HCEO WORKING PAPER SERIES

# Entry Through the Narrow Door: The Costs of Just Failing High Stakes Exams

Stephen Machin
Sandra McNally
Jenifer Ruiz-Valenzuela

Working Paper

**2018-029**

05/2018

HUMAN CAPITAL AND ECONOMIC OPPORTUNITY GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

# Entry Through the Narrow Door: The Costs of Just Failing High Stakes Exams

**Stephen Machin*, Sandra McNally** and Jenifer Ruiz-Valenzuela*****

April 2018

## Abstract

In many countries, important thresholds in examinations act as a gateway to higher levels of education and/or good employment prospects. This paper examines the consequences of just failing a key high stakes national examination in English taken at the end of compulsory schooling in England. It uses unique administrative data to show that students of the same ability have significantly different educational trajectories depending on whether or not 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 are therefore a source of educational inequality with high potential long-term consequences for those affected.

**Keywords**: high stakes examinations; manipulation; English.

**JEL codes:** I20, I21, I24

* Department of Economics and Centre for Economic Performance, London School of Economics (Houghton Street, WC2A 2AE, London, United Kingdom). Email: S.J.Machin@lse.ac.uk

** School of Economics, University of Surrey and Centre for Economic Performance, Centre for Vocational Education Research. London School of Economics (Houghton Street, WC2A 2AE, London, United Kingdom). Email: s.mcnally1@lse.ac.uk

*** Centre for Economic Performance, Centre for Vocational Education Research. London School of Economics (Houghton Street, WC2A 2AE, London, United Kingdom). Email: j.ruiz-valenzuela@lse.ac.uk

**Acknowledgements:** 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 of 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), and the National Institute of Economic and Social Research. We would also like to thank participants at the 2017 Royal Economic Society Conference (Bristol), the 2017 Meetings of the Society of Labor Economics (Raleigh), the 2017 RIDGE Workshop in Public Economics (Montevideo), the 2017 conference of the Centre for Economic Performance (London), the 2017 Workshop on Human Capital and Public Policy (Stockholm), the CESifo Area Conference on Economics of Education (Munich, especial thanks to Guido Schwerdt), the 2017 EALE conference (St Gallen) and the 2nd IZA Workshop on the Economics of Education (Bonn).

## 1. Introduction

Getting above or failing to reach thresholds in exams is an important feature of success or failure in many people's lives. Indeed, scoring above or below a particular threshold can prove important for longer term outcomes in many settings. Examples include different degree classifications, acquiring a high school diploma or reaching a certain grade point average, to name just a few.

In some contexts, achievement of particular qualification levels is vital from the perspective of educators, employers and governments. The need to obtain a grade C in English and maths in the age 16 school leaving examinations in England is one well known example. This is in part because achievement of good literacy and numeracy skills is recognised as an important output of the education system, especially as England has consistently underperformed in this regard.[1] It is also because achieving a 'good pass' (i.e. grade C or better) in these exams has long been recognised as a key requirement for employment.[2] In fact, this level of achievement is deemed so important that recently (since 2015), 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.[3]

Exam thresholds have also become increasingly important for incentivising teachers and school managers. This is especially true in decentralised education systems where mechanisms like pay for performance operate and where school rankings can play a role in this.[4] In such settings, worries have emerged that this can lead to manipulation of marks by

[1] The percentage of young people with low basic skills in literacy and numeracy is close to 30% in England according to the OECD survey of basic skills, and is one of few countries where there has been no improvement amongst the younger generation compared to the older generation (Kuczera et al. 2016).
[2] To give one example, it is now a requirement for nursery school teachers to have achieved a grade C in English and maths.
[3] However, the pass-rate for those students re-taking the GCSE exam is very low. According to official figures, 29.3% of the students failing to get a C grade in the cohort we study have re-taken the exam two years after failing to get a C grade; with 45% of those retaking it managing to secure a C grade or above (Department for Education, 2016).
[4] See the discussion by Johnes (2004) for example.

teachers. In addition, there is also the concern that disadvantaged students can directly lose out because of such manipulation. There is a growing literature (discussed below) which evaluates the consequences of such teacher bias.

This paper offers an empirical study of a high stakes exam and analyses the benefits (or costs) for students who just pass (or fail) to meet a key threshold. The context is national examinations taken at the end of compulsory age schooling in England where access to rich administrative data enables study of detailed grades and marks, together with institutional features of the grading system that may have led to manipulation of marks and grades. More specifically, evidence is presented on the importance of just obtaining a grade C – a good pass – in English in high stakes national examinations taken for the General Certificate of Secondary Education (or GCSEs) when students are 16 years of age.[5]

The administrative data covers a recent GCSE cohort and follows them for three years after their exam. Comparing students on the threshold of success/failure enables analysis of whether just passing/failing has consequences for them in relation to their probability of early drop-out from education (and employment) and their probability of accessing higher-level courses, which are known to have a positive wage return in the labour market. The analysis also looks at the effect on the probability of entering tertiary education. The question is not so much whether it is important to perform well in English, as to whether it is important to get past the specific threshold of a grade C. In other words, the focus is on isolating the effect of good or bad luck, which leads one to end up on either side of the C threshold. Up to now this

[5] 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 which accounts for over half of exams in English (discussed later in the paper and in Appendix A). We do not have comparable information for maths

has not been evaluated empirically, even though getting a grade C in English is given great weight within English institutions and in popular discourse.[6]

The paper makes use of the distribution of exact marks around the important threshold of grade C. The empirical challenge is to address potential endogeneity around who passes this threshold. This has some similar features to two recent papers, one studying a national examination in Sweden, the other a high school exit examination from New York, where possible teacher manipulation has been placed centre stage. In the former, Diamond and Persson (2016) report significant test score manipulation around known grade thresholds in the national mathematics tests taken by ninth graders in Sweden. They conclude that this generates an unexpected benefit to pupils manipulated across the threshold because they get longer-term improvements in education and earnings. In the latter, Dee at al. (2016) demonstrate that manipulation took place in the New York Regents exam taken by high school students, and that crossing the score cut-off due to this raised high school graduation (although it also lowered the likelihood of taking advanced coursework and of college enrolment). To show this, they exploit the reforms introduced by the education authorities to eliminate the test score manipulation.

Both Diamond and Persson (2016) and Dee et al. (2016) have teacher cheating or bias in mind as the underlying mechanism behind grade score manipulation. This relates to several other 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) as well as to a literature that examines the effect of teacher bias in marking more generally (e.g. Lavy and Sand, 2015; Terrier, 2016). Teacher bias is also behind the test-score manipulation analysed by

[6] Getting a grade C in English and/or maths is often a pre-requisite to higher-level courses in post-compulsory education and can affect whether a student is admitted to post-16 institutions. It is also something considered by universities in their admissions criteria. It forms part of the school-level indicators that are in published School Performance Tables. Additionally, much emphasis is given on articles in newspapers and on the Internet about GCSE results and what to expect afterwards.

Angrist et al. (2017) in a region of Italy and accounts for the observed relationship between class size and student achievement. In contrast to other papers on test score manipulation, we show that requests to re-mark externally administered scripts are behind the observed 'manipulation' in our context (excess mass in the distribution of marks to the right side of the C cut-off).[7]

There are some unique features of the data and the institutional setting used in this study. These enable a different methodological approach to be adopted and to generate a causal impact of just passing or failing a key high stakes exam that is free of any concern about manipulation bias. First, one key feature of English examinations is a right to appeal, and whilst the administrative data contains final (i.e. post-appeal) grades, we have also obtained access to student level data on the pre-appeal and post-appeal marks. This is important since we can use these data to ascertain whether or not what looks like manipulation in the data is actually due to the regrading process through appeals. Our paper is unique in having the 'pre-manipulation' and 'post-manipulation' distribution of marks for the same students. Second, the threshold we consider (grade C in English) is well known in an English context and is explicitly sought, not only by students, but also by schools.[8] Our context is unusual in that we are looking at the importance of passing this threshold at the end of compulsory education or lower secondary education (when students are about 16 years of age) rather than, for example, older students at the end of high school in other countries.

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 very different

[7] 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.

[8] It is not the only such indicator, as getting a grade C in maths is also important, as is achieving 5 or more grades at A*-C at GCSE. As documented above, these indicators are often used as pre-requisites for advancement in education and by some employers.

educational contexts. 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.[9] Canaan and Mouganie (2017) study the impact of marginally passing the French high school exit exam on choice of higher education institution and degree subject. Finally, in the educational context considered here, the ranking of schools (and their managers) by pupil performance has become a central feature of the school system. Competition has been promoted by such measures as the publication of school performance tables (since the mid-1990s) and more recently by large-scale school autonomy. Teachers and head-teachers are highly incentivised to make sure students perform well in high-stakes tests, making sure as many as possible pass important thresholds such as that considered here (e.g. see Cassen et al. 2015).

The findings reported in this paper show that failing to achieve a grade C in English has a large associated cost – or 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. As a result of just failing to obtain a grade C in English, students are more likely to drop out of education early and become classified as 'not in education, training or employment' (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). They are also 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.

[9] Other related examples include the effect of achieving a higher score on choice of major (Avery et al. 2016); the effects of class of degree on earnings (e.g. Feng and Graetz, 2017; Freier et al. 2015); and how test score labels affect human capital investment decisions (Papay et al. 2015).

We show some evidence on the mechanisms through which failing to obtain a grade C in English leads to poor outcomes. 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. Also, students end up in institutions with a worse academic environment (as measured by peer quality). In a well-functioning education system, there would be ladders for the marginal student – or at least alternative educational options with good prospects. This paper suggests that the marginal student who is unlucky pays a high price. This is consistent with descriptive evidence which suggests that the English educational system does not work well 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).[10]

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.

[10] In an English terminology, lower-level courses are 'level 2' (GCSE level) and higher level courses are 'level 3' (upper secondary education, equivalent to a post-compulsory high-school setting in other countries). The latter are generally pre-requisites for tertiary education and tend to be associated with positive earnings differentials in the labour market. We give further descriptive evidence about this in an exercise using the Labour Force Survey in Section 5.

## 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.

The GCSE exam is very important because getting a 'good grade' influences the level of the course that the student can start and potentially the type of institution the student can attend. 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 exams in these subjects.[11] Getting a C grade is often a pre-requisite for advanced academic or vocational courses. Universities will also consider students' GCSE grades (as well as subsequent advanced qualifications) when deciding whether to offer a place to an applicant. The C grade is also important for schools since the percentage of students who achieve grades above this threshold is a component of the (published) Schools Performance Tables.

[11] As referred to above, from 2015 onwards, it has been compulsory for students who do not achieve a C in English or maths to re-sit the exam over the next year (which is typically in a college of further education, where such students will be most likely enrolled in some form of vocational education). The cohort considered here were not compelled to repeat GCSE exams, although they had the option to do so.

GCSE exams are set and marked by different exam boards – of which there are four in England.[12] 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 (at the end of the school year).[13] In the year of relevance to our study (2013), 40% of the overall marks were accounted for by the standardised exam.[14] 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.[15] 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.[16] Thus, it is not possible for teachers to manipulate coursework assessments such that the marginal student just crosses the threshold for a grade C.[17]

---

[12] 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.

[13] In practice, students in our sample could sit the exam before the end of the academic year; although that was more frequent in maths than in English Language. In fact, most students sit English Language at the end of the school year (see Appendix A for further information)

[14] Information is based on the 2013 criteria set out by the AQA exam board, as this is the group for which we have data.

[15] From the year considered here, 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.

[16] 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

[17] 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 per cent 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 pupil-level 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 their gender, eligibility for free school meals, ethnicity and whether they speak English 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 also observe whether or not they enrol in any form of tertiary education by the age of 19. We are also able to link the education data to administrative data on employment and self-employment from the Longitudinal Educational Outcomes data set (LEO). We use data from students who undertook their GCSE exams in June 2013 (when they were aged 16) and can follow them for three years. Appendix A offers a thorough description of the data sources used, as well as the sample selection criteria and construction of variables.

We are able to merge the GCSE exam grade in English to information on pre-appeal and post-appeal marks from one of the four exam boards, the AQA.[18] This exam board accounts for well over half of all exam entries in GCSE English (61.6% of GCSE English Language

[18] Although we have this information for maths from the AQA, this is a much less important subject for this exam body. It only accounts for about 12% of all exam entries in this subject. Hence our focus upon examination performance in English.

entries, and 55.7% of GCSE English entries; see Table A1 in Appendix A).[19] 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 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.[20]

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 sitting the GCSE English Language exam with AQA are 2 percentage points more likely to achieve a C grade or above), AQA students are very similar in terms of predetermined characteristics to all students in the cohort (see columns (1) and (2) in Table 1, second panel). 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

---

[19] Analysis about awarding bodies suggests that schools choose exam boards predominantly on the basis of 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

[20] Students can choose between English and English language (which is normally taken together with the English literature GCSE). The English specification is preferable for those students who want to explore a range of literature and language topics but do not want to take separate GCSEs in Language and Literature. We obtain very similar results for students who undertake English rather than English language. Results are available on request. The vast majority of students undertaking English language take higher tier exams. For the smaller proportion of students taking lower tier exams, the maximum grade achievable is grade C. Results are very similar to the ones shown here for the higher tier students and are available on request.

the cohort than are foundation students. Given our focus on higher tier students, the remaining tables and figures refer to higher tier students only.

The data used are unique in that both the 'pre-manipulation' and 'post-manipulation' distribution 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. This has not been possible in other papers looking at related questions where estimating the counter-factual distribution has been necessary (Dee et al. 2016; Diamond and Persson, 2016).

Figure 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., 2016; Diamond and Persson, 2016). In the English context, however, this is not likely to be 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.[21] However, it may arise from many re-grading requests for students near the boundary. Furthermore, requests for remarking may be biased in relation to students or school characteristics (which we examine below). Figure 2 shows the original distribution of marks (i.e. before re-marking requests) and it overlays the final distribution. This shows that the

[21] There has been online marking since 2012 in which examiners are allocated 'clips' from scripts to mark (i.e. a specific question from a paper and not a whole paper). Thus, questions on each script will have been marked by different examiners (and this is also true for scripts that need to be re-marked because of an appeal by the school).

original distribution of marks is approximately normal.[22] 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.[23] 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 who 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).[24] 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).

Figure 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

[22] Although it is also evident that the distribution is not completely smooth and normal because there is not a one-to-one mapping between the raw scores and the scaled scores.
[23] 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.
[24] 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.

per cent. In contrast, the probability only rises to about 20 per cent near the thresholds for grades B (210 marks), A (240 marks) and A* (270 marks). This is illustrative of how important getting a grade C is within the English education system. The Figure also shows the probability of actually being upgraded. This shows that a high proportion of students who request a re-mark 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 students' 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; English spoken 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.[25] 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.

The average probability of requesting a re-mark is about 10 per cent. Re-marking 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 requesting a re-mark. When school fixed effects are included (column 2), the coefficients decline for both gender and

[25] The marginal effects from a Probit model give very similar results.

prior attainment (though for the latter it is still precisely estimated and statistically significant). This is likely to be a reflection of the fact that requests for re-marking come via the school and not the 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.[26]

## 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 well or badly on the day of assessment. We are interested in establishing the causal effect of getting a C grade 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.

To assess the effect of obtaining a C grade on later outcomes, we make use of the fact that we have the original (pre-appeal) mark distribution and can use this to build an instrument to predict whether a person actually obtains grade C. Figure 4 illustrates the first stage and shows that the original mark is a very strong predictor of whether grade C is finally obtained

[26] In most cases, re-marking is requested for exams and not the controlled assessment. Results are very similar whether we look only at exams or at both forms of assessment together (which is reported here).

(after the appeal process). It is not a perfect predictor because of the possibility of re-grading. The probability is 1 after the critical threshold by construction because 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). 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)). The pattern to the right of the cut-off arises because there is no incentive for schools to enter students for a re-mark if they are too far away from the threshold, since this is costly and there is also a possibility of being downgraded. This is reflected in the pattern of applications throughout the distribution in Figure 3. For students on the left of the 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.

Given the shape of the first stage, fuzzy regression discontinuity methods (Angrist and Lavy, 1999; Hahn et al. 2001) are used where a dummy indicating whether the student originally obtained a C grade (i.e. pre-appeal) is used to instrument for whether or not an individual receives a final C grade in 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 regressions where we limit the sample to individuals that are very close to the grade C threshold in the original (pre-appeal) distribution of marks. 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 fully parametric set-up (under assumptions of linearity). More formally, estimate the following equations can be estimated in a two stage least squares setting:

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

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

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 appeal process, denoted *CF*). Marks of the student are denoted by *M* (these are the original distribution of marks, i.e. pre-appeal) 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 (i.e. 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).[27] $\mu$ 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. $\epsilon_{is}$ and $\omega_{is}$ are error terms and we cluster at the level of the school, following Kolesár and Rothe (2017).[28]

We estimate 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 Figure 4). We then estimate linear regressions over a small range of the data ('local regressions') close to the C threshold (original marks ranging from +/-

[27] The inclusion or exclusion of these pre-determined characteristics makes no difference to any of the estimated effects.

[28] 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 alter standard errors significantly.

5 to +/- 1 points away from the original C threshold). Such students perform very similarly in English except those who pass the threshold of 180 get awarded the C grade.

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 equation (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). It is high stakes for schools because it forms the basis of the School Performance Tables for primary schools.

Figure 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 corresponds to the raw difference in the Key Stage 2 (age 11) test score between those with an original (pre-appeal) mark of between 179 and 180 in GCSE English language. There is no discontinuity around the Grade C threshold in GCSE English (which can be seen visually and by the reported estimate of the difference in the two lines at the discontinuity). 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 Figure B1 in Appendix B).

In Table 3, we report regression estimates where each baseline characteristic is regressed against a dummy variable measuring whether the student obtains a C grade (pre-appeal), controlling for the original (pre-appeal) mark, with and without including school fixed effects. Columns (1) and (2) show regressions estimated for the whole C and D range, whereas columns (3) and (4) show regressions estimated for the subsample of students that are +/- 10 points away from the original C threshold. 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 statistically significance (this is even more so close to the C threshold, see Table B1 for checks done using the +/-5 and +/-1 bandwidth). Hence, it is plausible to conclude that the marginal student who passes the (pre-appeal) 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 by the age of 18; (2) the probability of not being observed in education, employment or training (NEET) by the age of 18; (3) entering a higher-level academic or vocational qualification by the age of 19 (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.[29] Although this cohort is too young to observe labour market earnings, we show that having a level 3 qualification has a high wage premium in Section 5, even if young people do not subsequently go on to tertiary education.[30] Appendix A explains how we have constructed

[29] In England, this implies starting an undergraduate or foundation degree, or enrolling in any sort of high level (level 4 and above) vocational qualification.
[30] This has also been shown in many reports in the UK. See, for example, McIntosh (2006) and Patrignani et al. (2017).

these outcome variables. Table 4 shows summary statistics 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 with marks in the C-D range that are main interest here (column 3) and the subsample of foundation students in the English Language 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 in any of the five dimensions analysed here.

Before showing the regression results, the outcome variables are plotted in Figures 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 equation (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-appeal) 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 with a decrease in the probability of not dropping out of education at age 18 (Figure 6, top figures) as well as a lower probability of being observed as 'not in education, training or employment' (NEET) at age 18 (Figure 6, bottom figures). Figure 7 shows that students who just pass the original C cut-off have a higher probability of accessing (top figures) or achieving (middle figures) a higher qualification by age 19, and starting tertiary education by age 19 (bottom figures). This gives *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 (which is illustrated in Figure B2) or indeed at other points of the distribution.

### 4.2. Baseline Results

In Table 5 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 re-marking) on the outcome of interest. In the notation of equation (1), these correspond to the coefficient $\beta_1$, the (second stage) instrumental variable estimate. The sixth panel (panel F) shows estimated coefficients for the first stage (i.e. $\alpha_1$in equation 2), which is always very large and statistically significant.

Results are very similar across the different specifications (whether they include schools 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 whole sample of students obtaining either a grade C or D, about 9 percent of students have dropped out of any form of education by the age of 18 (rising to 11 percent of students within +/- 10 marks of the Grade C threshold). The effect of just achieving a C grade in GCSE English is to reduce this 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. Specifically this is 3.2% of the sample of students with marks between grade D and grade C, rising to 4% 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.

With regard to starting a higher-level academic or vocational level qualification within 3 years, the effect of marginally achieving a grade C is to increase this probability by between 6 and 9 percentage points. This is a big effect. About 90% of people (in the range of marks from grades C to D) 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).[31]

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 27 percent of this sample have started tertiary education by this age (20.5 percent for those within 10 marks of the C threshold).

### 4.3. Local Regressions with Varying Windows

In Table 6, we show results for subsamples of students who obtain a very narrow range of marks in the original (pre-appeal) 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. In the next section, 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

[31] Donut estimates (see Barreca, Guldi, Lindo and Waddell, 2011) excluding observations that are very close to the C threshold produce very similar results.

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). We show the results including school fixed effects in Table B2 in Appendix B. Table B2 shows the proportion of schools with only one student in each subsample (bottom row). In the sample of students within +/- 5 marks of the C threshold, about 16 percent of schools have one student in this subsample. This rises to half of all schools in the sample of students within +/- 1 mark of the threshold.

In both Table 6 and Table B2, column (1) show 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 6 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 (Table B2) but point estimates are always positive and slightly higher than for the global regressions reported in Table 5. The point estimates are usually consistent across specifications with a different number of students. The outcome showing whether a student enrols in study for a higher-level 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

We conduct several tests to assess the robustness of our results. We start by performing 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. Figure 8 shows the reduced-form ‘intention to treat’ graphs where the relationship between achieving a grade C or above in GCSE Mathematics and the original (pre-appeal) marks in GCSE English Language is shown. There is no visual evidence of a jump around the discontinuity. These results are confirmed in a regression setting. For instance, in the most restrictive case when we compare individuals just above and just below the C threshold, the fuzzy RD estimates –with or without the inclusion of school fixed effects- are of small magnitude and statistically insignificant.[32]

A series of other robustness checks are summarised in Table 7. We show the sensitivity of results to changing the specification in various ways. For simplicity, we show results for the +/-5 bandwidth (columns 1 to 4) and the +/-1 bandwidth (columns 5 and 6). The results in this table can be compared with those in Table 6. Firstly, we show that the results are virtually the same when we do not control for prior attainment or any of the baseline characteristics (Columns 1 and 5). In column (2), we include an exogenous interaction between the forcing variable and the instrument. The point estimates change slightly but the interpretation of the results is virtually the same. This is also true in column (3) when we introduce the forcing variable in a quadratic way in both the first and second stage.

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 would point towards the same

[32] In the +/-1 sample, the coefficient (standard error) when we do not include KS4 school fixed effects is -0.020 (0.027). Including KS4 school fixed effects the results are very similar: -0.025 (0.038).

conclusions. Columns (4) and (6) show marginal effects from Probit estimations of the reduced form equations. As expected given the size of the first stage coefficients, the reduced form estimates are slightly smaller but in line with the estimates that we would obtain for reduced form estimates in a linear setting. Overall, the evidence in Table 7 together with the results of the Frandsen (2017) manipulation tests (Section 2.1) and balancing tests (Section 3.2), 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. One possible reason is that students are set 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. In our context, it would be surprising if psychological effects alone could explain the relatively large effect that narrowly failing a grade C has on outcomes that are achieved by the vast majority of students within this ability range: namely continuing education beyond the age of 18 and (at least) starting a Level 3 qualification by the age of 19.

It is more plausible that the barriers arise because 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.[33]

In Table 8, 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 5. 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 90%). Thus, it can make the difference between achieving and failing to achieve another signal of performance at 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.[34]

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

[33] https://university.which.co.uk/advice/gcse-choices-university/how-important-are-my-gcse-grades
[34] The typical student undertakes 8-10 GCSEs.

GCSEs – where English is particularly important as a core subject. Panel B of Table 8 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 30%. The magnitude is little changed in the narrower windows (columns 3-6).

A bigger door that might close is whether students can attend an academic institution at all. In the sample of students with a grade C-D in English, just under half 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 row 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 reduces 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; those within plus or minus 5 points).

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.[35] In Row 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 52 percent of all students with a grade C-D in English are enrolled in such a qualification at age 17. Marginally failing to make grade C reduces this 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.

[35] Academic courses are also taught within institutions that specialise in vocational education (Further Education Colleges). Many students actually take a combination of academic and vocational subjects.

From the above analysis we see that more doors are closed to students who do not get a Grade C in English at GCSE (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. 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 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 9 where each of these measures of peer quality is the dependent variable. We use the same structure as in Table 8, 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 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 does not 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 even enter a level 3 qualification up to three years later and less likely to enrol in tertiary education. Put another way, the marginal student would have performed significantly better if obstacles had not been put in their path as a result of being unlucky in their English language GCSE exam.

This cohort are too young to evaluate impacts on wages. But using the Labour Force Survey, we can estimate the wage differential from achieving a level 3 qualification (versus a level 2 qualification, which is equivalent to GCSEs or a vocational equivalent). Estimates are shown in Table 10 for the working age population using three specifications with increasingly more detailed controls. The basic estimate (with limited controls) is an hourly wage differential of 11 percent.[36] This reduces to 9.5 percent after including 2-digit industry dummies and then to 6.2 percent in the most detailed specification, which also includes 1-digit occupational dummies. Of course, failing to achieve level 3 qualifications prevents students achieving qualifications higher than level 3 (e.g. university) as students who don't even have level 3 qualifications will not have the right pre-requisites. The high average return to university education in the UK is well established (e.g. Blundell et al. 2005). This implies that the earning losses for those who would have progressed to university education may well be higher than suggested by this simple exercise.

## 6. Concluding Remarks

This study uses one example of a context where examination grade thresholds may be important for future outcomes to identify the effect of narrowly passing (or failing to pass) the critical threshold. It has some similarities to recent papers that evaluate the effects of manipulation in high-stakes tests (Dee et al., 2016, Diamond and Persson, 2016) but is unique in that we have access to the marks of the same students before and after potential endogenous sorting of students across the relevant threshold. In our case, this is due to requests for re-marking, which happen for some students who obtain a mark very close to the relevant grade C threshold. This results in significant bunching of students near this threshold in the (post-

[36] The basic specification controls for potential experience, whether full-time employed, gender, region, ethnicity and time of interview dummies. The sample is restricted to individuals whose maximum level of qualification is either 2 or 3; are aged between 21 and 65, and are respondents in the first wave in the sample.

appeal) distribution of marks, an empirical feature of the distribution that looks like what has been characterised as manipulation in the Dee et al. (2016) and Diamond and Persson (2016) research. As we have data on the original distribution of marks (i.e. before any requests for a re-mark), we can eliminate possible manipulation bias due to regrades by using this to instrument the probability of obtaining a grade C in the English exam at the end of the compulsory phase of education. We are thus able to evaluate the causal effect of narrowly achieving (or failing to achieve) this important threshold.

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 pre-requisite for accessing higher-level courses and institutions (including university) and is a component of indicators published in the School Performance Tables (where performance in English and maths is specifically highlighted). However, up to now the importance of obtaining a grade C in English has never 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 or not they just pass the critical threshold or just fall short of it. Our analysis suggests that 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 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.

This impact on the outcomes considered in this paper matter 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 a higher-level 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, training or employment’ 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, Bindler and Machin, 2017).

More generally, this analysis does not suggest 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. However, if there are big consequences from narrowly missing out on a C grade, this suggests that there is something going wrong within the system. It suggests that young people are not getting the support they need if they fail to make the grade (even narrowly). It also suggests that other educational options available to people who cannot immediately enter higher academic/vocational education are failing to progress a significant proportion of young people up the educational ladder. Thus, it is symptomatic of an important source of inequality in education, with associated negative long-term economic consequences for individuals who just fail to pass such an important high stakes national examination taken at the end of compulsory schooling.

**References**


Altonji J., E. Blom and C. Meghir (2012). Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers. Annual Review of Economics, 4, 185-223.

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.

AQA (2017). A Basic Guide to Standard Setting. Version 1.3. Centre for Education Research and Policy.

Avery, C., O. Gurantz, M. Hurwitz, and J. Smith. (2016). Shifting College Majors in Response to Advanced Placement Exam Scores. NBER Working Paper 22841.

Angrist, J. and V. Lavy. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. Quarterly Journal of Economics. 114(2): 533-575

Angrist, J. D., E. Battisin and D. Vuri. (2017). In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno. American Economic Journal: Applied Economics, 9(4): 216–249.

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

Battistin. E., and E. Rettore (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. Journal of Econometrics, 142(2): 715-730.

Barreca, A., M. Guldi, J. Lindo and G. Waddell (2011). Saving babies? Revisiting the effect of very low birth weight classification. The Quarterly Journal of Economics, 126, 2117-2123.

Bell, B., A. Bindler and S. Machin (2017). Crime Scars: Recessions and the Making of Career Criminals. Review of Economics and Statistics, (forthcoming).

Borcan, O., M. Lindahl, and A. Mitrut (2017). Fighting Corruption in Education: What Works and Who Benefits? American Economic Journal: Economic Policy, 9(1): 180–209

Blundell, R., L. Dearden, and B. Sianesi, (2005). Evaluating the impact of education on earnings: Models, methods and results from the NCDS, Journal of the Royal Statistical Society Series A, 168(3): 473-512.

Canaan, S., and P. Mouganie (2017). Returns to Education Quality for Low-Skilled Students: Evidence from a Discontinuity. Journal of Labor Economics. Forthcoming.

Clark, D., and P. Martorell (2014). The Signaling Value of a High School Diploma. Journal of Political Economy, 122, 282-318.

Cassen, R., S. McNally and A. Vignoles. (2015). Making a Difference in Education: What the Evidence Says. Routledge.

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.

Dearden, L., S. McIntosh, M. Myck and A. Vignoles (2002). The Returns to Academic and Vocational Qualifications in the UK. Bulletin of Economic Research, 54, 249-274.

Dee, T., W. Dobbie, B. Jacob, and J. Rockoff (2016). The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations (No. w22165). National Bureau of Economic Research.

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., and P. Persson (2016). The long-term consequences of teacher discretion in grading of high-stakes tests (No. w22207). National Bureau of Economic Research.

Ebenstein, A., V. Lavy, and S. Roth (2016). The Long Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution. American Economic Journal: Applied Economics, 8(4), 36-65

Feng. A, and G. Graetz (2017). A Question of Degree: The Effects of Degree Class on Labour Market Outcomes. Economics of Education Review, 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. In Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38), ed. M. Cattaneo and J. Escanciano, 29-72. Emerald Group Publishing.

Freier, R., M. Schumann and T. Siedler (2015). The Earnings Returns to Graduating with Honors – Evidence from Law Graduates. Labour Economics. 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

Gregg, P. and E. Tominey (2005). The Wage Scar from Male Youth Unemployment. Labour Economics, 12, 487-509.

Hahn, P. J. Todd and W. van der Klaauw. (2001). Identification and Estimation of Treatments with a Regression Discontinuity Design. Econometrica 69(1): 201-209.

Hupkau, C., S. McNally, J. Ruiz-Valenzuela and G. Ventura (2017). Post-Compulsory Education in England: Choices and Implications', National Institute Economic Review, 240(1): 42-56.

Imbens, G. and T. Lemieux (2008). Regression Discontinuity Designs: A Guide to Practice. Journal of Econometrics, 142, 615-635.

Joint Council for Qualifications (2016). GCSE Full Couse UK by age 2016. Accessed on October 8th 2016 online: www.jcq.org.uk/examination-results/gcses/2016

Kolesár, M. and C. Rothe (2017). Inference in Regression Discontinuity Designs with a Discrete Running Variable. Working Paper.

Kuczera, M., S. Field, and H. Windisch, (2016). Building Skills for All: A Review of England Policy Insights from the Survey of Adult Skills. OECD Skills Studies.

Johnes, G., (2004). Standards and Grade Inflation. In G. Johnes and J. Johnes, International Handbook on the Economics of Education. Edward Elgar.

Lavy, V., and E. Sand. (2015). On the Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teacher Stereotypical Biases. NBER Working Paper. No. 20909.

Lee, D. and T. Lemieux (2010). Regression Discontinuity Designs in Economics, Journal of Economic Literature, 48, 281-355.

McIntosh, S. (2006). Further Analysis of the Returns to Academic and Vocational Qualifications. Oxford Bulletin of Economics and Statistics, 68, 225-51.

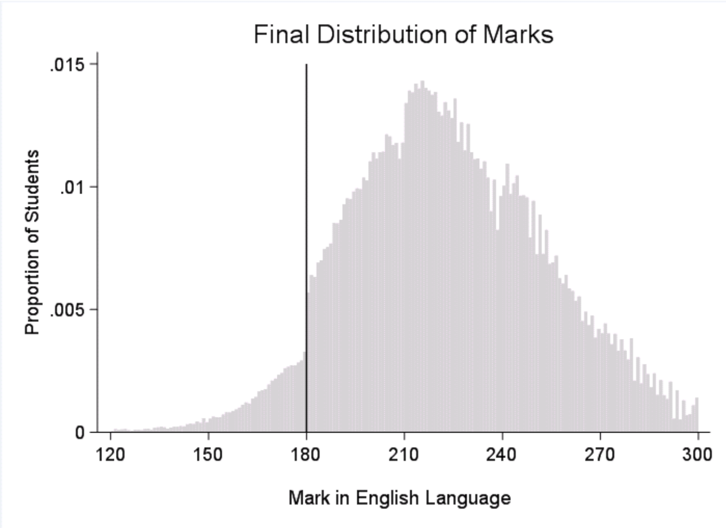
Office of Qualifications and Examinations Regulation (2013). Enquiries About Results for GCSE and A-level: Summer 2013 Exam Series. Statistical Release. Ofqual/13/5357.

Papay, J., R. Murnane and J. Willett (2015). The Impact of Test-Score Labels on Human-Capital Investment Decisions. Journal of Human Resources. 51(2): 357-388.

Terrier. C. (2016). Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement. IZA Discussion Paper No. 10343.

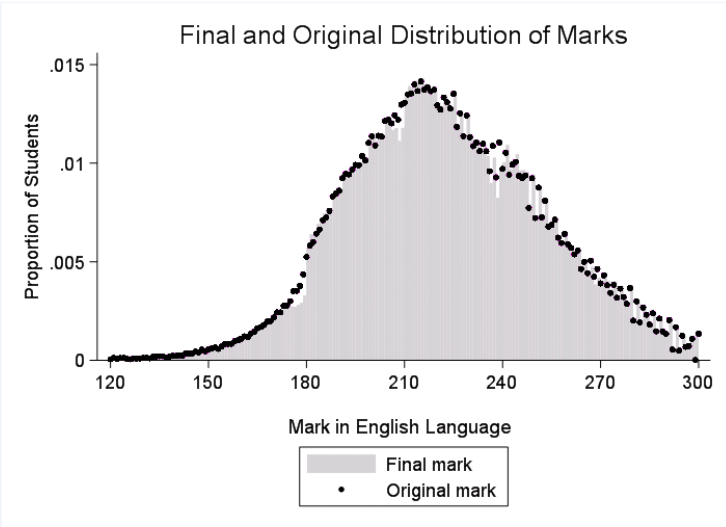
## Figures and Tables

**Figure 1. Final (post-appeal) Distribution of Marks**



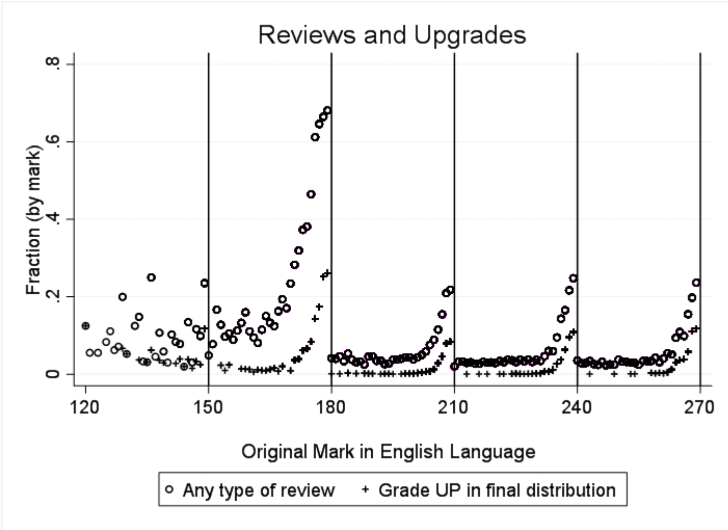

Note. Histogram showing the final (post-appeal) distribution of marks for Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on the data sample construction.

**Figure 2. Final (post-appeal) and Original (pre-appeal) Distribution of Marks**



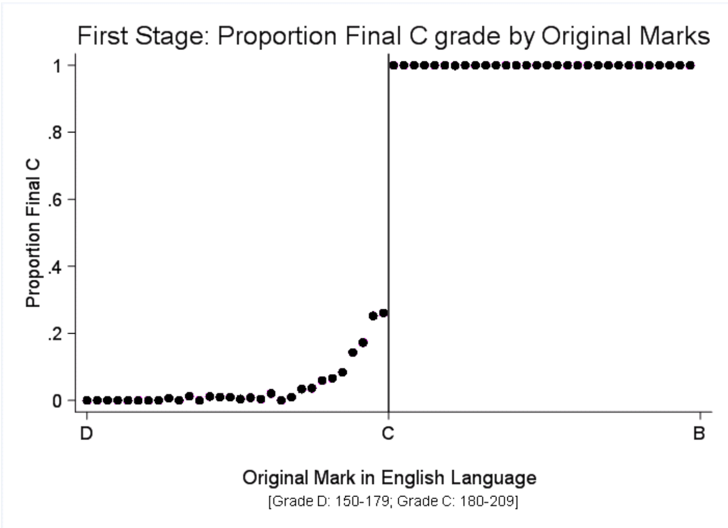

Note. Histogram showing the final (post-appeal) distribution of marks. The dotted line shows the original (pre-appeal) distribution of marks. Both distributions use data for Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on the data sample construction.

**Figure 3. Proportion of Students Asking for a Review and Being Upgraded, by Original Mark**



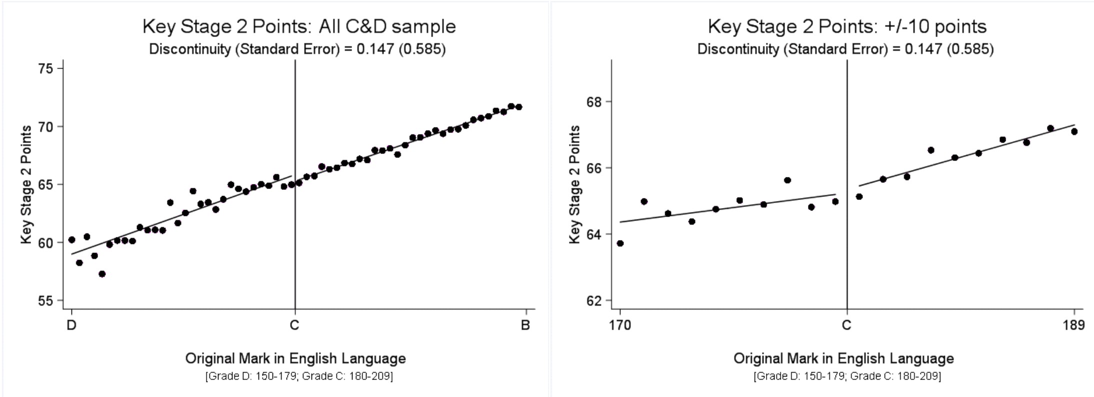

Note. Graph showing the fraction of students (within each original mark), asking for a review and being upgraded; for Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on review data.

**Figure 4. First Stage**



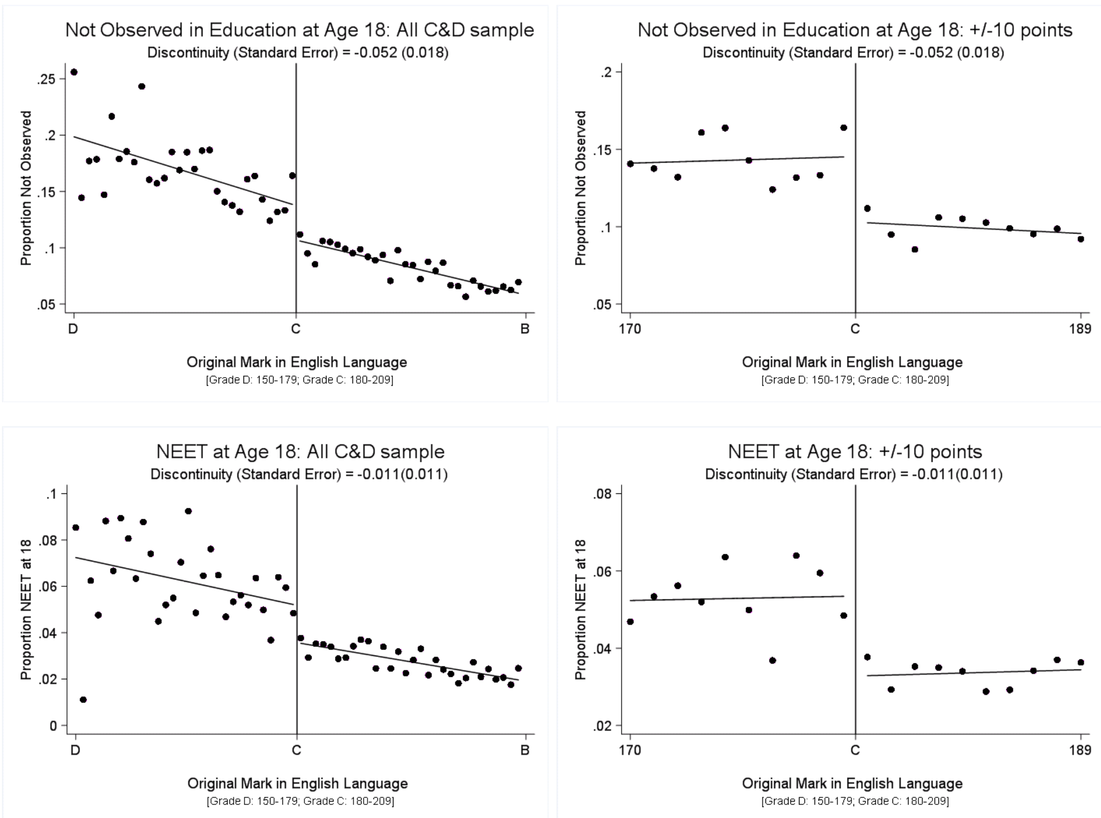

Note. Graph showing the first stage. Each dot represents the fraction of students obtaining a grade C (post-appeal) within each potential original mark (pre-appeal); for Higher Tier students (i.e. those sitting the Higher Tier paper in Unit 1). See Appendix A for further details on the data sample construction.

**Figure 5. Key Stage 2 Points by Forcing variable**



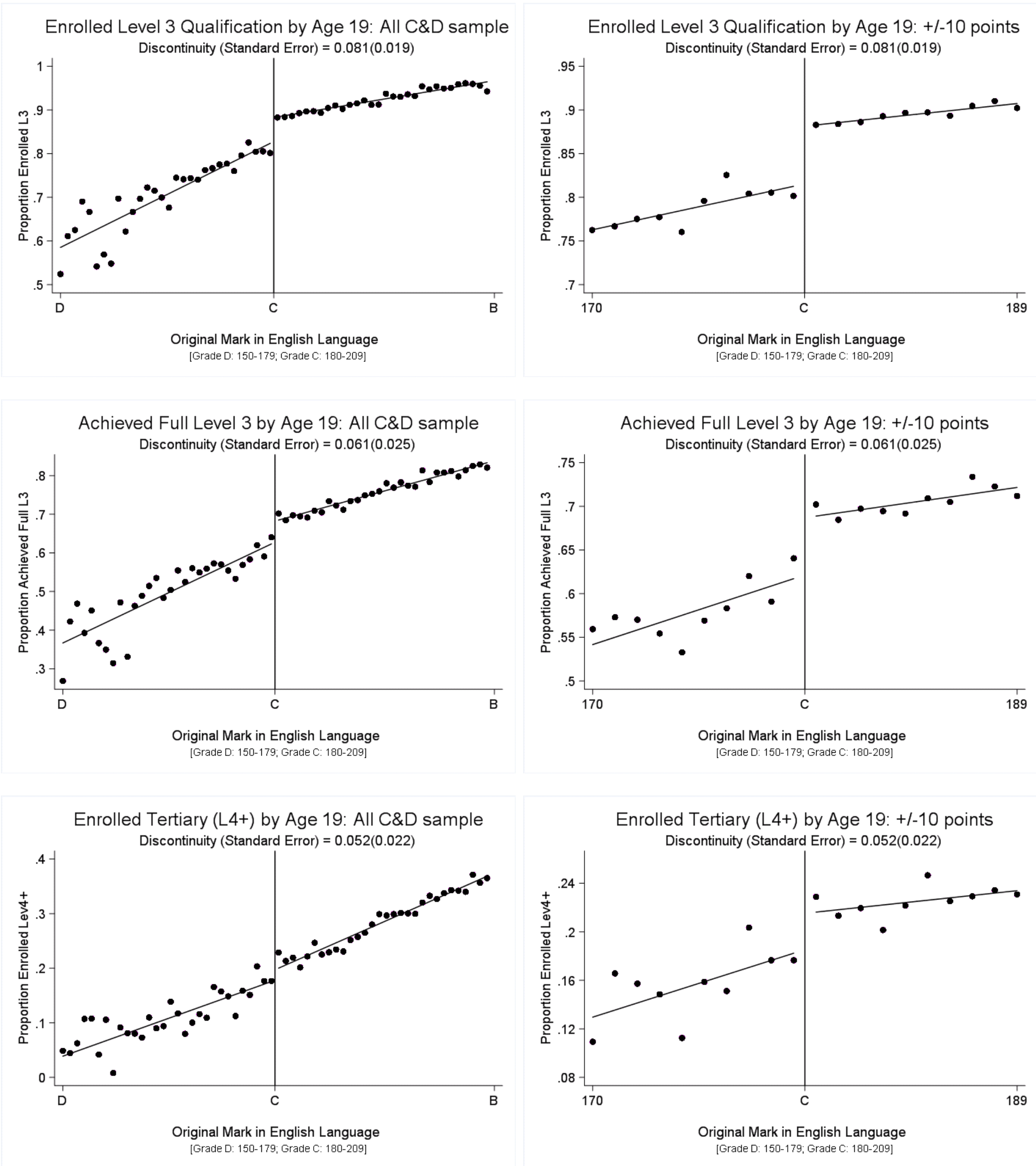

Note. Graph showing the relationship between prior student performance at Key Stage 2 National exams (age 11) and the original (pre-appeal) marks. Each dot represents the average score obtained in the Key Stage 2 examinations within each potential original mark (pre-appeal). Higher Tier students (i.e. 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 the raw differences between the 180 and 179 marks.

**Figure 6. Outcomes at Age 18 by Forcing Variable**



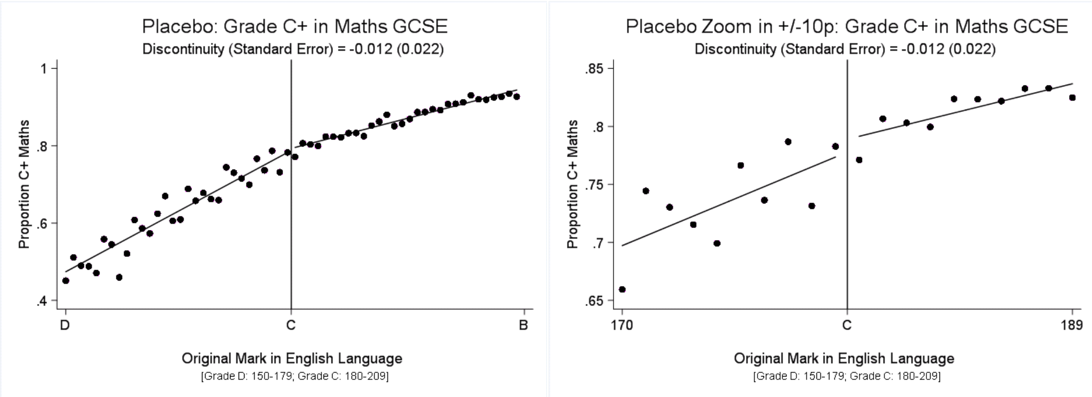

Note. Graph showing the relationship between outcomes at age 18 and the original (pre-appeal) marks. Each dot represents the proportion of students classified as Not Observed in Education/NEET at age 18 within each potential original mark (pre-appeal). Higher Tier students (i.e. those sitting the Higher 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 the raw differences between the 180 and 179 marks.

**Figure 7. Outcomes by Age 19 by Forcing Variable**




Note. Graph showing the relationship between outcomes by age 19 and the original (pre-appeal) marks. Each dot represents the proportion of students classified as achieving each outcome within each potential original mark (pre-appeal). A level 3 qualification is equivalent to an upper secondary qualification. A full-level 3 qualification is equivalent to two A-levels (the usual entry requirement for Higher Education entry in England). A level 4 and above qualification is equivalent to a tertiary education qualification. 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 the raw differences between the 180 and 179 marks.

**Figure 8. Placebo experiment: Probability of getting a C+ in GCSE Maths by Forcing Variable**




Note. Graph showing the relationship between achieving a grade C (or more) in GCSE Mathematics and the original (pre-appeal) marks in GCSE English Language. Each dot represents the proportion of students achieving a grade C or more in GCSE Mathematics within each potential original mark (pre-appeal) in GCSE English Language. 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 the raw differences between the 180 and 179 marks.

**Table 1. Descriptive Statistics**

| | (1) 2013 cohort sitting English Language GCSE | (2) AQA English Language sample | (3) AQA English Language C&D sample - Higher Tier | (4) AQA English Language C&D sample - Foundation Tier |
|---|---|---|---|---|
| Achieved C or above (Level 2) in GCSE English (%) | 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 | 383730 | 189485 | 49231 | 33034 |

Note. 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

**Table 2. Determinants of Asking for a Review and Being Upgraded**

| | (1) | (2) | (3) |
|---|---|---|---|
| Dependent variable: | 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 | 49231 | 49231 | 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 |

Note. 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). 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. More details about the sample and variable construction are given in Appendix A.

**Table 3. Baseline characteristics by Forcing Variable**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Window: All C&D | | Window: +/- 10 points | |
| **A. Dependent variable: Key Stage 2 Points** | | | | |
| Grade C (original) | -0.548** | -0.335 | 0.156 | 0.173 |
| | (0.240) | (0.217) | (0.375) | (0.355) |
| Mean dep variable | 67.275 | | 65.049 | |
| **B. Dependent variable: Whether student is of white ethnicity** | | | | |
| Grade C (original) | 0.001 | 0.004 | 0.013 | 0.022** |
| | (0.009) | (0.007) | (0.014) | (0.011) |
| Mean dep variable | 0.811 | | 0.814 | |
| **C. Dependent variable: Whether the student receives Free-School Meals** | | | | |
| Grade C (original) | -0.012* | -0.012* | -0.012 | -0.014 |
| | (0.007) | (0.007) | (0.012) | (0.011) |
| Mean dep variable | 0.103 | | 0.109 | |
| **D. Dependent variable: Whether English is the language spoken at home** | | | | |
| Grade C (original) | 0.005 | 0.007 | 0.002 | 0.007 |
| | (0.007) | (0.005) | (0.011) | (0.009) |
| Mean dep variable | 0.890 | | 0.891 | |
| **E. Dependent variable: Whether student is a female** | | | | |
| Grade C (original) | 0.009 | 0.013 | -0.009 | -0.002 |
| | (0.012) | (0.011) | (0.019) | (0.018) |
| Mean dep variable | 0.487 | | 0.470 | |
| Sample size | 49231 | 49231 | 14597 | 14597 |
| School Fixed effects | No | Yes | No | Yes |

Note. Each cell shows the results of a parametric sharp RD regression in which a baseline characteristic is regressed against a dummy variable that indicates whether the student originally got a C grade (i.e., pre-appeal) and the forcing variable (original distribution of marks). We let the slope of the forcing variable to vary on each side of the C threshold. The coefficient shown is the one corresponding to whether the student originally obtained a C dummy. Standard errors are clustered at the KS4 school level (i.e., school the student was attending in Year 11).

**Table 4. 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 Level4+ qualification by Age 19 (%) | 36.2 | 38.6 | 26.9 | 16.6 |
| Number of Pupils | 383730 | 189485 | 49231 | 33034 |

Note. 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.

**Table 5. Fuzzy RD estimates: Impact of getting C grade (post-appeal) 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 | 0.092 | | 0.113 | |
| **B. Outcome variable: NEET at age 18** | | | | |
| Grade C (final) | -0.019*** | -0.021*** | -0.028*** | -0.028** |
| | (0.006) | (0.006) | (0.011) | (0.011) |
| Mean dep variable | 0.032 | | 0.040 | |
| **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 | 0.900 | | 0.864 | |
| **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 | 0.732 | | 0.669 | |
| **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 | 0.269 | | 0.205 | |
| **F. Summary Main First Stage: Obtaining a C grade after the appeal process** | | | | |
| Grade C (original) | 0.827*** | 0.828*** | 0.723*** | 0.726*** |
| | (0.008) | (0.008) | (0.013) | (0.013) |
| Sample size | 49231 | 49231 | 14597 | 14597 |
| 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). 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-appeal distribution of marks). Standard errors are clustered at the KS4 school level. School fixed effects are also defined at the KS4 level (i.e. the school the student was attending in Year 11).

**Table 6. Fuzzy RD estimates narrowing the window: Impact of getting a C grade (post-appeal) on different outcomes**

| | (1)<br>+/-5 points | (2)<br>+/-4 points | (3)<br>+/-3 points | (4)<br>+/-2 points | (5)<br>+/-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*** |
| | (0.023) | (0.026) | (0.030) | (0.039) | (0.025) |
| Mean dep variable | 0.116 | 0.115 | 0.116 | 0.123 | 0.136 |
| **B. Outcome variable: NEET at 18** | | | | | |
| Grade C (final) | -0.027** | -0.028* | -0.015 | -0.001 | -0.012 |
| | (0.014) | (0.015) | (0.018) | (0.023) | (0.015) |
| Mean dep variable | 0.041 | 0.041 | 0.043 | 0.042 | 0.043 |
| **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) | (0.040) | (0.026) |
| Mean dep variable | 0.858 | 0.856 | 0.852 | 0.849 | 0.846 |
| **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** |
| | (0.032) | (0.036) | (0.041) | (0.054) | (0.034) |
| Mean dep variable | 0.660 | 0.661 | 0.664 | 0.661 | 0.674 |
| **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) | (0.030) | (0.036) | (0.044) | (0.028) |
| Mean dep variable | 0.201 | 0.200 | 0.206 | 0.202 | 0.205 |
| **F. Summary Main First Stage: Obtaining a C grade after the appeal process** | | | | | |
| Grade C (original) | 0.724*** | 0.720*** | 0.715*** | 0.735*** | 0.737*** |
| | (0.015) | (0.017) | (0.020) | (0.025) | (0.018) |
| 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). 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-appeal distribution of marks). School fixed effects are not included in the regressions (see Table B2 for the results of specifications that include school fixed effects). Standard errors are clustered at the KS4 school level (i.e. the school the student was attending in Year 11).

## Table 7. Robustness checks

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | | **Window: +/- 5 points** | | | **Window: +/- 1 point** | |
| | No controls | Exogenous interaction | Quadratic forcing variable | Reduced Form (probit) | No controls | Reduced Form (probit) |
| **A. Outcome variable: Not Observed in Education at Age 18** | | | | | | |
| Grade C | -0.071*** | -0.084*** | -0.075*** | -0.052*** | -0.071*** | -0.053*** |
| | (0.023) | (0.027) | (0.024) | (0.017) | (0.025) | (0.019) |
| **B. Outcome variable: NEET at 18** | | | | | | |
| Grade C | -0.028** | -0.031* | -0.030** | -0.019* | -0.015 | -0.008 |
| | (0.014) | (0.016) | (0.015) | (0.011) | (0.015) | (0.011) |
| **C. Outcome variable: Enrolled in any Level 3 (upper secondary) qualification by age 19** | | | | | | |
| Grade C | 0.102*** | 0.114*** | 0.108*** | 0.073*** | 0.110*** | 0.081*** |
| | (0.024) | (0.028) | (0.026) | (0.018) | (0.027) | (0.019) |
| **D. Outcome variable: Achieved a Full Level 3 qualification by age 19** | | | | | | |
| Grade C | 0.089*** | 0.073** | 0.082** | 0.066*** | 0.083** | 0.063** |
| | (0.032) | (0.036) | (0.034) | (0.023) | (0.034) | (0.025) |
| **E. Outcome variable: Enrolled in tertiary education (Level 4 or above) by age 19** | | | | | | |
| Grade C | 0.054** | 0.049* | 0.053* | 0.040** | 0.071** | 0.053*** |
| | (0.027) | (0.029) | (0.028) | (0.018) | (0.029) | (0.020) |
| **F. Summary Main First Stage: Obtaining a C grade after the appeal process** | | | | | | |
| Grade C (original) | 0.724*** | 0.679*** | 0.701*** | -- | 0.739*** | -- |
| | (0.015) | (0.019) | (0.017) | -- | (0.018) | -- |
| Sample size | | 7082 | | | 1409 | |
| Number of schools | | 1258 | | | 742 | |

Note. Panels A to E, columns 1, 2, 3 and 5: each cell shows the main coefficient of interest for the C dummy variable (endogenous variable in the second stage). Panels A to E, columns 4 and 6: each cell shows marginal effects from probit regressions of the main coefficient of interest, which in this case is the impact of the original C grade on the different outcomes. Panel F: each cell shows the main coefficient of interest in the first stage. Standard errors are clustered at the KS4 school level (i.e., school the student was attending in Year 11). All regressions include the set of controls described in the Data Appendix (except the ones in columns 1 and 5). The window restriction is based on the forcing variable (i.e., excluding +/- X points away from the C threshold as given by the pre-appeal distribution of marks). School fixed effects are not included in the regressions. Column 2 includes the interaction between the instrument and the forcing variable as an extra control variable (both in the first and second stage). All regressions in the +/-5 points window control for the forcing variable in a linear way (except in column 3). The regressions that are +/-1 point away do not control for the forcing variable because in this narrower sample the forcing variable is the same as the instrument.

**Table 8. 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 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) | (0.028) | (0.028) |
| Mean dep variable | 0.925 | | 0.885 | | 0.873 | |
| **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* |
| | (0.011) | (0.010) | (0.020) | (0.018) | (0.031) | (0.031) |
| Mean dep variable | 0.300 | | 0.235 | | 0.228 | |
| **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 |
| | (0.013) | (0.012) | (0.024) | (0.023) | (0.035) | (0.037) |
| Mean dep variable | 0.484 | | 0.391 | | 0.380 | |
| **D. Outcome variable: Enrolled in any A/AS/Applied GCE at 17** | | | | | | |
| Grade C (final) | 0.103*** | 0.114*** | 0.106*** | 0.111*** | 0.151*** | 0.150*** |
| | (0.013) | (0.012) | (0.023) | (0.023) | (0.035) | (0.037) |
| Mean dep variable | 0.524 | | 0.401 | | 0.384 | |
| Sample size | 49231 | 49231 | 14597 | 14597 | 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); (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 the Data Appendix. The window restriction is based on the forcing variable (i.e. excluding +/- X points away from the C threshold as given by the pre-appeal distribution of marks). Standard errors are clustered at the KS4 school level. School fixed effects are also defined at the KS4 level (i.e.the school the student was attending in Year 11).

**Table 9. Quality of the 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 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 |
| | (0.781) | (0.752) | (1.414) | (1.396) | (2.160) | (2.279) |
| Mean dep variable | 65.75 | | 59.724 | | 58.903 | |
| **B. Achieved a C grade in GCSE English** | | | | | | |
| Grade C (final) | 2.874*** | 3.001*** | 2.700** | 2.470** | 3.253* | 2.353 |
| | (0.688) | (0.666) | (1.253) | (1.241) | (1.905) | (2.015) |
| Mean dep variable | 73.014 | | 67.778 | | 67.077 | |
| **C. Achieved a C grade in GCSE Maths** | | | | | | |
| Grade C (final) | 2.141*** | 2.691*** | 2.053* | 2.144* | 2.864 | 2.323 |
| | (0.638) | (0.606) | (1.157) | (1.126) | (1.773) | (1.846) |
| Mean dep variable | 75.731 | | 71.17 | | 70.558 | |
| Sample size | 45526 | | 13187 | | 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). 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-appeal distribution of marks). Standard errors are clustered at the KS4 school level. School fixed effects are also defined at the KS4 level (i.e. the school the student attended in Year 11).

**Table 10. Wage differentials For level 3 qualifications (versus level 2 qualifications)**

| *Dependent variable: Log gross hourly wages* | (1) | (2) | (3) |
|---|---|---|---|
| Maximum attainment: level 3 qualification | 0.110*** | 0.095*** | 0.062*** |
| | (0.006) | (0.006) | (0.005) |
| Sample size | 23815 | 23752 | 23743 |
| *Controls:* | | | |
| Basic | ✓ | ✓ | ✓ |
| 2-digit Industry Sector dummies | | ✓ | ✓ |
| 1-digit Occupation dummies | | | ✓ |

Note: Earning differentials associated to obtaining level 3 qualifications (as the maximum level of qualification obtained), compared to obtaining level 2 qualifications. The data comes from the Quarterly Labour Force Survey (2015 to third quarter 2017). The sample is restricted to individuals whose maximum level of qualification is either 2 or 3; are aged between 21 and 65, and are respondents in the first wave in the sample. Small differences in the number of observations between the three specifications are due to missing values in the control variables. The dependent variable in all specifications is log gross hourly wages (i.e. log[gross weekly pay in main job/(basic usual hours+usual hours of paid overtime)]). Basic controls include: potential experience (calculated as age - age the individual left full time education), potential experience squared, a dummy for full-time employment, gender, regional dummies, white ethnicity dummy and year and month of interview dummies. Weighted regressions using the individual person weight.

## Appendix

### A. Data Appendix

#### A1. Key Stage 4 Results, Assessment, Qualifications Alliance (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.[37]

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 can be seen in Table A1, the majority of entries (72%) correspond to GCSE English Language entries. The biggest awarding body for both GCSE qualifications is 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. 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).

[37] 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).

**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 (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.[38] 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

[38] We also obtained access to information on who asked for a review on the 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).

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 is a rather common practice for other subjects (like Mathematics), this does not seem to happen very often for GCSE English Language qualifications. However, our first sample selection criteria follows the advice that we were given by the Department for Education (DfE) to deal with this issue. This consists in: (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 to 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 sample 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. 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 appeals process, i.e. we exclude students that suffer big jumps in their marks after the appeals process, since this might be due to measurement error). There are 49231 students fulfilling the underlined criteria and that will therefore constitute the main sample in our analysis.

**A2. 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).[39] Table A3 shows how raw marks for the three different units are translated into raw 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

[39] 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.

previous exam series.[40] 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 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 |

Notes: 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 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.

[40] 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

## A3. Other data

### *a. 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 use 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*).

### *b. 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].

### *c. 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.

### *d. 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.

### e. 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 the student is enrolled in, the second file has information at the learner level. 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.

### f. 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.

### g. Longitudinal Education Outcomes Dataset (LEO: P14 and Self-assessment)

We use information about annual earnings in tax year 2015 (i.e., from 6$^{th}$ April 2014 to 5$^{th}$ 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 PMR indicator for each of the records.

### A3. Construction of outcomes

***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). In particular, 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.

***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).

***Entry to a higher-level 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 a higher-level 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*.

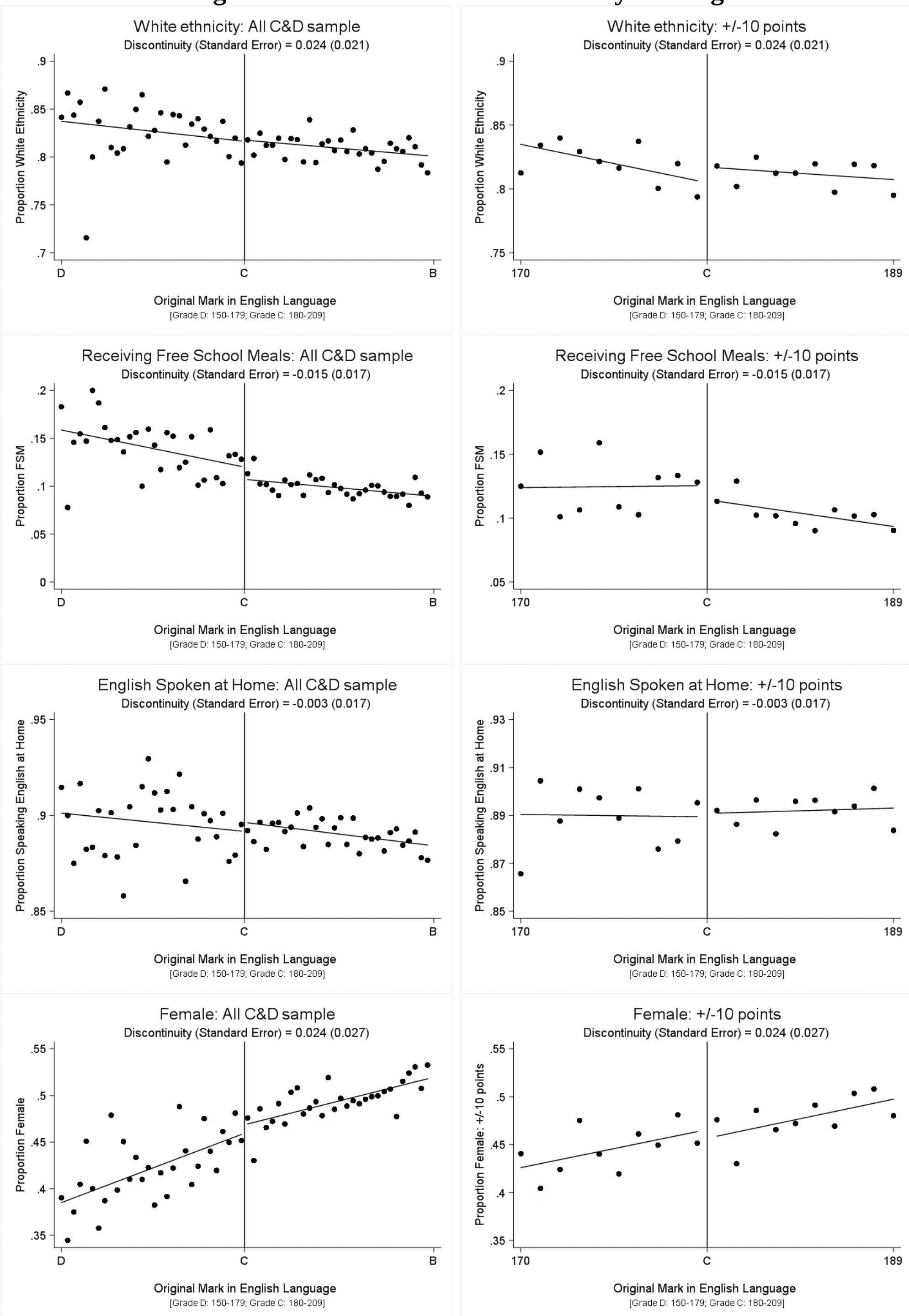
***Achieved a Full-Level 3 qualification (i.e. upper-secondary) by age 19:*** A full level 3 qualification is obtained when the student achieves at least two A-level (or equivalent qualifications) passes. In particular, we classify an individual as having fulfilled a full-level 3 qualification if at 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[41]; (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.

***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.

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

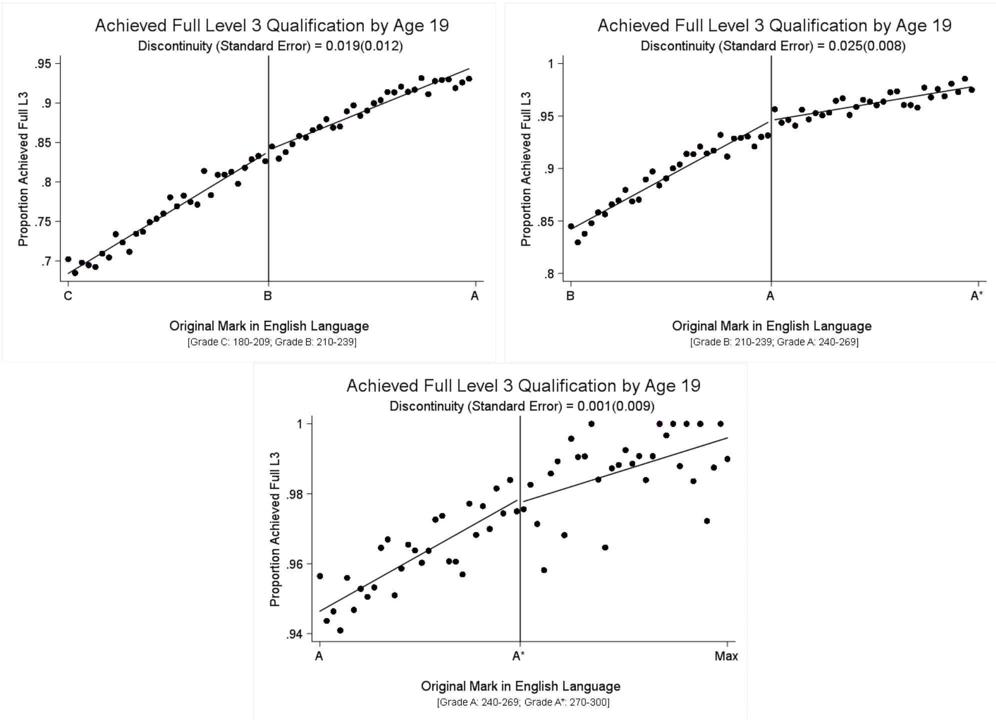
# B. Additional Figures and Tables

**Figure B1. Baseline Characteristics by Forcing variable**




Note. Graph showing the relationship between baseline characteristics and the original (pre-appeal) marks. Each dot represents the proportion of students that are white (or receive free school meals, speak english at home or are females, respectively, depending on the graph), within each potential original mark (pre-appeal). Sample of students defined in Appendix A. Linear regression lines are fitted separately on each side of the C threshold. The discontinuity and standard error shown correspond to the raw differences between the 180 and 179 marks.

**Figure B2. Achieving a Full Level 3 qualification by Age 19 at other grade thresholds**




Note. Graph showing the relationship between achieving a Level 3 (i.e., upper-secondary) by age 19 and the original (pre-appeal) marks. Each dot represents the proportion of students classified as achieving the outcome within each potential original mark (pre-appeal). 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 C threshold. The discontinuity and standard error shown correspond to the raw differences between the 180 and 179 marks.

**Table B1. Baseline characteristics by Forcing Variable (Narrower windows)**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Window: +/- 5 points | | Window: +/- 1 point | |
| **A. Dependent variable: Key Stage 2 Points** | | | | |
| Grade C (original) | -0.097 | -0.006 | 0.145 | -0.126 |
| | (0.517) | (0.537) | (0.581) | (0.810) |
| Mean dep variable | 64.664 | | 64.048 | |
| **B. Dependent variable: Whether student is of white ethnicity** | | | | |
| Grade C (original) | 0.014 | -0.000 | 0.024 | 0.001 |
| | (0.019) | (0.016) | (0.022) | (0.025) |
| Mean dep variable | 0.813 | | 0.807 | |
| **C. Dependent variable: Whether the student receives Free-School Meals** | | | | |
| Grade C (original) | -0.007 | 0.003 | -0.015 | 0.004 |
| | (0.016) | (0.016) | (0.018) | (0.023) |
| Mean dep variable | 0.113 | | 0.120 | |
| **D. Dependent variable: Whether English is the language spoken at home** | | | | |
| Grade C (original) | 0.002 | -0.004 | -0.003 | 0.003 |
| | (0.015) | (0.013) | (0.016) | (0.019) |
| Mean dep variable | 0.890 | | 0.894 | |
| **E. Dependent variable: Whether student is a female** | | | | |
| Grade C (original) | -0.011 | -0.011 | 0.024 | 0.037 |
| | (0.024) | (0.025) | (0.026) | (0.035) |
| Mean dep variable | 0.461 | | 0.465 | |
| Sample size | 7082 | 7082 | 1409 | 1409 |
| School Fixed effects | No | Yes | No | Yes |

Note. Window:+/-5 points: Each cell shows the results of a parametric sharp RD regression in which the baseline covariate is regressed against a dummy variable that indicates whether the student originally got a C grade (i.e., pre-appeal) and the forcing variable (original distribution of marks). The forcing variable is not interacted with the C dummy, but results are very similar if we let the slope of the forcing variable vary on each side of the C threshold. Window: +/-1 point: In this case the forcing variable collapses to the C dummy, so it is obviously not included in the regressions. In both windows the coefficient shown is the one corresponding to whether the student originally obtained a C dummy. Standard errors are clustered at the KS4 school level.

**Table B2. Fuzzy RD estimates narrowing the window: Impact of getting a C grade (post-appeal) on different outcomes (includes KS4 school fixed effects)**

| | (1)<br>+/-5 points | (2)<br>+/-4 points | (3)<br>+/-3 points | (4)<br>+/-2 points | (5)<br>+/-1 points |
|---|---|---|---|---|---|
| **A. Outcome variable: Not Observed in Education at Age 18** | | | | | |
| Grade C (final) | -0.058** | -0.072*** | -0.064* | -0.067 | -0.026 |
| | (0.024) | (0.028) | (0.034) | (0.047) | (0.034) |
| Mean dep variable | 0.116 | 0.115 | 0.116 | 0.123 | 0.136 |
| **B. Outcome variable: NEET at 18** | | | | | |
| Grade C (final) | -0.026* | -0.033** | -0.024 | -0.002 | 0.020 |
| | (0.014) | (0.017) | (0.020) | (0.026) | (0.019) |
| Mean dep variable | 0.041 | 0.041 | 0.043 | 0.042 | 0.043 |
| **C. Outcome variable: Enrolled in any Level 3 (upper secondary) qualification by age 19** | | | | | |
| Grade C (final) | 0.087*** | 0.094*** | 0.111*** | 0.109** | 0.103*** |
| | (0.025) | (0.030) | (0.038) | (0.049) | (0.037) |
| Mean dep variable | 0.858 | 0.856 | 0.852 | 0.849 | 0.846 |
| **D. Outcome variable: Achieved a Full Level 3 qualification by age 19** | | | | | |
| Grade C (final) | 0.067** | 0.053 | 0.067 | 0.042 | 0.037 |
| | (0.034) | (0.039) | (0.048) | (0.066) | (0.047) |
| Mean dep variable | 0.660 | 0.661 | 0.664 | 0.661 | 0.674 |
| **E. Outcome variable: Enrolled in tertiary education (Level 4 or above) by age 19** | | | | | |
| Grade C (final) | 0.037 | 0.042 | 0.054 | 0.053 | 0.024 |
| | (0.028) | (0.033) | (0.040) | (0.052) | (0.038) |
| Mean dep variable | 0.201 | 0.200 | 0.206 | 0.202 | 0.205 |
| **F. Summary Main First Stage: Obtaining a C grade after the appeal process** | | | | | |
| Grade C (original) | 0.731*** | 0.728*** | 0.730*** | 0.749*** | 0.775*** |
| | (0.016) | (0.018) | (0.021) | (0.028) | (0.023) |
| Sample size | 7082 | 5671 | 4212 | 2817 | 1409 |
| Number of schools | 1258 | 1201 | 1110 | 993 | 742 |
| Proportion schools with only 1 student (%) | 15.9 | 18.8 | 25.0 | 31.8 | 50.4 |

Note. Panels A to E: each cell shows the main coefficient of interest for the C dummy variable (endogenous variable in the second stage). 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 and KS4 school fixed effects. The window restriction is based on the forcing variable (i.e., excluding +/- X points away from the C threshold as given by the pre-appeal distribution of marks). Standard errors are clustered at the KS4 school level (i.e., school the student was attending in Year 11).