}$. I use the choice rule in Imbens and Kalyanaraman (2009) to obtain the bandwidth of the polynomial. Significance levels are as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Section offers 3 offers additional information on the research design and section 4 on the results in the empirical analysis. Summary statistics of the outcome variables are in Table 2.

Figure A-1: Timeline of students' tests at Key Stage 2 and of the collection of survey data risky behaviour after the disclosure of tests results



Note: The figure shows the timeline of the events and choices that students face from Key Stage 2 tests in May 2001 onwards. 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 the 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 A-2: Probability of risky behaviour by students' average achievement level in tests at Key Stage 2



Note: The figure shows the mean probability that students engage in different types of risky behaviour by the categorical achievement level in tests at Key Stage 2 that ranges from 2 to 5. Each achievement level is defined by using cutoffs 3, 4 and 5 in the average test score at Key Stage 2. Section 2 offers additional information on the institutional setting of compulsory education in England and on the linked dataset.