Journal of Public Economics 190 (2020) 104224

Contents lists available at ScienceDirect

Journal of Public Economics

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

Entry through the narrow door: The costs of just failing high stakes exams

Stephen Machin [a], Sandra McNally [b,c,*], Jenifer Ruiz-Valenzuela [c]

[a] Department of Economics and Centre for Economic Performance, London School of Economics, Houghton Street, WC2A 2AE London, United Kingdom
[b] School of Economics, University of Surrey, United Kingdom
[c] Centre for Economic Performance, Centre for Vocational Education Research, London School of Economics, Houghton Street, WC2A 2AE London, United Kingdom

ARTICLE INFO

Article history:
Received 13 June 2019
Received in revised form 5 June 2020
Accepted 5 June 2020
Available online xxxx

Keywords:
High stakes examinations
Manipulation
English

JEL classifications:
I20
I21
I24

ABSTRACT

In many countries, important thresholds in examinations act as a gateway to higher levels of education and/or improved employment prospects. This paper examines the consequences of just failing a particularly important high stakes national examination taken at the end of compulsory schooling in England. It uses unique administrative data, including full information on both initial and regraded exam marks, to show that students of the same ability have significantly different educational trajectories depending on whether they just pass or fail this exam. Three years later, students who just fail to achieve the required threshold have a lower probability of entering an upper-secondary high-level academic or vocational track and of starting tertiary education. Those who fail to pass the threshold are also more likely to drop out of education by age 18, without some form of employment. The moderately high effects of just passing or failing to pass the threshold in this high-stakes exam has high potential long-term consequences for those affected.

1. Introduction

The long-term economic and social impacts of passing or failing exams have long been of interest. Narrowly passing or failing exams and the effect on individuals with highly similar abilities and observable characteristics has recently been a prominent focus for research. This has come about, at least in part, because the availability of rich administrative data from different settings permits study of scoring just above or just below a particular threshold. Examples include different degree classifications, acquiring a high school diploma or reaching a certain grade point average, to name just a few.

Why is honing in on near passes and failures important? One reason is that, in many contexts, achievement of particular exam thresholds, especially those leading to key qualifications, is seen as vital by educators, employers and governments. This makes just passing the threshold a substantively interesting outcome to consider. Moreover, in decentralised education systems where mechanisms like pay for performance operate and where school rankings are important, exam thresholds can play a role in incentivising teachers and school managers. In such contexts, the 'pass/fail' threshold may have longer-term consequences for students with approximately the same marks and level of effort. As shown by Costrell (1994), under quite general conditions, even if educational standards are chosen optimally, it may be welfare enhancing to improve information flows (i.e. on the underlying marks) such that achievement of a binary threshold (the standard) becomes less important. In some contexts (such as the one considered here) it is also important to consider the counterfactual for students who do not pass the threshold because individuals are still very young and expected to proceed to upper secondary education.

This paper offers an empirical study of the consequences of just passing or failing a particularly important national high stakes exam using a (fuzzy) regression discontinuity design. The context is the national examinations taken by all students at the end of compulsory schooling in England. More specifically, evidence is presented on the importance of just obtaining a grade C in English – a good pass – in high stakes national examinations taken for the General Certificate of Secondary Education (or GCSEs) when students are 16 years of age.[1] Arguably, this is much more important than obtaining certain grade scores in other countries. Although external examinations at the end of compulsory

[*] Corresponding author at: School of Economics, University of Surrey, United Kingdom.
E-mail addresses: s.j.machin@lse.ac.uk (S. Machin), s.mcnally1@lse.ac.uk (S. McNally), j.ruiz-valenzuela@lse.ac.uk (J. Ruiz-Valenzuela).

[1] GCSEs grades are awarded on a scale of A*-G where fails are given the letter U. Marks are the overall points received in the subject. For GCSE English and the cohort under study, marks can take values from 0 to 300. More details can be found in Section 2. We focus on English rather than maths because we have detailed data on English marks for an exam board that accounts for over half of exams in English (discussed later in the paper and in Appendix A).

https://doi.org/10.1016/j.jpubeco.2020.104224

education are not uncommon across countries, the English system places an unusually high weight on the GCSE exam, not least because GCSE indicators appear in published school performance tables. In England, the grade C threshold at the end of lower secondary education can affect access to the quality of upper secondary education available to the student (via the programme or institution) and therefore whether he/she has good preparation for tertiary education. For instance, a grade C is typically necessary to access some qualifications, though not all, at the upper secondary education level. Alongside its use by educators, a grade C in English and/or maths is also important for employers. Since 2015 this level of achievement has been deemed so important that it has become mandatory for students to repeat the school leaving exam if they fail to get a C grade in English or maths and wish to continue in some form of publicly funded education thereafter.

Despite the general acknowledgement that obtaining a C grade in GCSE English and maths matters for future outcomes, there is no causal evidence to date to substantiate this claim and assess consequences for those who narrowly miss. The main empirical challenge is to address potential endogeneity around who passes this threshold. This paper makes use of a novel linked (administrative) dataset with student level information on the distribution of exact marks around the important threshold of grade C. We use this to estimate the effect of just passing the grade C threshold on later outcomes using a fuzzy regression discontinuity design. Specifically, the probability of achieving a grade C is instrumented with the original mark (i.e. before any appeal has been made by the school on behalf of the student to consider a regrade).

The data and methodological challenges have some similarities to two recent papers studying the effects of possible teacher manipulation of exam scores - one studying a national examination in Sweden, the other a high school exit examination from New York. In the Swedish paper, Diamond and Persson (2016) report significant test score manipulation around known grade thresholds in the national mathematics tests taken by ninth graders. In the New York study, Dee et al. (2019) demonstrate that manipulation took place in the New York Regents exam taken by high school students.[2] In both cases, students who cross the threshold benefit in terms of later educational and/or labour market outcomes. Diamond and Persson (2016) find that test score manipulation around a given threshold raises the student's likelihood of achieving a grade in maths which is necessary for admittance to any high school and also find that manipulated students perform better in high school, which in turn translates to higher income at age 23. They argue that these effects, which are particularly pronounced at the higher end of the ability distribution, are driven either by increased self-confidence of students or potentially by signalling to other teachers to subsequently give higher grades as well. Dee et al. (2019) finds that having an exam score manipulated to fall above a performance cutoff has a substantial positive effect on the probability of graduating from high school, but a negative effect on the probability of meeting the requirements for a more advanced high school diploma. They argue that students at the margin of dropping out are 'helped' by not having to retake classes but those on the margin of the advanced diploma are 'hurt' by not being pushed to do necessary preparation for more advanced coursework. Whereas Diamond and Persson (2016) and Dee et al. (2019) have teacher cheating or bias in mind as the underlying mechanism behind grade score manipulation, this is not the case in the English context (as we explain below). Furthermore, while they have to impute a 'pre-adjusted' distribution of marks, we have access both to initial marks and (where relevant) regraded exam marks. This gives a big modelling advantage compared to other papers. Like these other papers, we show that passing the threshold has important consequences for students – but in our case, not because of any teacher manipulation.

There are other papers that analyse the effect of obtaining an important educational signal (as a consequence of luck), but they are for older students and in different educational contexts. The population of interest in these other papers have already selected into post-compulsory education, and therefore their results do not apply to all school aged children, as is the case here. For example, Clark and Martorell (2014) evaluate the signalling value of a high school diploma in the US for earnings later in life. Ebenstein et al. (2016) evaluate the effect of shocks (or bad luck) in the context of high stakes exams in Israel, using transitory variation that comes from pollution exposure. Canaan and Mouganie (2018) study the impact of marginally passing the French high school exit exam on choice of higher education institution and degree subject. There is also a literature that looks at the impact of passing college admission exams on subsequent outcomes using regression discontinuity methods (e.g. Anelli, 2016; Avery et al., 2018; Goodman et al., 2017; Kaufmann et al., 2013; Smith et al., 2017; Zimmerman, 2014).[3]

The findings reported in this paper show that failing to achieve a grade C in English has a large associated cost. Put another way, the marginal student would have performed significantly better in the longer term had he/she not been so unlucky at this point. This is not necessarily a natural consequence of failing to reach an important threshold in a high stakes exam. For example, Clark and Martorell (2014) find that marginal students do not benefit from marginally passing high school exit exams in the US (as reflected in earnings). Furthermore, many of the outcomes considered here are achieved by the vast majority of students of this ability level (i.e. we are not, for example, talking about rationing of places in elite institutions).

In our context, students who just fail to obtain a grade C in English are more likely to drop out of education early and become classified as 'not in education, employment or training' (or NEET) at age 18. They are much less likely to have entered a high-level course in upper secondary education up to 3 years after having sat the GCSE exams, by the age of 19 (which is the age by which most English students will have entered upper secondary education if they are going to start at all). This is in spite of the fact that there are opportunities to enrol in other courses (and in principle progress to the next level) and at the same time retake GCSE exams in subsequent academic years.[4] We also find that students are less likely to enter tertiary education by the age of 19. All these indicators make poor employment and earnings prospects more likely in the longer term.

Evidence on the mechanisms through which failing to obtain a grade C in English leads to poor outcomes is presented. These involve a narrowing of opportunities that arise within the educational system on the choice of post-16 institution and course the year after failing to get a C grade in GCSE English. This does not mean that students are permanently excluded from particular institutions or courses. Those who marginally fail should be able to progress in their education and get back on track the subsequent year. However, we find that a significant minority fail to do so as much as 3 years after the event. In addition, students end up in institutions with a worse academic environment (as measured by peer quality).

Compared to the previous literature, this paper is the first to offer insights on the negative consequences of marginally failing to reach an

[2] Several other recent papers that involve analysing the consequences of teacher/examiner bias in high-stakes exams for student outcomes (such as Apperson et al., 2016 and Borcan et al., 2017) examine the effect of teacher bias in marking more generally (e.g. Angrist et al., 2017; Lavy and Sand, 2018; Terrier, 2016). Battistin and Neri (2017) is another paper concerned with manipulation of test scores in an English context. They use an anomaly in the marking system with regard to primary schools in England (which existed prior to 2007) to identify the relationship between (randomly induced) signalling in test scores and house prices. They show that publicly available information on test scores yields a significant house price differential.

[3] Other related examples include the effects of class of degree on earnings (e.g. Feng and Graetz, 2017; Freier et al., 2015); how test score labels affect human capital investment decisions (Papay et al., 2015) and how individuals' choice of educational quality—measured by college reputation—may likewise signal their ability (MacLeod et al., 2017).

[4] However, the pass-rate for those students re-taking the GCSE English exam in our cohort is very low. We offer more details in Section 2.

institutionally set threshold in national high-stake exams taken by the overall student population. This paper suggests that the marginal student who is unlucky pays a high price. This is consistent with descriptive evidence that suggests that the English educational system offers limited prospects for those who leave compulsory education without good grades. For example, Hupkau et al. (2017) show that the probability of progression from lower level to higher level courses is relatively low and several studies also show non-existent wage returns to lower-level courses (Dearden et al., 2002; McIntosh, 2006). The more general message is that failure to pass a 'high stakes' threshold can have very serious consequences for students at the margin in the absence of well-designed courses for those who fail. While the existence of such thresholds may be necessary to incentivise effort by students and institutions, policy makers and administrators have a choice on what other information to provide and on what resources to invest in the upper secondary education of students who do not pass the threshold. In England, students and schools do find out the exact marks (as well as the threshold) but such information is not systematically provided to Further Education providers (i.e. the institution to which students move to for their upper secondary education). As the new institution does not know who the marginal student is, it may be difficult to offer appropriate support.

The rest of the paper is structured as follows. First, we provide some information on the institutional background of relevant parts of the education system in England, with a special focus on the school-leaving exams, the empirical distribution of pre-appeal and post-appeal marks, and a descriptive analysis of who gets regraded (Section 2). Then we discuss the research design and discuss its validity (Section 3), before presenting our results (Section 4); and discussing the potential mechanisms and implications (Section 5). We conclude in Section 6.

2. Grades in high stakes examinations

2.1. End of school-leaving examinations

In its compulsory phases, the English education system is organised into four Key Stages (KS). There are external assessments at the end of primary school (at Key Stage 2) and at the end of compulsory full-time education (at Key Stage 4 – the GCSE examinations), when students are aged 16 (as grade repetition very rarely occurs). The typical student takes 8–10 GCSE exams and it is compulsory to sit exams in English, maths and science. After this time, most students pursue post-secondary courses for at least two years, which may be at the same school or in an institution specialising in academic education (e.g. Sixth Form Colleges) or in vocational education or some combination of vocational and academic courses (typically Further Education Colleges). The cohort considered here was the first under an obligation to stay in some form of education (which can be part-time) up to the age of 17. In practice, most students were already doing this, though drop-out is more common at age 18.

GCSEs are marked on a scale of A*-G where fails are given the letter U. A 'good' grade at GCSE is regarded as being at least a C, with particular emphasis on achieving this standard in English and maths. Students who do not get a grade C may re-sit GCSE exams in these subjects in subsequent academic years. Since 2015 it has been compulsory for students who do not achieve a grade C in English or maths to re-sit the exam, but the cohort considered here was not affected by this reform. In this cohort, 29.3% of the students failing to get a grade C retook the exam up to two years later, with 45% managing to secure a grade C or above (Department for Education, 2016).

GCSE exams are set and marked by different exam boards – of which there are four in England.[5] There is a regulator (the Office of Qualifications and Examinations Regulation, Ofqual) that is responsible for ensuring that standards are maintained across boards and over time. A number of assessment units feed into the overall GCSE grade in English. Some of these are teacher assessed (and moderated by the exam board) and some are based on a standardised exam that is corrected (anonymously) by external examiners that perform online marking on separate questions of the exam (not the whole script). Exams take place after the coursework assessment (usually at the end of the school year). In the year of relevance to our study (2013), 40% of the overall marks were accounted for by the standardised exam. Crucially, for teacher-assessed units, teachers are not given advance information on how raw marks on the different assessment units are translated to the 'unified marking scheme' (UMS) which is the format of the final marks (and is on a scale of 0–300; where 180 is the threshold of a C grade). Marks vary from year to year on the various units that make up a student's overall assessment. Furthermore, grade boundaries are not decided in advance of the exam. This is decided by an external committee that engages in a process of inspecting papers (e.g. comparing them to previous years) and statistical analysis (more detail is given in Appendix A). Thus, it is not possible for teachers to manipulate coursework assessments such that the marginal student just crosses the threshold for a grade C. Moreover, the exam board issues strict grading guidelines for units that are teacher assessed, and this marking can also be subject to reviews if inconsistencies are detected.

After the standardised exam, requests for a re-mark of scripts can only come through the school (i.e. not from the individual student) and at a price of roughly £40 per script. At this point, there is a possibility that different schools will vary in their propensity to request re-grading for marginal students. In 2013, there were appeals for about 2% of all GCSE exams, with about one in six appeals leading to a grade change (Office of Qualifications and Examinations Regulation, 2013).

2.2. English language grades

We use administrative data on the census of school students in state schools where we have information as they progress through different stages of education. We use data from students who were in their final year of compulsory education, undertaking their GCSE exams in June 2013 (when they were aged 16). We use data on the grades in their various GCSE exams, their prior attainment (e.g. test scores in their national Key Stage 2 exams taken at age 11), the school attended, and some personal characteristics such as gender, eligibility for free school meals, ethnicity and whether English is spoken as a first language. We are able to follow students up to three years later, as they pursue upper-secondary post-compulsory education ('Key Stage 5') and we observe whether they enrol in any form of tertiary education by the age of 19. We link the education data to administrative data on employment and self-employment from the Longitudinal Educational Outcomes data set (LEO). Appendix A offers a thorough description of the data sources used, as well as describing the sample selection criteria and construction of variables.

We are able to merge the GCSE exam grade in English to information on pre-review and post-review marks from one of the four exam boards, the AQA. This exam board accounts for well over half of all exam entries in GCSE English (61.6% of GCSE English Language entries, and 55.7% of GCSE English entries; see Table A1 and Section A1 in Appendix A for more details on these qualifications).[6] To ensure we are considering only those students taking the same assessment, we focus on the form of English exam that is undertaken by 72% of students ('English Language') and on those students taking the higher tier exam within this group (77% of students). However, we observe similar

[5] There has been a variety of exam boards in the UK since at least the early 1900s, with some modifications over time as the education system has changed. They have regional roots but are nationwide.

[6] Analysis about awarding bodies suggests that schools choose exam boards predominantly based on the perceived quality of the syllabus on offer and seldom change providers (Frontier Economics, 2015). Media reports suggest that perceptions of difficulty are relevant. https://www.theguardian.com/education/2009/aug/25/teachers-choosing-exam-boards-gcse.

patterns if we consider the other type of English exam which students might sit as an alternative and if we consider those taking the lower tier (English language) exam paper.

The characteristics of entrants sitting the GCSE English Language examination with AQA in June 2013 are shown in Table 1 (column 2). We compare their characteristics with those for the whole cohort of students that sat GCSE English Language in June 2013 (column 1). Even though they perform slightly better (students in our sample are about 2 percentage points more likely to achieve a C grade or above), they are very similar in terms of predetermined characteristics to all students in the cohort. In columns (3) and (4), we focus on the students that are of main interest for this paper: those in the C-D range. For the reasons outlined above, we divide students in the C-D range into those that sat the Higher Tier paper (column 3) and those that sat the Foundation exam paper (column 4). As is expected, higher tier students are much better performing than lower tier students: whereas 85% of higher tier students achieve a C grade, only 57.5% do so in the Foundation tier. In terms of predetermined characteristics, higher tier students in the C-D range are more similar to the average student in the cohort. The remaining analysis refers to the higher tier students.

The data used are unique in that both the ‘pre-manipulation’ and ‘post-manipulation’ distributions of marks are available for the same students (i.e. before and after re-marking is requested). We also know who has applied for a re-mark and the outcome of this process. Hence, we can use the data to directly calculate and infer why the distributions differ. Importantly, this has not been possible in other papers looking at related questions where estimation of the counter-factual distribution has been necessary (Dee et al., 2019; Diamond and Persson, 2016).

Fig. 1 shows the final distribution of marks after re-marking has taken place. Specifically, the marks combine the various units of assessment to the ‘unified marking scheme’ (which is on a scale of 0–300; where 180 is the threshold of a C grade). There is clear bunching at the threshold for grade C. In fact, this aspect of the distribution has strong similarities to the exam mark distributions in other countries where manipulation has been identified close to important thresholds (Dee et al., 2019; Diamond and Persson, 2016). In the English context, however, this is not a consequence of teacher bias in marking because teachers do not know how their coursework assessments will contribute to the final mark, nor where the grade boundary will be set. It is also not possible for examiners to manipulate total marks because they correct specific questions rather than whole scripts. However, it may arise from many re-grading requests for students near the boundary. Furthermore, requests for remarking may be non-random with respect to student or school characteristics (which we examine below). Fig. 2a shows the original distribution of marks (i.e. before any review takes place) overlaying the final distribution. This shows that the original distribution of marks is approximately normal (note that there is not a one-to-one mapping between the raw scores and the scaled scores, leading to some lack of smoothness). Fig. 2b zooms in to the area of interest. We test for the presence of manipulation around the C cut-off in both distributions using the test proposed by Frandsen (2017) in the context of regression discontinuity designs with a discrete running variable, since marks only change in increments of 1 point from 0 to 300.[7] As expected, the results of the test under $k = 0$ lead us to reject the null of absence of manipulation in the post-appeal distribution (p-value = 0.000); whereas we cannot reject the null (p-value = 0.489) of absence of manipulation in the original (i.e. pre-appeal) distribution of marks.

2.3. Regrading

As mentioned above and described in Appendix A, we also know the students for whom the school has applied for any kind of review and the outcome of this process. We can use these data to directly calculate and infer why the distributions differ. Reviews can be requested for controlled assessments in unit 3 (teacher assessed unit evaluating ‘extended reading and creative writing’, accounting for a 40% of the overall mark) and for the external exam (unit 1, accounting for another 40% of the mark).[8] Most reviews correspond to remarking requests of the latter (i.e. 70% of review requests in the AQA language sample of higher tier students are due to requests to remark unit 1 – increasing to 74% in the D-C range).

Fig. 3 shows the probability of requesting any kind of review within each original mark. The probability is generally very small but rises close to cut-offs to grade thresholds. This is much more prominent for grade C than for any other grade threshold. For those very close to the grade C threshold (180 marks), the probability of requesting a review is over 60%. In contrast, the probability only rises to about 20% near the thresholds for grades B (210 marks), A (240 marks) and A* (270 marks). This is illustrative of the perceived importance attached to getting a grade C within the English education system. The figure also shows the probability of actually being upgraded. This shows that a high proportion of students for whom a re-mark is requested do not actually cross the relevant threshold, and that crossing it is only likely for those students that originally scored a mark very close to the threshold.

We examine the probability of requesting a review and the conditional probability of being upgraded in Table 2. We use only those students whose original marks were in the range of a C-D grade and we always control for the student's original mark. We regress whether or not a review is made (and an upgrade received) against available student demographics and their achievement in national tests at primary school. Specifically, the variables are whether the student is white; eligible to receive free school meals; speaks English as a first language; female; and the standardised test score in national tests (a composite of English, maths and science) at age 11. The results are similar whether these variables are included separately or together. Column (1) shows results for the Linear Probability Model where the dependent variable is whether any kind of review is requested for a student. In column (2), we re-estimate the regression including school fixed effects. In column (3), the dependent variable is whether the student is upgraded from D to C (conditional on a request having been made) and the regression controls for school fixed effects. Results in all specifications are very similar when we control for the original marks with a quadratic functional form and when we additionally let the slope of the student's original marks to vary on each side of the original C threshold (see Table B1 in the online Appendix). They are also very similar when restricting the sample to those originally scoring very close to the threshold, but under the C threshold (5 points below), and without controlling for the original marks.

The average probability of requesting a review in the C&D range is about 10%. Reviewing of scripts is less likely to be requested for females (by close to 1 percentage point) and more likely to be requested for those with higher scores in primary school. Otherwise, there is no relationship between demographic characteristics and the probability of a review being requested. When school fixed effects are included (column 2), the coefficients decline for both gender and prior attainment (though for the latter it is still precisely estimated and statistically significant). This is likely to reflect the fact that requests for re-marking come via the school and not the

[7] We implement the test using the Stata command *rddisttestk*. See Frandsen (2017) for more details. We choose the parameter k (that determines the maximal degree of nonlinearity in the probability mass function that is still considered to be compatible with no manipulation) to be able to detect manipulation in the most stringent situation (when $k = 0$). As Frandsen (2017) points out, a large k means that the mass at the threshold can deviate substantially from linearity before the test will reject with high probability, while a small k means even small deviations from linearity will lead the test to reject with high probability. Choosing k to be conservatively high will therefore reduce the test's power to detect manipulation.

[8] Unit 2 (‘Speaking and Listening’, accounting for a 20% of the mark for the cohort completing GCSEs in the academic year 2013) cannot be subject to any reviews. See Appendix A (Section A2) for more details.

Table 1
Descriptive statistics.

	(1)	(2)	(3)	(4)
	2013 cohort sitting English Language GCSE	AQA English Language sample	AQA English Language C&D sample - Higher Tier	AQA English Language C&D sample - Foundation Tier
Achieved C or above (Level 2) in GCSE English[a] (%)	81.9	83.8	85.2	57.5
Predetermined characteristics and prior Key Stage 2 performance				
White ethnicity (%)	81.2	79.9	81.1	78.3
Eligible for free school meal (%)	11.1	10.3	10.3	16.7
English spoken at home (%)	88.9	88.2	89.0	86.3
Female (%)	52.9	53.7	48.7	43.6
KS2 total points	70.3	71.1	68.1	60.0
Number of pupils	383,730	189,485	49,231	33,034

Note.

[a] This is calculated from the variable *ks4_lev2eng* in the KS4 Candidate Indicator dataset (more details on the datasets used are given in Appendix A). This indicator includes all qualifications counting towards GCSE English in school performance tables (this includes both GCSE English and GCSE English Language). 2013 cohort: those in the KS4 Candidate/Indicator tables that belong to year group 11 (derived from birth date) and appear in the Census data (i.e. we have data on pre-determined characteristics). Students sitting English Language GCSE in the 2013 cohort are those students that are observed in the 2013 KS4 Results tables as having sat a full GCSE qualification in English Language with any of the awarding bodies. More details about the sample and variable construction are given in Appendix A.

individual. The probability of being upgraded to a C grade (which happens for 12% of students for whom a review is requested in our sample) is not related to any demographic characteristic of students, and only marginally to prior attainment. This is not surprising given that examiners doing the re-marking of the externally examined unit know nothing about the students or the school they attend.

3. Research design and descriptive analysis

3.1. Research design

The institutional setting has imposed an important threshold at grade C from which similar students will fall either side simply because they perform relatively well or badly on the day of assessment. We are interested in establishing the causal effect of getting a C grade (at the end of compulsory education) at age 16 on later outcomes for students who otherwise look the same based on observable characteristics. In other words, what is the effect of getting a C grade in English language GCSE when this is simply a matter of good luck? However, because who enters the appeals process is not a random draw (i.e. schools make a decision to apply for a re-mark in the case of certain students), who ultimately gets a C grade is potentially endogenous. Hence, we need a strategy to overcome this problem.

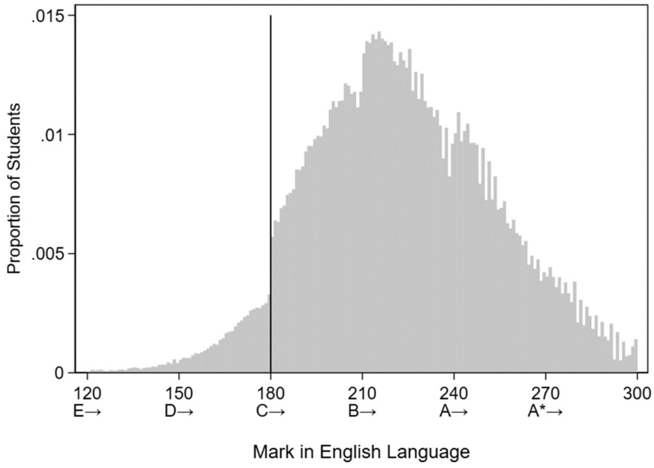

Fig. 1. Final (post-review) distribution of marks. Note. Histogram showing the final (post-appeal) distribution of marks for Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on the data sample construction.

To assess the effect of obtaining a C grade on later outcomes, we make use of the fact that we have the original (pre-review) mark distribution and can use this to build an instrument to predict whether a person actually obtains grade C by the end of their compulsory education (Key Stage 4). Fig. 4 illustrates the first stage and shows that the original mark is a very strong predictor of whether grade C is finally obtained (after the review process). It is not a perfect predictor because of the possibility of re-grading. The probability is 1 beyond the critical threshold for two reasons. First, the pattern just to the right of the C cut-off arises because there is no incentive for schools to enter students for a re-mark if they are just above the threshold, since this is costly and there is a possibility of being downgraded. This is reflected in the pattern of applications throughout the distribution in Fig. 3. Second, this sample only contains students who eventually obtain a grade C or grade D in their English language exam (i.e. it does not contain those who get upgraded from grade C to B).

For students on the left of the C cut-off, the incentive to apply for a re-mark becomes much stronger, the closer the student's original mark is to the C threshold. Thus, to the left of the cut-off, the probability of obtaining a C grade gradually increases from about 10 marks away from the C threshold, whereas to the right of the cut-off, the probability of getting a C grade is 1 (i.e. a partially fuzzy regression discontinuity design (Battistin and Rettore, 2008)).

Given the shape of the first stage, fuzzy regression discontinuity methods (Angrist and Lavy, 1999; Hahn et al., 2001) are used. A dummy indicating whether the student originally obtained a C grade (i.e. pre-review) is used to instrument for whether or not an individual receives a final C grade, in parametric models that control for the original distribution of marks (centred at 180 marks) as the forcing variable. Changes in slope on either side of the cut-off are modelled through an interaction between the forcing variable and the instrument, as suggested by Imbens and Lemieux (2008). We also estimate parametric regressions where we limit the sample to individuals that are very close to the grade C threshold in the original (pre-review) distribution of marks, and non-parametric regressions. We test whether any other observable characteristic of students (such as prior attainment) varies discontinuously at this threshold and show that this can be ruled out.

As Battistin and Rettore (2008) show, the impact of treatment in this partially fuzzy regression discontinuity design can be estimated in a

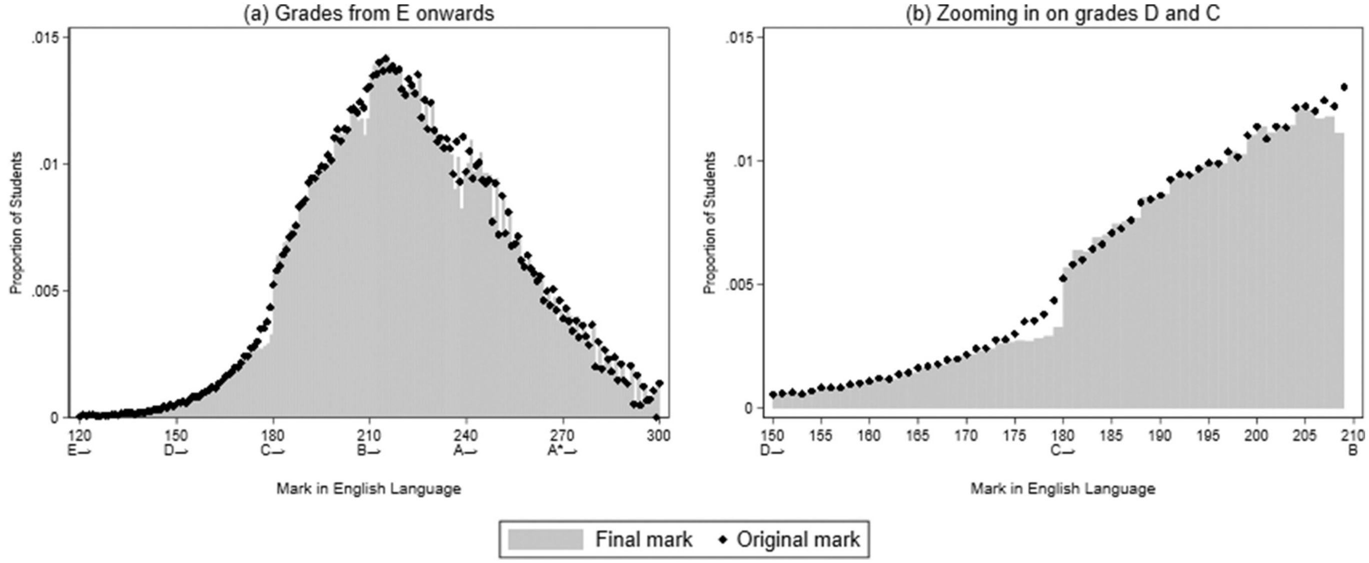

Fig. 2. Final (post-review) and original (pre-review) distribution of marks. Note. (a) Histogram showing the final (post-review) distribution of marks from grade E onwards; (b) Zooms in on the D & C area of the histogram depicted in (a). In both graphs, the dotted line shows the original (pre-review) distribution of marks. Both distributions use data for Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on the data sample construction.

fully parametric set-up (under assumptions of linearity). More formally, the following equations can be estimated in a two stage least squares setting:

$$Second\ stage : Y_{is} = \beta_0 + \beta_1 CF_{is} + \beta_2 M_{is} + \beta_3 CO_{is} * M_{is} + \beta_4 X_{is} + \mu_s + \epsilon_{is} \quad (1)$$

$$First\ stage : CF_{is} = \alpha_0 + \alpha_1 CO_{is} + \alpha_2 M_{is} + \alpha_3 CO_{is} * M_{is} + \alpha_4 X_{is} + \mu_s + \omega_{is} \quad (2)$$

where outcome Y of individual i in school s (the school where the individual completed Key Stage 4) is related to a dummy variable indicating whether he/she achieves a C grade in the English language GSCE exam (after the review process, denoted CF). Marks of the student are denoted by M (these are the original distribution of marks, i.e. pre-review) and CO is a dummy variable indicating if the student originally was awarded a C grade (before any remarking). X is a set of pre-determined characteristics that we are using throughout the analysis, although their inclusion or exclusion makes no difference to estimated effects. Specifically, we include the student's ethnicity, gender, whether he/she is eligible to receive free school meals, whether he/she speaks English as a first language and the test score obtained in the examinations at the end of primary school. μ denotes a school fixed effect. Our main results introduce the forcing variable in a linear way, but we show that results barely change when using a quadratic functional form. ϵ_{is} and ω_{is} are error terms and we use robust standard errors, following Kolesár and Rothe (2018).[9]

Table 2
Determinants of asking for a review and getting an upgrade.

Dependent variable	(1)	(2)	(3)
	Any review	Any review	Grade up after reviews
White	−0.001	−0.003	−0.002
	(0.007)	(0.004)	(0.016)
Free School Meals	−0.006	−0.006	−0.010
	(0.007)	(0.004)	(0.017)
English Language	−0.002	−0.003	0.007
	(0.007)	(0.005)	(0.020)
Female	−0.007*	−0.004	0.002
	(0.004)	(0.003)	(0.010)
KS2 total points (std)	0.036***	0.013***	0.028*
	(0.006)	(0.003)	(0.014)
Original marks	−0.004***	−0.003***	−0.004***
	(0.000)	(0.000)	(0.000)
Mean dependent variable	0.101	0.101	0.122
Sample size	49,231	49,231	4966
Sample	All higher tier (C&D)	All higher tier (C&D)	Students involved in any kind of review (C&D)
School fixed effects	No	Yes	Yes

Notes: The dependent variables in all regressions are dummy variables. In the first 2 columns, the dependent variable is equal to 1 if any of the units contributing to the final mark was subject to any kind of review (units subject to review are units 1 and 3). The dependent variable in Column 3 is equal to 1 if the grade goes from D to C after the review process. Standard errors are clustered at the KS4 school level (i.e., school the student was attending in Year 11), with *p < 0.10; **p < 0.05; ***p < 0.01. Columns 2 and 3 include KS4 school fixed effects. Marginal effects coming from probit estimates are almost identical to the coefficients shown in Column 1 in this table.

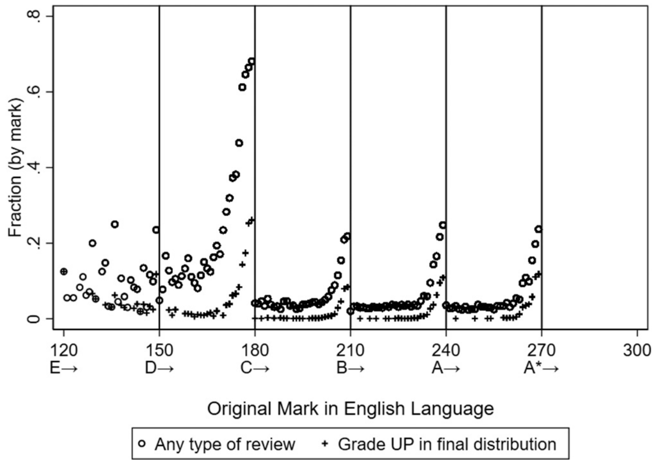

Fig. 3. Proportion of students being reviewed and being upgraded, by original mark. Note. Graph showing the fraction of students (within each original mark), being subject to any kind of review and being upgraded; for Higher Tier students. See Appendix A for further details on review data.

[9] When we use the whole C-D range for estimations (where we have a reasonably big number of clusters as given by the forcing variable – i.e. marks are grouped into 60 clusters), clustering standard errors at the level of the forcing variable does not make much difference to the standard errors. Also, the results do not change in any substantive way when we cluster standard errors at the school level.

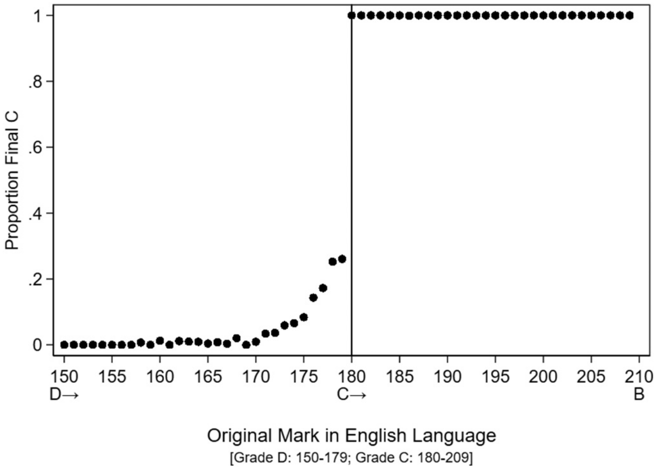

Fig. 4. First stage: proportion final C grade by original marks. Note. Graph showing the first stage. Each dot represents the fraction of students obtaining a grade C (post-review) within each potential original mark (pre-review); for Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on the data sample construction.

We estimate parametric regressions using the full range of scores between grades C and D, and zooming in to ±10 points from the original C threshold (since it is from −10 points away from the left of the original C threshold when the probability of getting a final C grade starts becoming strictly positive – see Fig. 4). We then estimate linear parametric regressions over a small range of the data close to the C threshold (original marks ranging from ±5 to ±1 points away from the original C threshold), as well as local polynomial (fuzzy) regression-discontinuity point estimators with robust-bias corrected confidence intervals (Calonico et al., 2014).

For this approach to estimate the true causal relationship between obtaining a grade C and individual outcomes, passing the threshold must be quasi-randomly assigned. The validity of this assumption is examined in detail below.

3.2. Validity

As discussed previously, the examination process is sufficiently rigorous to ensure that teachers and examiners are not able to manipulate students close to the C threshold in the original mark distribution. If this is the case, then we should observe that predetermined variables vary smoothly across the threshold corresponding to a C grade in the original distribution (i.e. *CO* in the notation of Eq. (2) above). Prior performance at Key Stage 2 is measured using results from a national test that takes place at the end of primary school (at age 11).

Fig. 5 plots the relationship between prior student performance at age 11 (Key Stage 2) and the original (pre-appeal) distribution of marks. The graph on the left covers the entire C and D range, whereas the graph on the right zooms in at ±10 points away from the C cut-off. Linear regression lines are fitted separately on each side of the C threshold. The discontinuity and standard error shown correspond to Intention-to-Treat estimates coming from a sharp regression discontinuity design with the original marks as the forcing variable, letting the slope change on each side of the original C threshold and without any controls. There is no visual discontinuity around the Grade C threshold in GCSE English. The same is true for the other baseline characteristics considered here: the student's ethnicity, gender, whether he/she is eligible to receive free school meals and whether he/she speaks English as a first language (see Fig. B1 in the online Appendix).

In Table B2 of the Online Appendix, we report regression estimates where each baseline characteristic is regressed against a dummy variable measuring whether the student obtains a C grade (pre-review), controlling for the original (pre-review) mark, letting the slope change on each side of the original C threshold, and with and without including school fixed effects. In almost all cases, the relationship between the baseline characteristic and whether or not the student obtains a C grade is small in magnitude and does not reach statistical significance (this is even more so close to the C threshold, see Table B3 in the Online Appendix for checks done using the ±5 and ±1 bandwidth). Hence, it is plausible to conclude that the marginal student who passes the (pre-review) threshold appears to be quasi-randomly assigned.

4. Results

4.1. Outcomes

We consider the following outcomes: (1) the probability of dropping out of education at the age of 18; (2) the probability of not being observed in education, employment or training (NEET) at the age of 18; (3) entering an upper secondary academic or vocational qualification by the age of 19, which is the age by which most English students will have entered upper secondary education if they are going to start at all (i.e. a 'level 3' qualification which is A-levels or other vocational qualifications); (4) the probability of achieving a full level 3 qualification by the age of 19 (i.e. the typical requirement for a university entrant); (5) the probability of enrolling in tertiary education by age 19 (i.e. undergraduate or foundation university degree or high-level vocational education). Although this cohort is too young to observe labour market earnings, having a level 3 qualification is associated with a high wage premium, even if young people do not subsequently go on to tertiary education (e.g. McIntosh, 2006; Patrignani et al., 2017).

Appendix A explains how we have constructed these outcome variables. Table 3 shows summary statistics of outcome variables for the whole cohort sitting GCSE English Language (column 1), the AQA English language sample (column 2), the subsample of higher tier students in the English Language sample with marks in the C-D range that are main interest here (column 3) and the subsample of foundation students in the English Language sample with marks in the C-D range (column 4). The patterns described when discussing predetermined characteristics for the same groups in Table 1 also emerge here: AQA English Language students (column 2) have slightly better outcomes than the average student in the cohort (column 1), and higher tier students perform much better than foundation students on any of the five dimensions analysed here.

Before showing the regression results, the outcome variables are plotted in Figs. 6 and 7 according to whether or not students obtain a C grade in the original distribution of marks (i.e. *CO* in the notation of Eq. (2) above). These plots are therefore a depiction of an 'intention to treat' type of analysis with graphical evidence of the reduced form impact. The graphs are for all students who obtained marks (pre-review) within the range of a D and a C grade (i.e. marks between 150 and 209), where the threshold is at 180 marks (see Appendix A for more details on the sample construction). These show that the discontinuity around the C grade corresponds to a decrease in the probability of not dropping out of education at age 18 (Fig. 6a and b) as well as a lower probability of being observed as 'not in education, employment or training' (NEET) at age 18 (Fig. 6c and d). Fig. 7 shows that students who just pass the original C cut-off have a higher probability of accessing (Fig. 7a and b) or achieving (Fig. 7c and d) upper secondary education by age 19, and starting tertiary education by age 19 (Fig. 7e and f).[10] This gives

[10] Dispersion below the C threshold (i.e. among D students) is always larger than above the threshold (i.e. for C students), because there are more observations above the threshold (84% of the observations are above the original, pre-remark C threshold).

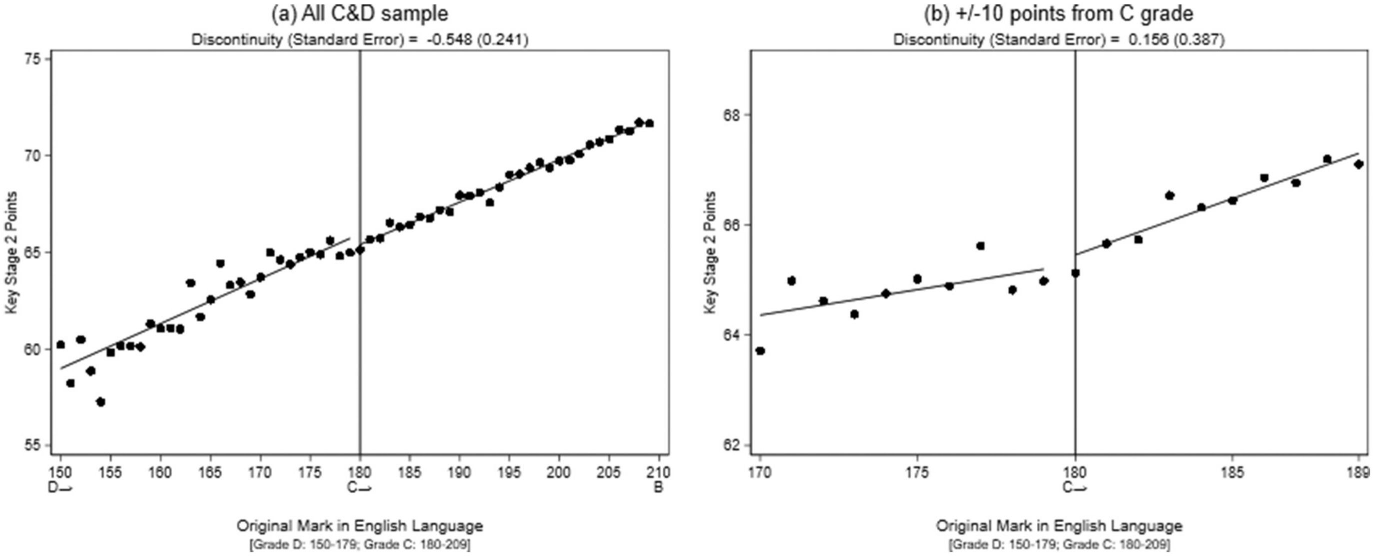

Fig. 5. Key Stage 2 points by forcing variable. Note. Graph showing the relationship between prior student performance at Key Stage 2 National exams (age 11) and the original (pre-review) marks. Each dot represents the average score obtained in the Key Stage 2 examinations within each potential original mark (pre-review). Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on the sample construction. Linear regression lines are fitted separately on each side of the C threshold. The discontinuity and standard error shown correspond to Intention-to-Treat estimates coming from a sharp regression discontinuity design with the original marks as the forcing variable, letting the slope change on each side of the original C threshold and without any controls.

prima facie evidence of the effects of narrowly passing the threshold. This is not evident across other grade thresholds (i.e. C/B, B/A, A/A*) for any of these outcomes or indeed at other points of the distribution. Fig. 8a, b, and c show this for the outcome measuring achievement of upper secondary education. Graphs for the remaining outcomes do not show any discontinuities either.

4.2. Baseline results

In Table 4 we show regressions estimated for two different specifications for the full sample of interest (columns 1–2, without and including KS4 school fixed effects, respectively) and for the subsample within ±10 points of the grade C threshold (columns 3–4). There are five panels for the different outcome variables (panels A to E). Each coefficient shows the estimated effect of achieving a grade C (after any review) on the outcome of interest. In the notation of Eq. (1), these correspond to the coefficient β_1, the (second stage) instrumental variable estimate. The sixth panel (panel F) shows estimated coefficients for the first stage (i.e. α_1 in Eq. (2)), which is always very large and statistically significant.

Results are very similar across the different specifications (whether they include school fixed effects or not, and whether they consider the whole C-D range or the sample within ±10 points from the original C threshold) and are statistically significant (apart from one of the specifications where commencing tertiary education is the dependent variable).

Overall, the magnitude of the results is slightly bigger in the ±10 sample, but with all the regressions suggesting a sizeable effect of marginally achieving (or failing to achieve) a C grade. In the sample of students originally obtaining a grade D, about 16% of students have dropped out of any form of education by the age of 18 (14% of students within −10 marks of the Grade C threshold). The effect of just achieving a C grade in GCSE English is to reduce the probability by almost 4 percentage points, with a slightly higher point estimate for the smaller subsample of students.

A smaller number of students in this subsample are classified as 'not in education, employment or training' (NEET) at age 18. This is 6% of the sample of students with grade D, and about 5% of students within −10 marks of the original C threshold. The regression estimates suggest that just achieving a C grade can have a big effect relative to this sample average. It reduces the probability by about 2 percentage points, rising to almost 3 percentage points in the smaller sub-sample.

Table 3
Descriptive statistics: outcomes.

	(1)	(2)	(3)	(4)
	2013 cohort sitting English Language GCSE	AQA English Language sample	AQA English Language C&D sample - Higher Tier	AQA English Language C&D sample - Foundation Tier
Not observed in education at age 18	8.3	7.9	9.2	14.2
Not observed in education, employment or training (NEET) at age 18	3.3	3.0	3.2	5.3
Enrolled in a level 3 qualification by age 19	87.7	89.0	90.0	75.9
Achieved a full level 3 qualification by age 19	75.3	77.4	73.2	56.7
Enrolled in any level 4+ qualification by age 19	36.2	38.6	26.9	16.6
Number of pupils	383,730	189,485	49,231	33,034

Note. Figures are in %. 2013 cohort: those in the KS4 Candidate/Indicator tables that belong to year group 11 (derived from birth date) and appear in the Census data (i.e. we have data on pre-determined characteristics). Students sitting English Language GCSE in the 2013 cohort are those students that are observed in the 2013 KS4 Results tables as having sat a full GCSE qualification in English Language with any of the awarding bodies. More details about the sample and data construction are given in Appendix A.

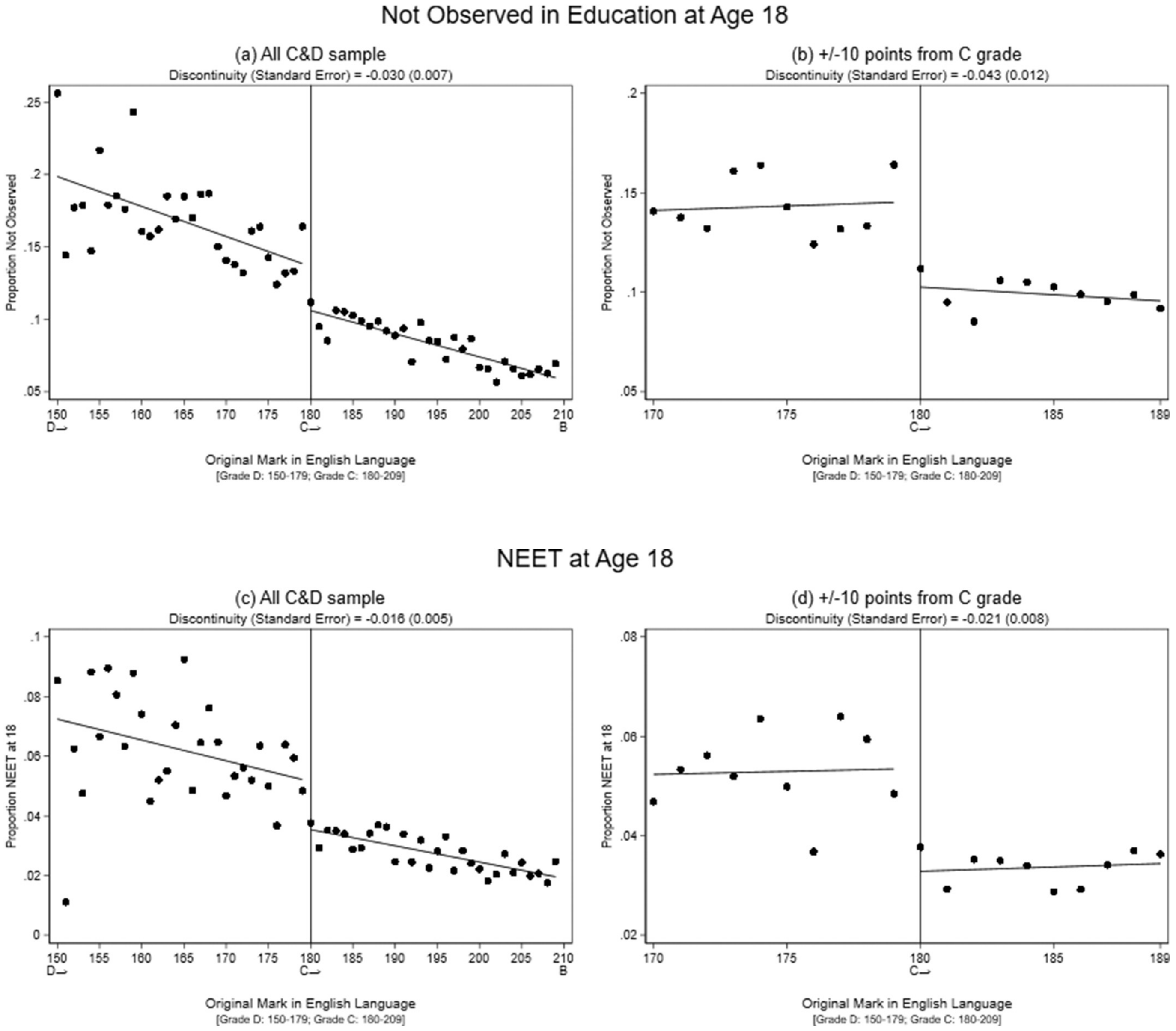

Fig. 6. Outcomes at age 18 by forcing variable. Note. Graph showing the relationship between outcomes at age 18 and the original (pre-review) marks. Each dot represents the proportion of students classified as Not Observed in Education (figures (a) and (b)) and NEET (figures (c) and (d)) at age 18, within each original mark (pre-review). Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on the sample construction. Linear regression lines are fitted separately on each side of the C threshold. The discontinuity and standard error shown correspond to Intention-to-Treat estimates coming from a sharp regression discontinuity design with the original marks as the forcing variable, letting the slope change on each side of the original C threshold and without any controls.

With regard to starting an upper secondary academic or vocational level qualification within 3 years of taking the GCSE examinations, the effect of marginally achieving a grade C is to increase this probability by between 6.4 and almost 9 percentage points. This is a big effect. About 75% of people originally scoring a D grade manage to start a high-level qualification within this time and thus it is not a very high yard-stick of achievement. Yet, just failing to get a C grade manifestly has a huge effect on the probability of getting back on track within 3 years. The next panel shows very similar effects on whether a student is able to achieve a 'full-level' 3 qualification within 3 years (whereas the expectation would be that most people would achieve this within 2 years of the end of compulsory education). As a robustness check, we obtain very similar results if we exclude observations that are very close to the C threshold (following Barreca et al., 2011; see results in Table B4 in the online Appendix).

Panel E shows that just managing to obtain a grade C affects the probability of enrolling in tertiary education. Marginally achieving a C grade increases the probability of commencing tertiary education by 2.5 to 4 percentage points in a context where about 13% of students originally scoring a D grade have started tertiary education by this age (16% percent for those below 10 marks of the C threshold).

The effects discussed so far rely on a partially fuzzy RD design, equivalent to following an instrumental variable strategy. As such, they capture the effects of marginally achieving a C grade on compliers. One might wonder whether the effects for this subpopulation are of interest. It turns out that these local average treatment effects are very similar to the average treatment effects that would be obtained from implementing a sharp RD design using the post-review marks as the forcing variable and controlling for school fixed effects. This is not surprising because most of the variation in who applies for an upgrade is driven by the school attended. That these results turn out to be very similar can already be gleaned from observing two sets of results shown here: (1) the strong magnitude of the first stage; and (2) the fact that who applies for a review and gets upgraded depends very little on observable characteristics once school fixed effects have been taken into account in the regression

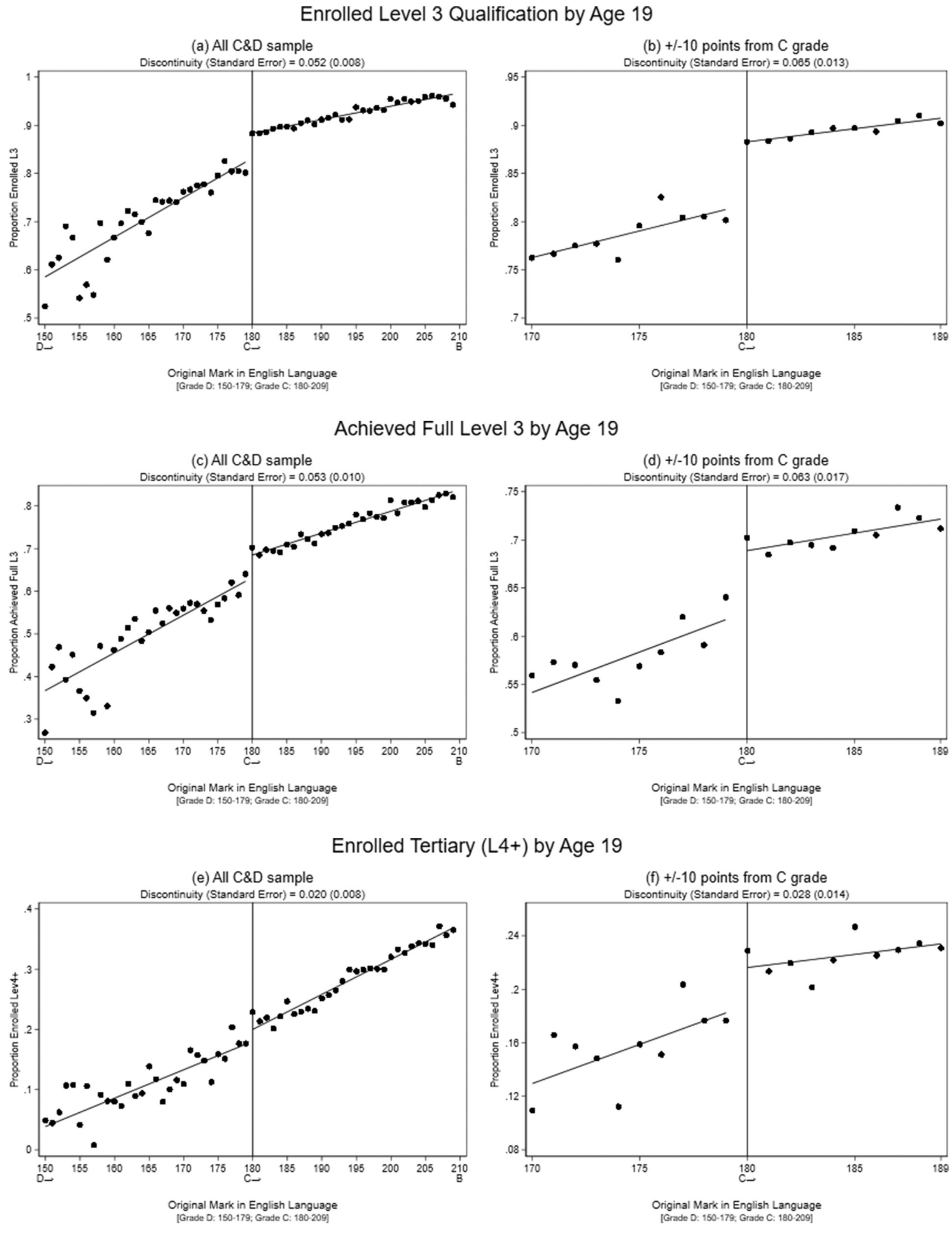

Fig. 7. Outcomes by age 19 by forcing variable. Note. Graph showing the relationship between outcomes by age 19 and the original (pre-review) marks. Each dot represents the proportion of students classified as achieving each outcome within each original mark (pre-review). See Appendix A for further details on the sample and variable construction. Linear regression lines are fitted separately on each side of the C threshold. The discontinuity and standard error shown correspond to Intention-to-Treat estimates coming from a sharp regression discontinuity design with the original marks as the forcing variable, letting the slope change on each side of the original C threshold and without any controls.

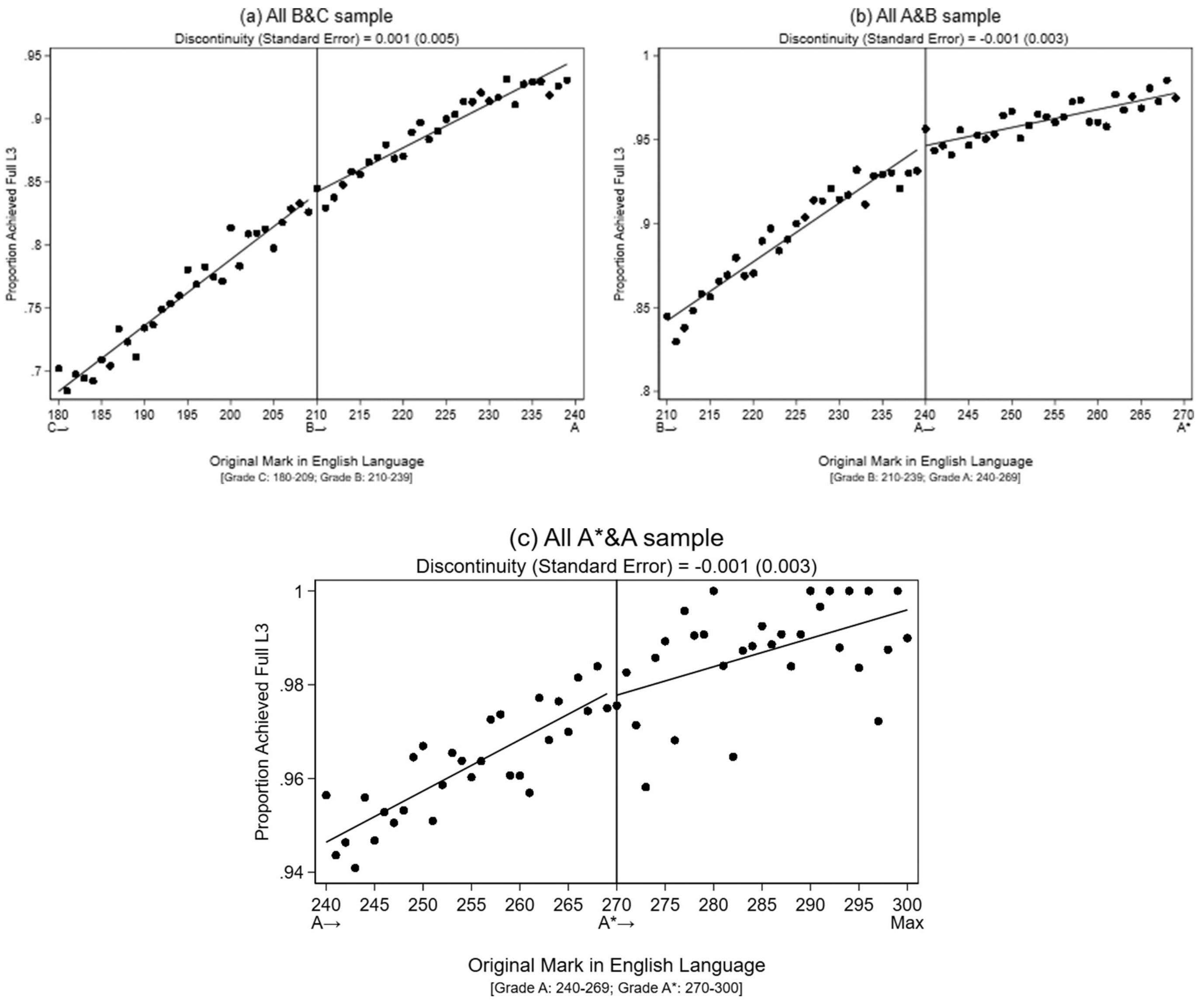

Fig. 8. Achieving a full level 3 qualification by age 19 at other grade thresholds. Note. Graph showing the relationship between achieving a full level 3 qualification (i.e., upper-secondary) by age 19 and the original (pre-review) marks. Each dot represents the proportion of students classified as achieving the outcome within each original mark (pre-review). Higher Tier students (i.e., those sitting the Higher Tier paper in Unit 1). See the Data appendix for further details on the sample construction. Linear regression lines are fitted separately on each side of the relevant threshold. The discontinuity and standard error shown correspond to Intention-to-Treat estimates coming from a sharp regression discontinuity design with the original marks as the forcing variable, letting the slope change on each side of the original relevant threshold and without any controls.

analysis. The reduced form effects are also very similar, as will be shown in Section 4.4.

4.3. Local regressions with varying windows

In Table 5, we show results for subsamples of students who obtain a very narrow range of marks in the original (pre-review) distribution. We use the same linear model described in Section 3.1, with the only difference that we are not including exogenous interactions between the forcing variable and the instrument in any of the equations. However, we show that results are robust to its inclusion. Again, there are five panels for the different outcome variables (the sixth showing results from first stage regressions) and five columns, each of which shows the estimated effect of achieving grade C on the outcome of interest. We saw in the previous section that results are barely affected by the inclusion of school fixed effects. In these set of regressions we do not include school fixed effects. The reason we estimate the regressions without school fixed effects is because as the sample size reduces, there are more schools with only one student in the specified mark range and hence not used for 'within school' estimates (i.e. they are dummied out by the school fixed effect). For instance, in the sample of students within ±5 marks of the C threshold, about 16% of schools have one student. This rises to half of all schools in the sample of students within ±1 mark of the threshold. Results including school fixed effects are available in Table B5 in the online Appendix.

In Table 5, column (1) shows estimates of regressions for the subsample of students within ±5 marks from the original grade C threshold. Column (2) replicates the regressions for the sample of students within ±4 marks of the threshold. Then the sample is gradually narrowed to ±3 marks (column 3), ±2 marks (column 4) and ±1 marks (column 5).

The results in Table 5 are consistent with those shown for the larger sample and are qualitatively similar. They are generally statistically significant. The variable denoting enrolment in tertiary education is never statistically significant when school fixed effects are included but point estimates are always positive and slightly higher than for the global regressions reported in Table 4. The point estimates are usually consistent across specifications with a different number of students. The outcome showing whether a student enrols in study for an upper-secondary

Table 4
Fuzzy RD estimates: impact of getting a C grade (post-review) on different outcomes.

	(1)	(2)	(3)	(4)
	Window: All C&D		Window: ±10 points	
A. Outcome variable: not observed in education at age 18				
Grade C (final)	−0.036***	−0.037***	−0.059***	−0.052***
	(0.009)	(0.009)	(0.017)	(0.017)
Mean dep variable, D group	0.158		0.143	
B. Outcome variable: NEET at age 18				
Grade C (final)	−0.019***	−0.021***	−0.028***	−0.028**
	(0.006)	(0.006)	(0.010)	(0.011)
Mean dep variable, D group	0.060		0.053	
C. Outcome variable: enrolled in any Level 3 (upper secondary) qualification by age 19				
Grade C (final)	0.064***	0.068***	0.088***	0.087***
	(0.010)	(0.010)	(0.018)	(0.018)
Mean dep variable, D group	0.745		0.791	
D. Outcome variable: achieved a full level 3 qualification by age 19				
Grade C (final)	0.064***	0.071***	0.087***	0.089***
	(0.013)	(0.013)	(0.023)	(0.024)
Mean dep variable, D group	0.538		0.585	
E. Outcome variable: enrolled in tertiary education (level 4 or above) by age 19				
Grade C (final)	0.025***	0.025**	0.040**	0.031
	(0.010)	(0.010)	(0.019)	(0.019)
Mean dep variable, D group	0.130		0.160	
F. Summary first stage: obtaining a C grade after the review process				
Grade C (original)	0.827***	0.828***	0.723***	0.726***
	(0.007)	(0.008)	(0.012)	(0.013)
Sample size	49,231	49,231	14,597	14,597
School Fixed effects	No	Yes	No	Yes

Note. Panels A to E: each cell shows the main coefficient of interest for the C dummy variable (endogenous variable in the second stage, i.e., β_1 in Eq. (1)). Panel F: each cell shows the main coefficient of interest in the first stage. All regressions control for the forcing variable in a linear way. The slope of the forcing variable is allowed to vary on each side of the C threshold in all cases. All regressions include the set of controls described in Appendix A. The window restriction is based on the forcing variable (i.e. excluding 10 points away from the C threshold as given by the pre-review distribution of marks). School fixed effects are defined at the KS4 level (i.e. the school the student was attending in Year 11). Robust standard errors, with $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 5
Fuzzy RD estimates narrowing the window: impact of getting a C grade (post-review) on different outcomes.

	(1)	(2)	(3)	(4)	(5)
	±5 points	±4 points	±3 points	±2 points	±1 points
A. Outcome variable: not observed in education at age 18					
Grade C (final)	−0.071***	−0.077***	−0.063**	−0.076*	−0.071***
	(0.022)	(0.025)	(0.030)	(0.039)	(0.025)
Mean dep variable, D group	0.140	0.140	0.144	0.150	0.164
B. Outcome variable: NEET at age 18					
Grade C (final)	−0.027**	−0.028*	−0.015	−0.001	−0.012
	(0.013)	(0.015)	(0.018)	(0.023)	(0.015)
Mean dep variable, D group	0.052	0.052	0.057	0.054	0.048
C. Outcome variable: enrolled in any level 3 (upper secondary) qualification by age 19					
Grade C (final)	0.101***	0.112***	0.108***	0.113***	0.110***
	(0.024)	(0.027)	(0.032)	(0.041)	(0.026)
Mean dep variable, D group	0.810	0.810	0.804	0.803	0.802
D. Outcome variable: achieved a full level 3 qualification by age 19					
Grade C (final)	0.091***	0.089**	0.090**	0.076	0.086**
	(0.031)	(0.035)	(0.042)	(0.053)	(0.034)
Mean dep variable, D group	0.603	0.610	0.618	0.618	0.641
E. Outcome variable: enrolled in tertiary education (level 4 or above) by age 19					
Grade C (final)	0.057**	0.072**	0.081**	0.090**	0.073***
	(0.026)	(0.029)	(0.035)	(0.045)	(0.028)
Mean dep variable, D group	0.174	0.177	0.185	0.177	0.177
F. Summary first stage: obtaining a C grade after the review process					
Grade C (original)	0.724***	0.720***	0.715***	0.735***	0.737***
	(0.014)	(0.016)	(0.019)	(0.025)	(0.017)
Sample size	7082	5671	4212	2817	1409
Number of schools	1258	1201	1110	993	742

Note. Panels A to E: each cell shows the main coefficient of interest for the C dummy variable (endogenous variable in the second stage, i.e., β_1 in Eq. (1)). Panel F: each cell shows the main coefficient of interest in the first stage. All regressions control for the forcing variable in a linear way. All regressions include the set of controls described in Appendix A. The window restriction is based on the forcing variable (i.e. excluding ±X points away from the C threshold as given by the pre-review distribution of marks). School fixed effects are not included in the regressions. Robust standard errors, with $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

academic or vocational qualification by the age of 19 is positive, significant and large in every specification. Thus, these specifications show the robustness of our findings to using fewer students (who are a priori more and more similar) to identify the causal effect of obtaining a grade C in GCSE English language.

4.4. Robustness

The robustness checks are discussed in detail in the online Appendix. In summary, the results presented above are robust to all of the specification checks. They are as follows: (1) local polynomial (fuzzy) regression-discontinuity point estimators with conventional and robust-bias corrected confidence intervals (Calonico et al., 2014); (2) A placebo test based on the following intuition: in the absence of manipulation of original marks, marginally obtaining a C grade in English Language should not have an impact on the likelihood of obtaining a C grade (or above) in GCSE Mathematics. We find that it does not. (3) Further checks of the sensitivity of results to changing the specification in various ways such as whether baseline characteristics are controlled for; whether we include an exogenous interaction between the forcing variable and the instrument; whether we introduce the forcing variable in a quadratic way in both the first and second stage. Although point estimates change slightly in some of these checks, the interpretation of results is unchanged. (4) Finally, our partially fuzzy RD framework requires the linearity assumption for estimation purposes (see Battistin and Rettore, 2008). We can nonetheless estimate the reduced form equations in a non-linear setting and assess whether results point towards the same conclusions. As expected, given the size of the first stage coefficients, the reduced form estimates are slightly smaller but in line with the estimates obtained for reduced form estimates in a linear setting. Overall, the evidence suggests that our results satisfy the assumptions for partially fuzzy RD estimation and that our results are not driven by a specific choice of bandwidth, inclusion of controls, or the functional form of the forcing variable.

5. Mechanisms and implications

It is clear that failing to obtain a grade C in GCSE English can have serious consequences for students. Students' grades in high stakes exams may affect their incentives to invest further in their education (Hvidman and Sievertsen, 2019). In this context, one possible reason is that students are held back by the psychological effect that perceived failure can have on self-evaluation of abilities (as discussed by Papay et al., 2015). However, it is not a universal finding that failing to achieve significant thresholds in exams has negative consequences. For example, in their paper about test-based accountability in Massachusetts, Papay et al. (2015) only found effects for a specific sub-group with regard to maths (and nothing for English). Clark and Martorell (2014) found no wage penalty attributable to barely failing to obtain a high school diploma in the US. Our data do not enable us to directly consider the potential importance of psychological effects as a mechanism – except to note that to the extent they exist, they are only evident for those students who fail to achieve a Grade C and not at other grade thresholds (as shown in Fig. 8 and discussed in Section 4.1 where we show that there is no evidence of a discontinuity for any outcome variable at other grade thresholds). Yet one might expect some psychological effect here too as it is likely that students who just fail to achieve thresholds for B or A* are also disappointed and aware that this information will be on record for applications to universities or to employers in the future.

A plausible potential explanation for the large consequences of just failing to obtain a grade C is that the range of post-16 opportunities narrow without this educational credential or signal. The grade C in English is important as a credential in itself (as a core subject) and has implications for another oft-used signal of educational performance: whether a student has at least 5 'good' grades in GCSE (i.e. A*-C). Both the number of 'good' GCSEs and the grade in specific GCSEs can affect the post-16 educational institution that the student is able to attend as well as the course he/she can choose.[11]

In Table 6, we show regressions with the following outcome variables: whether the student obtains 5 or more GCSEs at grades A*-C; whether he/she stays at the same school at age 17; whether he/she attends an academic institution at age 17; and whether he/she enrols in qualifications that are pre-requisites for university entry at age 17 (i.e. A-levels; AS-levels; Applied Generals). We show regressions with and without school fixed effects for three samples: all those obtaining a grade C or D in English (columns 1 and 2); all those within plus or minus 10 points of the grade C threshold (columns 3 and 4); and all those within plus or minus 5 points of the grade C threshold (columns 5 and 6). The approach is analogous to that shown for Table 4. Specifically, we report estimates from the fuzzy regression discontinuity design of the effect of getting a grade C on various intermediary outcomes.

For each outcome variable, a consistent story is shown across all six specifications. Panel A shows that getting a grade C in English makes it more likely that a student will obtain 5 or more 'good' GCSEs by about 10 percentage points (from a baseline of close to 72% for those students originally scoring a D grade). Thus, it can make the difference between achieving and failing to achieve another signal of performance at age 16. A marginal student may face the double whammy of failing to obtain a 'good' grade in a core subject and failing to achieve a sufficient number of 'good' GCSEs.

One door that might close to students is the possibility of staying on at the same school they attended up to age 18. If schools cater for 16–18 year olds, this is usually only in academic subjects (such as A-levels) and are likely to have selection criteria based on performance in GCSEs – where English is particularly important as a core subject. Panel B of Table 6 shows that students without a grade C in English are indeed less likely to stay on at the same school. In the sample of all students with grade C-D, the magnitude is 4–5 percentage points from a baseline of 15% for those students originally scoring a D grade. The magnitude is little changed in the narrower windows (columns 3–6). It might seem surprising that schools do not systematically make exceptions for these marginal candidates – and certainly implies some rigidity on their part. The school might be concerned about their (publically available) performance table ranking for A-level courses (2 years later) and worry that marginal candidates at GCSE could underperform in A-level courses (even though this is just as likely for those who marginally pass grade C at GCSE).

A bigger door that might close is whether students can attend an academic institution at all. In the sample of students with an original grade D in English, just under 30% attend a school or a sixth form college at age 17. The latter are small institutions that cater for students of age 16–18 and focus on academic subjects. In panel C, we consider how obtaining a grade C in English affects the probability of attending an academic institution (i.e. a school or sixth form college). Obtaining a grade C in English increases this probability by about 4–5 percentage points in the bigger sample of all C-D students. The point estimate is 3–4 percentage points and 5–7 percentage points in the smaller windows (namely those within plus or minus 10 points; and those within plus or minus 5 points, respectively).

Table 6
Potential mechanisms.

	(1)	(2)	(3)	(4)	(5)	(6)
	Window: All C&D		Window: ±10 points		Window: ±5 points	
A. Outcome variable: getting 5 or more GCSEs (and equivalents) at grades A*-C						
Grade C (final)	0.106^{***}	0.102^{***}	0.095^{***}	0.089^{***}	0.094^{***}	0.103^{***}
	(0.010)	(0.010)	(0.018)	(0.018)	(0.027)	(0.028)
Mean dep variable, D group	0.722		0.783		0.801	
B. Outcome variable: staying in same school at age 17						
Grade C (final)	0.042^{***}	0.049^{***}	0.030	0.039^{**}	0.045	0.052^{*}
	(0.010)	(0.010)	(0.020)	(0.018)	(0.031)	(0.031)
Mean dep variable, D group	0.155		0.181		0.196	
C. Outcome variable: attending an academic institution at age 17						
Grade C (final)	0.039^{***}	0.049^{***}	0.043^{*}	0.034	0.070^{**}	0.055
	(0.012)	(0.012)	(0.023)	(0.023)	(0.035)	(0.037)
Mean dep variable, D group	0.285		0.320		0.339	
D. Outcome variable: enrolled in any A/AS/applied GCE at age 17						
Grade C (final)	0.103^{***}	0.114^{***}	0.106^{***}	0.111^{***}	0.151^{***}	0.150^{***}
	(0.012)	(0.012)	(0.023)	(0.023)	(0.035)	(0.037)
Mean dep variable, D group	0.239		0.287		0.310	
Sample size	49,231	49,231	14,597	14,597	7082	7082
School Fixed effects	No	Yes	No	Yes	No	Yes

Note: The outcome variable in Panel A is computed using the variable ks4_level2 from the KS4 Candidate Indicator dataset. The outcome variable in Panel B is equal to 1 if the institution attended at age 17 has the same school identifier (Unique Reference Number: URN) as the institution attended at age 16. The outcome variable in Panel C is equal to 1 if the student is attending a school or sixth form college and 0 otherwise. The outcome variable in Panel D is equal to 1 if the student is observed enrolled in any A/AS/Applied General Certificate of Education (GCE) qualification at age 17. For each specification, we show: (1) first row: the main coefficient of interest for the C dummy variable (endogenous variable in the second stage, i.e., β_1 in Eq. (1)); (2) second row: associated standard error; (3) third row: mean dependent variable. All regressions control for the forcing variable in a linear way. The slope of the forcing variable is allowed to vary on each side of the C threshold in all cases. All regressions include the set of controls described in Appendix A. The window restriction is based on the forcing variable (i.e. excluding ±X points away from the C threshold as given by the pre-review distribution of marks). Robust standard errors, with $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. School fixed effects are defined at the KS4 level (i.e. the school the student was attending in Year 11).

Students who fail to get a grade C in English might find it difficult to enrol for an academic qualification. This is not only because of difficulty in accessing academic institutions but also because of pre-requisites for some academic courses. It should be noted that many students move to another educational institution for their post-secondary education (such as a Further Education College) and these institutions do not necessarily know whether the student is marginal or not (i.e. they do not have their exact mark). In panel D, we analyse the effect of obtaining a grade C in English on the probability of being enrolled in a broadly-defined academic qualification at age 17 (specifically A-levels, AS-levels or Applied General qualifications). About 24% of all students with a D in English are enrolled in such a qualification at age 17. Marginally succeeding to make grade C increases the probability by 10–11 percentage points in the sample of C-D students. The point estimate is either the same or higher in the smaller windows.

From the above analysis, we see that more doors are closed to students who do not get a Grade C in English at age 16 (i.e. in the academic year after failure). Another mechanism leading to poor later outcomes may be the quality of the environment to which they are exposed in the receiving institution. We measure this by 'peer quality' within the institution that the student attends at age 17. Of course, we can only measure peer quality in this way for those students who have not

[11] https://university.which.co.uk/advice/gcse-choices-university/how-important-are-my-gcse-grades.

dropped out of education at age 17, so caution needs to be used when interpreting these results since there may be some degree of sample selection bias. However, given that we only observe those students that were not induced to abandon education at age 17 having failed to get the C grade the first time around, the consequence would be downward bias in the estimates. In this case, the effects that we discuss are potentially lower bounds.

We construct the following measures of peer quality, using well known indicators of performance at age 16: the fraction of peers achieving 5 or more grades at A*-C including English and maths; the fraction achieving a C grade or more in English; and the fraction achieving a C grade or more in maths. We show regressions in panels A-C of Table 7 where each of these measures of peer quality is the dependent variable. We use the same structure as in Table 6, first presenting fuzzy regression discontinuity estimates for the full sample of C-D students (columns 1 and 2), before considering those within 10 points either side of the threshold (columns 3 and 4) and within 5 points (columns 5 and 6). The estimates are qualitatively similar across all the proxies of peer quality and within the different windows. In summary, those who get a grade C (and have not dropped out of education by age 17) are more likely to attend an institution with 'good peers' (according to any of the proxies) by 2–3 percentage points.

From this analysis, we can see that if a student marginally fails to obtain a grade C in English at GCSE, various doors are shut to them the following year (and many marginal students do not recover from this). They may not be able to access particular institutions or courses and end up in institutions with lower quality peers. The consequence is that they face a relatively high probability of dropping out of education at age 18 or even being 'not in education, employment or training'. They are less likely to enter an upper secondary (i.e. level 3) qualification up to three years later and less likely to enrol in tertiary education.

Table 7
Quality of peers in the receiving institution at age 17.

	(1)	(2)	(3)	(4)	(5)	(6)
	Window: All C&D		Window: ±10 points		Window: ±5 points	
Fraction of peers in receiving institution (URN) at age 17 that...						
A. Achieved five or more GCSEs at grades A*-C incl. English and maths						
Grade C (final)	3.012***	3.287***	2.718**	2.425*	3.530*	2.702
	(0.721)	(0.752)	(1.379)	(1.396)	(2.129)	(2.279)
Mean dep variable, D group	51.922		54.698		56.025	
B. Achieved a C grade in GCSE English						
Grade C (final)	2.874***	3.001***	2.700**	2.470**	3.253*	2.353
	(0.640)	(0.666)	(1.221)	(1.241)	(1.885)	(2.015)
Mean dep variable, D group	60.603		63.288		64.398	
C. Achieved a C grade in GCSE maths						
Grade C (final)	2.141***	2.691***	2.053*	2.144*	2.864*	2.323
	(0.586)	(0.606)	(1.117)	(1.126)	(1.721)	(1.846)
Mean dep variable, D group	65.253		67.419		68.454	
Sample size	45,526		13,187		6350	
School fixed effects	No	Yes	No	Yes	No	Yes

Note: The outcome variable is given by the percent of peers (belonging to the same KS4 cohort) in the receiving institution at age 17 that: achieved 5 or more GCSEs or equivalents at grades A*-C including English and maths (Panel A); achieved a grade C in GCSE English (Panel B); achieved a grade C in GCSE Maths. Each cell shows the main coefficient of interest for the C dummy variable (endogenous variable in the second stage, i.e., β_1 in Eq. (1)). All regressions control for the forcing variable in a linear way. The slope of the forcing variable is allowed to vary on each side of the C threshold in all cases. All regressions include the set of controls described in the Data Appendix. The window restriction is based on the forcing variable (i.e. excluding 10 (or 5) points away from the C threshold as given by the pre-review distribution of marks). Robust standard errors, with *$p < 0.10$; **$p < 0.05$; ***$p < 0.01$. School fixed effects are defined at the KS4 level (i.e. the school the student attended in Year 11).

6. Concluding remarks

This study considers an important high-stakes national examination to identify the effect of narrowly passing (or failing to pass) a critical grade threshold. In England, achieving a grade C in English (in the GCSE exam) is widely considered to be important for a variety of reasons including the fact that is often used as a pre-requisite for accessing upper secondary courses and certain institutions, and is a component of indicators published in the School Performance Tables (where performance in English and maths is specifically highlighted).

Up to now the importance of obtaining a grade C in English has not been empirically evaluated. The results reported in this paper show that students of approximately the same ability can have very different educational trajectories depending on whether they just pass the critical threshold or just fall short of it. An important mechanism for explaining this is the way that this threshold is used as a signalling device within the education system. Just failing to obtain a grade C significantly narrows the range of opportunities open to students immediately afterwards in terms of the courses, institutions and quality of institution they can attend. We show that many marginal students do not recover from this, even if students can retake the exams leading to this qualification in the following years.

This matters for a number of reasons. Firstly, one might expect someone who just misses a C grade to get back on track fairly easily and enter an upper-secondary higher-level course (at most) three years later. This does not happen for a significant minority of people. The results show that narrowly missing the C grade in English language decreases the probability of enrolling in an upper secondary qualification by at least 9 percentage points. There is a similarly large effect on the probability of achieving a higher ('full level 3') academic or vocational qualification by age 19 – which is needed as a pre-requisite for university or getting a job with good wage prospects. There is also an effect on the probability of entering tertiary education. Perhaps most surprisingly, narrowly missing a grade C increases the probability of dropping out of education at age 18 by about 4 percentage points (in a context where the national average is 12%) and becoming 'not in education, employment or training' by about 2 percentage points. Those entering employment at this age (and without a grade C in English), are unlikely to be in jobs with good progression possibilities. If they are 'not in education, employment or training', this puts them at a high risk of wage scarring effects and crime participation resulting from youth unemployment in the longer term (Gregg and Tominey, 2005; Bell et al., 2018).

More generally, this analysis does not mean that having pass/fail thresholds are undesirable. Achievement of a minimum level of literacy and numeracy in the population is an important social and economic objective. Moreover, thresholds can incentivise students to work towards achieving them. Although many countries do have exams at the end of compulsory education, arguably the GCSE exam has higher stakes both for students and for schools because of the strong accountability system of the education system in which it functions. The history is that most young people used to leave school at 16, and some measure of attainment at that point made sense. Nowadays few leave at 16 and all are supposed to be in some form of education until they are 18. GCSEs have become just one more sorting mechanism. We show that there are big consequences from narrowly missing out on a C grade in English language for outcomes that are, at least in principle, achievable for most people.

Is there a trade-off between having a national standard and providing for the needs of those who fall short of it? One cannot simply abandon standards because of the need to incentivise student and teacher effort, though it can be difficult to discern what the optimal standard is for incentivising effort and performance (as shown by Betts, 1998 and Costrell, 1994). However, one might argue that a good counterfactual is needed for those who fail to make the threshold if they are very young (as in the English case) and still need to pursue upper secondary education (which is expected for everyone in England and funded by the tax payer). Vocational education in England at lower levels is complex and has no clear future trajectory (Hupkau et al., 2017). This may explain why it matters so much in England to have doors shut on account of failing to make the threshold. Furthermore, and of particular relevance to the marginal learner, the institution to which the individual enters for their upper secondary education does not necessarily know the individual's exact marks in their GCSE exam. In fact, Higton et al. (2017) survey providers of upper secondary education about effective practice and they report that detailed information on actual exam marks would be very useful but is not readily available from a central source when it is needed earlier in the year. This means that providers are not able to target students effectively for any special support. This is an example of where perfect information supersedes a binary credential (as in Costrell's (1994) model of educational standards). The finding from this study of more general relevance is that there are risks attached to putting too much weight on passing a threshold. They may be mitigated by offering transparency in providing full details of marks to all relevant stakeholders and by ensuring that the counterfactual for students who fail to meet the threshold is structured and resourced adequately.

Declaration of competing interest

None.

Acknowledgements

We are grateful to the Department for Education, UK Government and the Economic and Social Research Council for funding this research. We thank the editor and two anonymous referees for very useful comments. We are grateful to the Joint Council for Qualifications, and in particular, to Stuart Cadwallader (formerly at AQA), John Croker, Simon Eason, Ben Jones, and Alex Scharaschkin at AQA, for giving us access to some of the data used in this paper and to the NPD team at the Department for Education. We thank Michelle Meadows and Beth Black at Ofqual for useful conversations about exam grading and appeals. We thank seminar participants at the Université Catholique de Louvain, University of Warwick, University of Essex, the London School of Economics, the Institute for Evaluation of Labour Market and Education Policy (IFAU), CUNEF and the National Institute of Economic and Social Research. We would also like to thank participants at the Royal Economic Society Conference (Bristol), the Meetings of the Society of Labor Economics (Raleigh), the RIDGE Workshop in Public Economics (Montevideo), the 2017 conference of the Centre for Economic Performance (London), the Workshop on Human Capital and Public Policy (Stockholm), the CESifo Area Conference on Economics of Education (Munich, especial thanks to Guido Schwerdt), the EALE conference (St Gallen), the 2nd IZA Workshop on the Economics of Education (Bonn), the 17th IZA/SOLE Transatlantic Meeting of Labour Economists, and the Conference of the European Economic Association (Cologne).

Appendix A. Data Appendix

A.1. Key Stage 4 Results, AQA data and sample construction

We use the National Pupil Database (NPD) to build our sample. This is a census of all students attending state schools in England. We use information for the whole cohort of students that completed compulsory schooling (at age 16) in 2012/13. The English education system is organised around various 'Key Stages'. At age 16 students complete Key Stage 4 (KS4) which ends with GCSE exams (General Certificate of Secondary Education). The *KS4 results* files (files with information at the subject level) provide information on the grade obtained by students. Table A1 shows the number of General Certificate of Secondary Education (GCSE) Full Course entries in English Language and English over the summer season (June 2013), distributed by awarding organisation.[12]

Both GCSE English and GCSE English Language count towards the school performance indicators for GCSE English that is published in the school performance tables. Students can choose between English and English language (which is normally taken together with GCSE English literature). The former course is normally taken by those students who want to explore a range of literature and language topics but do not want to take separate GCSEs in English Language and English Literature. As explained in the main text, to ensure we are considering only those students taking the same assessment, we focus on the form of English exam that is undertaken by the majority of students (i.e. English Language GCSE entries account for 72% of all GCSE English and English Language entries, see Table A1). We obtain very similar results for students who undertake English rather than English language.

The biggest awarding body for both GCSE qualifications is The Assessment and Qualifications Alliance (AQA). Over 60 and 55 percent of entries are taken with this awarding body for GCSE English Language and GCSE English, respectively.

Table A1
Number of GCSE Full Course entries by Awarding Body (KS4 Results tables, 2014)

	(1)	(2)	(3)	(4)
	English Language		English	
	Frequency	Percent	Frequency	Percent
AQA	241539	61.6	84742	55.7
WJEC	83219	21.2	39650	26.1
Pearson	37194	9.5	18815	12.4
OCR	30061	7.7	8818	5.8
Total	392015		152025	

Note. Number of GCSE Full Course entries in the summer season of the academic year 2012-2013. AQA (The Assessment and Qualifications Alliance); WJEC (Welsh Joint Education Committee); OCR (Oxford, Cambridge and RSA Examinations); CCEA (Council for the Curriculum, Examinations and Assessment). We do not show the information of an additional awarding body that accounts for almost no entries.

These KS4 results files do not include, however, information on the exact marks obtained by students. We are able to merge a novel dataset including detailed information on pre-appeal and post-appeal marks from AQA. We also have access to information on who asked for a review on the

[12] Awarding organisations (also called awarding bodies or exam boards) design, develop, deliver and award the recognition of learning outcomes (knowledge, skills and/or competences) of an individual following an assessment and quality assurance process that is valued by employers, learners or stakeholders (Federation of Awarding Bodies: http://www.awarding.org.uk/about-us/about-awarding-bodies). Awarding bodies are regulated and overseen by Ofqual (a non-ministerial government department with jurisdiction in England).

different units of the GCSE English and GCSE English Language qualifications, and on the tier of the externally assessed unit (i.e. whether Foundation or Higher Tier). The first row in Table A2 shows that the number of AQA entries that we are able to match to KS4 entries is lower than the recorded AQA entries in the KS4 results dataset (shown in row 1 of Table A1). This is for four main reasons. First, this is due to technical problems in providing Unique Candidate Numbers (UPN) for all candidates.

Table A2
GCSE English Language. Working Sample

	Observations
1. Matched AQA-NPD entries	208177
2. Candidates with no discounted entries (and no duplicates)	201073
3. Candidates with no inconsistency in grades across datasets	200983
4. Candidates with data for all controls	189485
a. Higher Tier all (of which C&D)	146747 (49231)
b. Foundation Tier all (of which C & D)	42738 (33034)

Note. NPD entries refer to the entries for AQA GCSE Full Courses found in the KS4 results dataset for academic year 2012-2013, summer season sittings.

Second, not all entries provided by AQA that had a candidate UPN could be matched to the NPD. Third, there could be mistakes in the UPN or the date of birth registered by AQA or the NPD that would make a match impossible in these cases. Finally, candidates taking the examinations with AQA overseas (i.e., Isle of Man, Jersey, Guernsey) would not be matched to the NPD data. All in all, the number of AQA entries that were matched (208177) to AQA entries in the KS4 results file (241539) account for 86.2% of all the KS4 GCSE English Language qualifications taken with AQA.

Students can attempt GCSE qualifications in the same subject (also called discounting group in the data) more than once. While this was a rather common practice for other subjects (like mathematics), this did not seem to happen very often for GCSE English Language qualifications. However, our first sample selection criteria follows the advice given by the Department for Education (DfE) to deal with this issue. This consists of: (1) keeping those entries that are undiscounted (i.e., this is normally the best entry in terms of achievement in the discounting group for exam year 2013, *ks4_disc3=0*); (2) keeping those entries associated with students at the end of KS4 (*ks4_endks=1*); (3) keeping those entries that should be included in national results calculations (*ks4_natres=1*); and (4) keeping those entries that are included in school performance calculations (*ks4_include=1*). After applying these restrictions, we are left with a sample that accounts for almost 97% of the initial sample (see row 2, Table A2).

We detected inconsistencies between the grades in the different datasets (i.e. AQA supplied data *versus* KS4 data) in a small number of cases. The observations available after dropping those entries from the sample barely changes (see row 3, Table A2). The last sample restriction is given by the availability of data to construct controls from the Student Census dataset, which is also part of the National Pupil Database. This involves a bigger cut to the initial sample, and is explained by the fact that only students in state schools are included in the student census. The final number of candidates for which we have data for all controls is about 91% of those initially available (see Table A2, row 4).

A number of assessment units feed into the overall GCSE grade. In 2013, Units 2 (Speaking and Listening (accounting for 20% of the final grade) and Unit 3 (Extended reading and creative writing, 40%) were teacher assessed (although grading was moderated by the exam board). Unit 1 (40%) is based on a standardised exam that is corrected (anonymously) by an external examiner. Exams take place after the coursework assessment (at the end of the school year). We can divide the sample available into two groups, depending on the type of exam that students sat for Unit 1, since students can sit either the Higher Tier or the Foundation Tier exam, and these two exams vary in their content (i.e., both the texts under study and the questions are different). Students sitting the Higher Tier exam can only score grades from A* to D for that particular unit; whereas students sitting the Foundation Exam can achieve a C grade at most for Unit 1. Marks for the three units are added up and make the final GCSE English Language grade, that can range from A* to G, where fails (below G) are awarded the letter U (for ungraded). Most students sit the Higher Tier exam (about 77% of the sample). These students are the main group of interest throughout the paper. Finally, given the nature of the identification strategy and the focus on students marginally failing to achieve a C grade, we restrict our attention to students that obtained either a C or a D grade, before and after the review process (i.e. we exclude students that suffer big jumps in their marks after the review process, since this might be due to measurement error). There are 49,231 students fulfilling the underlined criteria and that will therefore constitute the main sample in our analysis.

A.2. Grade setting in English Language GCSE

As explained in the previous section, three units feed into the overall GCSE English Language mark. Teachers (for the teacher-assessed units 2 and 3) and external markers (for unit 1) are not given advance information on how raw marks on the different assessment units are translated to the 'unified marking scheme' (UMS), which is the format of the final marks (and is on a scale of 0-300; where 180 is the threshold of a C grade).[13] Table A3 shows how raw marks for the three different units are translated into UMS marks, in June 2013 (Panel A) and in June 2012 (Panel B). The raw mark that corresponds to the C grade in each of the three units changes from year to year, making it very difficult for teachers to accurately guess where the (180 UMS) C threshold would be in terms of raw marks. Moreover, for teacher-assessed units, the exam board issues strict grading guidelines, and this marking can also be subject to reviews if inconsistencies are detected. For the externally examined unit, AQA employs online marking since 2012. With this system, markers are not given whole scripts from specific centres but instead, are allocated 'clips' from scripts to mark (i.e. a specific question from a paper). Thus, for example, an individual candidate will not have her entire English Language script marked by a single examiner. Instead, the questions on that script will have been marked by different examiners.

Grade boundaries are not decided in advance of the exam. When setting grade boundaries, exam boards consider: (1) student's work; (2) reports from senior exam officials about how well the units worked in practice; (3) examples of typical performance expected of students at certain grades; (4) statistics; and (5) archived exam papers at the grade boundaries from previous exam series.[14] The awarding committee does not look at work at every grade of each paper, but scrutinises work and explicitly recommends grade boundaries for specific grades only. These are called the judgemental grades in recognition of the fact that awarders' judgements are directly involved in the boundary setting. For the GCSE AQA English

[13] From 2013, teachers did not know how raw grades would translate into UMS marks for the controlled assessments. This was a change from the previous year when there had been controversy about potential teacher bias.
[14] https://www.gov.uk/government/publications/gcse-and-a-level-exams-how-marking-and-grading-works/marking-and-grading-in-gcse-and-a-level-exams

Language higher tier qualification, the awarding committee looks at the boundary between grades C and D first. Next, the boundary between grades A and B is considered. Any remaining grade boundaries are called arithmetic boundaries because they are determined by calculation, without any judgement involved (AQA, 2017).

Table A3
Raw and Uniform Mark Scale marks

	Unit 1		Unit 2		Unit 3			
	Raw Mark	UMS	Raw Mark	UMS	Raw Mark	UMS	Total Raw marks	Total UMS
				A. June 2013				
A*	58	108	41	54	72	108	171	270
A	53	96	38	48	65	96	156	240
B	48	84	34	42	56	84	138	210
C	43	72	30	36	47	72	120	180
D	38	60	25	30	37	60	100	150
				B. June 2012				
A*	61	108	41	54	72	108	174	270
A	55	96	38	48	64	96	157	240
B	49	84	33	42	55	84	137	210
C	44	72	28	36	46	72	118	180
D	39	60	23	30	36	60	98	150

Note. Marks correspond to GCSE English language, June 2013 and June 2012 sittings; higher tier students. The maximum raw mark in Unit 1 is 80; the maximum raw mark in Unit 2 is 45; and the maximum raw mark in Unit 3 is 80. The data is for the AQA awarding body. Unit 1 is externally assessed, whereas Units 2 and 3 are teacher assessed.

After the exam, requests for a review (i.e. re-mark) of scripts can only come through the school (i.e. not from the individual student) and at a price of roughly £40 per script. At this point, there is a possibility that different schools will vary in their propensity to request re-grading for marginal students. In 2013, there were appeals for about 2 per cent of all GCSE exams, with about one in six appeals leading to a grade change (Office of Qualifications and Examinations Regulation, 2013). Marks can either increase or decrease through the appealing process.

A.3. Other data

A.3.1. Student Census

We use the spring pupil-level census (PLASC) dataset for the academic year 2012-2013 to incorporate predetermined characteristics that we use throughout the paper. This dataset has information on pupils attending state schools, and is one of the datasets within the National Pupil Database. The controls that we construct from this dataset are as follows: (1) a dummy variable indicating whether the student is of white ethnicity (*ethnicgroupmajor_spr13='WHIT'*); (2) a dummy variable indicating whether English is the pupil's major language group (*ethnicgroupmajor_spr13='1_ENG'*); (3) a variable indicating whether the student is eligible to receive Free School Meals (*fsmeligible_spr13=1*).

A.3.2. Key Stage 2 (KS2)

We use Key Stage 2 data corresponding to our cohort to construct prior attainment outcomes. This marks the end of primary school education, where there is an externally assessed test in English, maths and science. This forms the basis of the performance tables for primary schools. We use Key Stage 2 raw test scores to build a variable of prior attainment at age 11. The raw test score is graded out of 80 for science and is the sum of two separate science papers each marked out of 40 (total mark is given in the KS2 datasets as *ks2_scitotmrk*). The English test score is marked out of 100 and is composed of the sum of two separate test scores, each marked out of 50, in reading and writing (*ks2_engtotmrk*). Finally, maths is composed of two marks out of 50 with one of the tests being in mental arithmetic (*ks2_mattotmrk*). We construct the measure as follows: [(*ks2_mattotmrk+ks2_engtotmrk+ ks2_scitotmrk**(5/4))/3].

A.3.3. Key Stage 4 (KS4) Candidate Indicator dataset

The Key Stage 4 Candidate/Indicator dataset contains information on the assessment of learners at the end of their years of compulsory schooling (when they are aged 16, in Year 11). Whereas the KS4 Results dataset contains information at the subject level, this data set contains information at the pupil level. We use this dataset to obtain indicators of performance in GCSE Mathematics. We additionally construct a gender variable with the information contained in the KS4 Candidate Indicator dataset.

A.3.4. Key Stage 5 (KS5)

We use Key Stage 5 data to construct outcomes (see section A3). This dataset has information on the post-16 assessment of learners in school sixth forms, sixth form colleges and General and Tertiary Further Education Colleges. We use the files that contain information about the 2013/14 to 2015/16 academic years. See Hupkau et al (2017) for a more in-depth description of the post-16 education landscape in England. We also use this dataset to obtain information on the educational institution attended at 17 (together with the below dataset). Specifically, we construct indicators on the type of institution attended as well as the quality of the institution attended at 17. The latter variable uses information in the Key Stage 4 Candidate Indicator dataset described in (c) above. The quality of the institution attended for each student is measured by the fraction of students (excluding the student him/herself) attending the same institution at age 17, that achieved five GCSEs (or equivalent) at grades A*-C including English and maths (using the variable *ks4_level2_em*). We also construct measures of peer quality as the fraction of students attending the same institution at age 17 that achieved a C grade in GCSE English, and in GCSE Maths.

A.3.5. Individual Learner Records (ILR)

The Individualised Learner Record (ILR) dataset consists of two main datasets: the aims and the learner files. Whereas the former collects information on each of the aims (or subjects/qualifications) the student is enrolled in, the second file has information at the learner level (i.e., provider/

school attended, gender, and other relevant learner characteristics). These pertain to post-16 education and need to be used in conjunction with the Key Stage 5 file (described above). We use data from 2013/14 to 2015/16 in order to construct outcomes (see section A3). As in Section A2d, we use this dataset to obtain information on the educational institution attended at 17.

A.3.6. Higher Education Statistics Agency Dataset (HESA)

HESA records contain information on Higher Education Participation and outcomes. We merge information for the academic year 2015/16 (the first year that, by age, this cohort can be observed participating in Higher Education). All the datasets described so far can be merged by using the Pupil Matching Reference (PMR) indicator number that is present across all of them.

A.3.7. Longitudinal Education Outcomes Dataset (LEO: P14 and Self-assessment)

We use information about annual earnings in tax year 2015 (i.e., from 6th April 2014 to 5th April 2015) and income coming from the Self-Assessment files in tax year 2015 from the Longitudinal Education Outcomes (LEO) dataset. This information comes from HMRC tax records. More specifically, the earnings information comes from the annual statement of total earnings subject to taxes and national insurance that is issued at the end of each financial year (P14 form). These two datasets are used to construct an indicator of whether the student is a NEET at age 18 (i.e., not observed in education, employment or training at age 18). A detailed explanation of the construction of this variable is given in section A3 below. The files in the LEO dataset can be merged to the NPD, ILR and HESA datasets by using two look-up tables provided by the Department for Education (previously Department for Business, Innovation and Skills) that allow recovering the Pupil Matching Reference (PMR) indicator for each of the records.

A.4. Construction of outcomes

A.4.1. Not observed in education at age 18

We create a dummy variable that is equal to 1 if the student is not observed in any of the education datasets that the student should be registered in if he/she was enrolled in any sort of qualification during the academic year 2014/15. This corresponds to the year when the student is 18 years of age, – that is, two years after the completion of compulsory education (or Key Stage 4). Specifically, we construct the variable as equal to zero if the student does not appear in the 2014/15 KS5 Candidate indicator dataset; and he/she does not appear as taking any subjects (aims) in the KS5 Results dataset (ILR Aims dataset) in exam year 2014/15. The dummy variable is equal to one otherwise.

A.4.2. Not observed in education, employment or training (NEET) at age 18

We amend the previous variable to construct a proxy indicator for whether the individual is classified in the NEET category two years after having undertaken GCSEs. Specifically, we create a dummy variable that is equal to 1 if the individual is not observed in education at age 18 (during academic year 2014-15), and the individual has zero total annual earnings in the P14 files and zero income coming from the Self-Assessment files in the tax year 2015. The dummy variable is equal to one otherwise (i.e. the individual is observed in any form of education in the academic year 2014/15 or the individual has positive earnings or income in the P14 or Self-Assessment files).

A.4.3. Entry to an upper secondary academic or vocational qualification by age 19 (i.e. Observed in any Level 3 qualification)

We use the information in the KS5 datasets and in the ILR aims dataset to construct an indicator for whether the individual has ever enrolled in any Level 3 qualification (independently of the size of the qualification). This is a measure of whether the individual enters an upper secondary academic or vocational qualification by the age of 19. We classify an individual as having enrolled in any Level 3 qualification by age 19 if in any of the three academic years after KS4 completion (i.e, 2013/14, 2014/15 or 2015/16), at least one of the following is true: (1) the individual appears in any of the KS5 datasets for any of the three academic years after KS4 completion *and* the sum across subjects of *ks5_asize* is strictly bigger than zero (i.e., ks5_asize is a variable indicating whether any of the subjects that the student is enrolled in is equivalent to A-levels); (2) the individual appears in the ILR AIMS dataset with at least one aim – in any of the three academic years after KS4 completion – at Level 3 or above. The information about the level of an aim is obtained from merging the files from the Learning Aim Reference Service Datasets that are publicly available online. This information can be merged based on a variable that contains information on the *learning aim reference*.

A.4.4. Achieved a Full-Level 3 qualification (i.e. upper-secondary) by age 19

A full level 3 qualification is obtained when the student achieves at least two A-level (or equivalent qualifications) passes. In particular, we classify an individual as having fulfilled a full-level 3 qualification if at least one of the following is true: (1) the individual is observed as having a value of 1 in the variable *ks5_pass2lv3* in the KS5 Candidate Indicator dataset, in academic years 2013/14 or 2014/15; (2) the individual is observed as having 2 or more passes in the variable *ks5_passes_tot* in academic year 2015/16[15]; (3) the individual is observed in the ILR Learner files in any of the 3 academic years following KS4 completion with a value of the variable *ill_l_fulllevel3ach* that is equal to one.

A.4.5. Enrolled in tertiary education (i.e. a qualification of Level 4 or above) at age 19

This outcome is an indicator of whether the individual has enrolled in any Level 4 or above qualification (i.e. tertiary education) three years after the completion of KS4 (in academic year 2015/16). We classify an individual as being enrolled in any Level 4+ qualification (irrespective of the size of the qualification) if at least one of the following is true: (1) the student is observed in the HESA dataset with values of *he_xlev501* different than five (i.e. in practice, this implies that the student has started a university degree); (2) the individual appears in the ILR AIMS dataset with at least one aim in academic year 2015/16 at Level 4 or above.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpubeco.2020.104224.

[15] The variable *ks5_pass2lv3* is not available in academic year 2015/16, so we have to define the variable using an alternative approximation.

References

Anelli, M., 2016. The Returns to Elite College Education: A Quasi-experimental Analysis CESifo Working Paper Series. (No. 6076).

Angrist, J., Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Q. J. Econ. 114 (2), 533–575.

Angrist, J.D., Battistin, E., Vuri, D., 2017. In a small moment: class size and moral hazard in the Mezzogiorno. Am. Econ. J. Appl. Econ. 9 (4), 216–249.

Apperson, J., C. Bueno and T. Sass. (2016). Do the cheated ever prosper? The long-run effects of test-score manipulation by teachers on student outcomes. CALDER working paper no. 155. National Center for Analysis of Longitudinal Data in Education Research. US.

Avery, C., O. Gurantz, M. Hurwitz, and J. Smith. (2018). Shifting college majors in response to advanced placement exam scores. Journal of Human Resources, forthcoming.

Barreca, A., Guldi, M., Lindo, J., Waddell, G., 2011. Saving babies? Revisiting the effect of very low birth weight classification. Q. J. Econ. 126, 2117–2123.

Battistin, E., Neri, L., 2017. School Accountability, Score Manipulation and Economic Geography. Queen Mary University, Mimeo.

Battistin, E., Rettore, E., 2008. Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. J. Econ. 142 (2), 715–730.

Bell, B., Bindler, A., Machin, S., 2018. Crime scars: recessions and the making of career criminals. Rev. Econ. Stat. 100 (3), 392–404.

Betts, J.R., 1998. The impact of educational standards on the level and distribution of earnings. Am. Econ. Rev. 88 (1), 266–275.

Borcan, O., Lindahl, M., Mitrut, A., 2017. Fighting corruption in education: what works and who benefits? Am. Econ. J. Econ. Pol. 9 (1), 180–209.

Calonico, S., Cattaneo, M.D., Titiunik, R., 2014. Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica 82 (6), 2295–2326.

Canaan, S., Mouganie, P., 2018. Returns to education quality for low-skilled students: evidence from a discontinuity. J. Labor Econ. 36 (2), 395–436.

Clark, D., Martorell, P., 2014. The signaling value of a high school diploma. J. Polit. Econ. 122, 282–318.

Costrell, R.M., 1994. A simple model of educational standards. Am. Econ. Rev. 84 (4), 956–971.

Dearden, L., McIntosh, S., Myck, M., Vignoles, A., 2002. The returns to academic and vocational qualifications in the UK. Bull. Econ. Res. 54, 249–274.

Dee, T., Dobbie, W., Jacob, B., Rockoff, J., 2019. The causes and consequences of test score manipulation: evidence from the New York regents examinations. Am. Econ. J. Appl. Econ. 11 (3), 382–423.

Department for Education, 2016. Level 1 and Level 2 Attainment in English and Maths by Students Aged 16–18: Academic Year 2014/15, 24 May 2016 (Statistical First Release. Department for Education).

Diamond, R., Persson, P., 2016. The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests (No. w22207). National Bureau of Economic Research.

Ebenstein, A., Lavy, V., Roth, S., 2016. The long run economic consequences of high-stakes examinations: evidence from transitory variation in pollution. Am. Econ. J. Appl. Econ. 8 (4), 36–65.

Feng, A., Graetz, G., 2017. A question of degree: the effects of degree class on labour market outcomes. Econ. Educ. Rev. 61, 140–161.

Frandsen, B., 2017. Party bias in union representation elections: testing for manipulation in the regression discontinuity design when the running variable is discrete. Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, Volume 38). M. Cattaneo and J. Escanciano. Emerald Group Publishing, pp. 29–72.

Freier, R., Schumann, M., Siedler, T., 2015. The earnings returns to graduating with honors – evidence from law graduates. Labour Econ. 34, 39–50.

Frontier Economics (2015). Understanding Awarding Organisations' Commercial Behaviour Before and After the GCSE and A-level reforms. Report Prepared for the Office of Qualifications and Examinations Regulation. Ofqual/15/5596.

Goodman, J., Hurwitz, M., Smith, J., 2017. Access to 4-year public colleges and degree completion. Journal of Labor Economics 35 (3), 829–867.

Gregg, P., Tominey, E., 2005. The wage scar from male youth unemployment. Labour Econ. 12, 487–509.

Hahn, P., Todd, J., van der Klaauw, W., 2001. Identification and estimation of treatments with a regression discontinuity design. Econometrica 69 (1), 201–209.

Higton, J. R. Archer, D. Dalby, S. Robinson, G. Birkin, A. Stutz, R. Smith, V. Duckworth. (2017). Effective practice in the delivery and teaching of English and mathematics to 16–18 year olds. Report for Department for Education, UK. DFE-RR742.

Hupkau, C., McNally, S., Ruiz-Valenzuela, J., Ventura, G., 2017. Post-compulsory education in England: choices and implications. Natl. Inst. Econ. Rev. 240 (1), 42–56.

Hvidman, U., Sievertsen, H.H., 2019. High-stakes grades and student behavior. J. Hum. Resour. https://doi.org/10.3368/jhr.56.3.0718-9620R2.

Imbens, G., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. J. Econ. 142, 615–635.

Kaufmann, K.M., Messner, M., Solis, A., 2013. Returns to Elite Higher Education in the Marriage Market: Evidence from Chile, Working Papers 489, IGIER (Innocenzo Gasparini Institute for Economic Research). University, Bocconi.

Kolesár, M., Rothe, C., 2018. Inference in regression discontinuity designs with a discrete running variable. Am. Econ. Rev. 108 (8), 2277–2304.

Lavy, V., Sand, E., 2018. On the origins of gender gaps in human capital: short- and long-term consequences of teachers' biases. J. Public Econ. 167, 263–279.

MacLeod, W.B., Riehl, E., Saavedra, J.E., Urquiola, M., 2017. The big sort: college reputation and labor market outcomes. Am. Econ. J. Appl. Econ. 9 (3), 223–261.

McIntosh, S., 2006. Further analysis of the returns to academic and vocational qualifications. Oxf. Bull. Econ. Stat. 68, 225–251.

Office of Qualifications and Examinations Regulation (2013). Enquiries About Results for GCSE and A-level: Summer 2013 Exam Series. Statistical Release. Ofqual/13/5357.

Papay, J., Murnane, R., Willett, J., 2015. The impact of test-score labels on human- capital investment decisions. J. Hum. Resour. 51 (2), 357–388.

Patrignani, P., G. Conlon and S. Hedges. (2017). The earnings differentials associated with vocational education and training using the longitudinal education outcomes data, Centre for Vocational Education Research, London School of Economics. Discussion Paper 007.

Smith, J., Hurwitz, M., Avery, C., 2017. Giving college credit where it is due: advanced placement exam scores and college outcomes. J. Labor Econ. 35 (1), 67–147.

Terrier, C., 2020. Boys lag behind: How teachers' gender biases affect student achievement. Econ. Educ. Rev. 77. https://doi.org/10.1016/j.econedurev.2020.101981.

Zimmerman, S.D., 2014. The returns to college admission for academically marginal students. J. Labor Econ. 32 (4), 711–754.