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

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

Preliminary draft

This version: 12 September 2017

Abstract

In many countries, there are important thresholds in examinations that act as a gateway to higher levels of education and/or good employment prospects. This paper examines the consequences of just failing a key national examination in English taken at the end of compulsory schooling in England. It uses unique pupil level 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 exam; 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 evidence seminar series at the Department for Education (DfE). We would also like to thank participants at the 2017 Royal Economic Society Conference, the 2017 Meetings of the Society of Labor Economics, the 2017 RIDGE Workshop in Public Economics and the 2017 conference of the Centre for Economic Performance. All remaining errors are our own.

1. Introduction

Getting above or failing to reach thresholds in exams or tests 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 qualifications is been deemed 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 such example. This is in part because achievement of good literacy and numeracy skills is recognised as an important output of the education system, with England consistently underperforming in this regard.[1] It is also because achieving a 'good pass' 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]

At the same time, exam thresholds have 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 teachers. In addition, there is also the concern that disadvantaged students can directly

[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 the only 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 England and maths.
[3] However, the pass-rate for those students re-taking the GCSE exam is less than 30 percent.
[4] See the discussion by Johnes (2004) for example.

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 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 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]

[5] 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). We do not have comparable information for maths.

[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 also forms part of the school-level indicators that are in published School Performance Tables. It is given much emphasis on articles in newspapers and on the Internet about GCSE results and what to expect afterwards.

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 there to be significant test score manipulation around known grade thresholds in the national mathematics tests taken by ninth graders in Sweden. This, they conclude, actually 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 both high school graduation and the probability of taking advanced coursework (though also lowered the likelihood of college enrolment). To show this, they exploit the reforms that were introduced by the education authorities to deliberately get rid of the test score manipulation that was observed taking place.

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 Borchan 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 Angrist et al. (2015) 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 teacher or school 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 that 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 worries about manipulation bias. First, one key feature of English examinations is a right to appeal, and whilst the administrative data we use 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 what looks like manipulation in the data is actually due to the regrading process through appeals or not. Our paper is unique in having the 'pre-manipulation' and 'post-manipulation' distribution 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 some 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 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 transitory shocks (or bad luck) in the context of

[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.

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. 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. The fact that students who narrowly fail to get a grade C in English have a moderate to high risk of negative consequences reflects badly on the mechanisms within the education system to give support to these students and on the availability of suitable education programmes for students of weaker ability. Indeed, Hupkau et al. (2017) show that the probability of progression from lower level to higher level courses

[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, 2013; Freier et al. 2015); and how test score labels affect human capital investment decisions (Papay et al. 2015).

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 and explain our data (Section 2). Then we discuss the empirical distribution of pre-appeal and post-appeal marks and the methodological approach (Section 3), before presenting our results (Section 4). We conclude in Section 5.

2. Institutional Background and Data

2.1. The English Education System

In England, the GCSE (General Certificate of Secondary Education) examinations mark the end of compulsory education, when students are aged 16 (as grade repetition very rarely occurs). The typical student takes 8-10 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

[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.

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 or not 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.

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. Some of these are teacher assessed (and moderated by the exam board) and some are based on a standardised exam which is corrected (anonymously) by an external examiner. Exams take place after the coursework assessment (at the end of the school year). In the year of relevance to our study (2013), 40% of the overall marks were accounted for by the standardised exam.[13] Crucially, 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.[14] Furthermore, grade boundaries are not decided in advance of the exam. This is decided by an external committee that engages in a

[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.

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

[14] 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.

process of inspecting papers (e.g. comparing them to previous years) and statistical analysis.[15] Thus, it is not possible for teachers to manipulate coursework assessments such that the marginal student just crosses the threshold for a Grade C.[16]

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).

2.2. Data

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. In its compulsory phases, the English education system is organised into four Key Stages (KS), where there are external assessment 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 exam). 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.

[15] 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

[16] 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.

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.[17] This exam board accounts for over half of all exam entries in GCSE English.[18] The characteristics of entrants are shown in Table 1 (column 2). Compared to the cohort as a whole (column 1), they are more likely to have higher prior attainment and less likely to be disadvantaged. To ensure we are considering only those students taking the same assessment, we focus on the form of English exam that is undertaken by 75% 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 also if we consider those taking the lower tier (English language) exam paper.[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. 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

[17] 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.

[18] 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

[19] 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.

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.[20] Table 1 shows summary statistics for the whole cohort (column 1), the AQA English language sample (column 2), and the subsample of students that are main interest here (column 3).

3. Empirical Distribution and Methodology

3.1. Empirical Distribution of Marks

We have both the final distribution of marks and the original distribution of marks (i.e. before re-marking is requested) for the same students. 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

[20] In England, this implies starting an undergraduate or foundation degree, or enrolling in any sort of high level (level 4 and above) vocational qualification.

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 original distribution of marks is approximately normal.[22]

Figure 3 shows the probability of requesting a re-mark 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, the probability of requesting a re-mark is close to 60 per cent. In contrast, the probability only rises to about 20 per cent near the thresholds for Grades, B, A and A*. This is illustrative of how important getting a Grade C is within the English education system. The Figure also shows the probability of actually getting 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 re-mark and the conditional probability of getting upgraded in Table 2. We use only those students whose original marks were in the range of a C or a D grade and we always control for the students' original mark. We regress whether or not a request 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

[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).
[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.

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 a remark is requested for a student.[23] 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 close to 10 per cent. Re-marking of scripts is less likely to be requested for females (by about 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 prior attainment (and are close to zero, though are 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. In fact, if we regress the final distribution of marks on school fixed effects and plot the residuals, we get a distribution that is very similar to the original distribution (i.e. the bunching close to the C threshold disappears). The probability of being upgraded to a C grade (which happens for 11% of students for whom a re-mark is requested in our sample) is not related to any demographic characteristic of students or to their prior attainment. This is not surprising given that examiners doing the re-marking know nothing about the students and are given different questions to re-examine (i.e. the same person does not re-evaluate the whole script).[24]

[23] The marginal effects from a Probit model give identical results.

[24] 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).

3.2. 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 marginally obtaining a C grade on later outcomes, we make use of the fact that we have the original (pre-appeal) score 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 score is a very strong predictor of whether grade C is finally obtained (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 – although including them does not alter our results). 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 half-fuzzy, half-sharp regression discontinuity design). 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, we use fuzzy regression discontinuity methods (Angrist and Lavy, 1999; Hahn et al. 2001) 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 as the forcing variable. In some specifications, we control for flexible functions of the final score, whereas in other specifications, we let the slope of the treatment variable vary on either side of the C cut-off. We also estimate regressions where we limit the sample to individuals that were 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 their prior attainment) varies discontinuously at this threshold and show that this can be ruled out.

More formally, we estimate the following equations:

$$Y_{is} = \beta_0 + \beta_1 CF_{is} + f(M_{is}) + \beta_2 X_{is} + \mu_s + \epsilon_{is} \quad (1)$$

$$CF_{is} = \alpha_0 + \alpha_1 CO_{is} + g(M_{is}) + \alpha_2 X_{is} + \mu_s + \omega_{is} \quad (2)$$

where outcome *Y* of individual *i* in school *s* is related to a dummy variable indicating whether or not 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).[25] μ denotes a school fixed

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

effect. *f(m)* and *g(M)* are functions that capture the underlying relationship between the so-called running or forcing variable (the original distribution of marks) and the treatment and outcome variable, respectively. ϵ_{is} and ω_{is} are error terms and we cluster at the level of the school.[26]

We estimate these regressions in two ways. First, we use the global polynomial approach in which the full range of scores between Grades C and D is used and *f* and *g* are specified in various different ways to approximate the relationship between the original mark obtained by the student and the dependent variable. In some specifications, we allow *f* and *g* to take a different shape on either side of the grade C threshold.

We also estimate linear regressions over a small range of the data ('local regressions') close to the C threshold (original marks ranging from +/- 5 to +/- 1). 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. We examine this assumption below in detail.

4. Results

4.1. Validity of Approach

As discussed in Section 2.1 and 3.1, 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). Figure 5 shows a series of graphs that plot pre-determined characteristics for students who obtain a C or a D grade according to

[26] Clustering standard errors at the level of the forcing variable does not alter standard errors significantly.

their original marks. A line is fitted to the data before and after the threshold for Grade C. One of these variables is the student's test score at age 11. This test is national and takes place at the end of primary school. It is high stakes for schools because it forms the basis of the School Performance Tables for primary schools. There is no discontinuity around the Grade C threshold in GCSE English (which can be seen visually and also 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, whether he/she speaks English as a first language.

In Table 3, we report regression estimates where each baseline characteristic is regressed against whether the student obtains a C grade (pre-appeal), controlling for the original (pre-appeal) mark and school fixed effects. Columns (1) and (2) show regressions estimated for a subsample of students very close to the (original) Grade C threshold (+/- 5 marks in column 1; +/- 1 mark in column 2). Columns (3)-(5) show results for the full range of marks between Grades C and D, controlling for a polynomial function of the original marks which is different in each column: linear (1); quadratic (2); cubic (3) and quartic (4). In almost all cases, the relationship between the baseline characteristic and whether or not the student obtains a C grade is small and not statistically significant. Hence, it is plausible to conclude that the marginal student who passes the (pre-appeal) threshold is quasi-randomly assigned.

4.2. A Graphical Illustration

Before showing the results of our regressions, we plot our outcome variables in Figures 6 to 10 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). This is for all students who obtained marks (pre-appeal) within the range of a C and a D grade (i.e. marks between 150 and 210), where the threshold is at 180 marks. 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) as well as a

lower probability of being observed as 'not in education, training or employment' (NEET) at age 18 (Figure 7). Figure 8 shows that students who just pass the original C cut-off have a higher probability of accessing or achieving a higher qualification by age 19 (Figure 9), and starting tertiary education by age 19 (Figure 10). 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 Figures A1 to A6 in the Appendix) or indeed at other points of the distribution.[27]

4.3. Regression Estimates: Global Polynomial Approach

In Table 4, we show regressions estimated for three different specifications for the full sample of interest (columns 1-3) and for the subsample within +/- 10 points of the Grade C threshold (columns 4-6). There are five panels for the different outcome variables. 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), they show the estimated coefficient β_1, the (second stage) IV estimate. The sixth panel shows estimated coefficients for the first stage (i.e. α_1 and interaction terms where relevant), which is always very large and statistically significant.

The first specification controls for the linear forcing variable (columns 1 and 4). The second and third specifications allow the distribution of final marks (or the forcing variable) to vary at either side of the threshold. Specifically, the second specification involves controlling for the "endogenous interaction" between the post-appeal threshold dummy and the forcing variable (i.e. the original mark). This is instrumented using the interaction between the pre-appeal threshold dummy and the original mark (columns 2 and 5). The third specification involves controlling for the "exogenous interaction" between the pre-appeal threshold dummy and the original mark (columns 3 and 6). In Appendix Table 1 we show the results when

[27] We have followed Imbens and Lemieux (2008) tests for jumps at non-discontinuity points (i.e. at the median of each side of the C threshold) and can easily reject the hypothesis of discontinuities at these other points.

controlling for different functional forms of the forcing variable (quadratic, cubic and quartic). Results are very similar across these different specifications and are generally statistically significant (apart from some of the specifications where commencing tertiary education is the dependent variable).

The regressions all suggest a sizeable effect of marginally achieving (or failing to achieve) a C grade. In this 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 about 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 C and Grade D, 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 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 10 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' qualification within 3 years (whereas the expectation would be that most people would achieve this within 2 years of the end of compulsory education).

The final row 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 1.5 to 4 percentage points in a context where about 27 percent of this sample have started tertiary education by this age (21 percent for those within 10 marks of the C threshold).

4.3. Regression Estimates: Local Linear Approach With Varying Windows

In Table 5, we show results for the local linear model of these regressions (discussed above), where they are estimated only on the subsample of students who obtain a very narrow range of marks in the original (pre-appeal) distribution. Again, there are five panels for the different outcome variables (the sixth showing results from first stage regressions) and ten columns, each of which shows the estimated effect of achieving Grade C on the outcome of interest. Columns (1) to (5) show results including fixed effects for the secondary school attended at the time of the exams and columns (6) to (10) show results without including school fixed effects. Columns (1) and (6) show estimates of regressions for the subsample of students within +/- 5 marks from the Grade C threshold. Columns (2) and (7) replicate the regressions for the sample of students within +/- 4 marks of the threshold. Then the sample is gradually narrowed to +/- 3 marks (columns 3 and 8), +/- 2 marks (columns 4 and 9) and +/- 1 mark (columns 5 and 10). The reason we estimate the regressions without fixed effects (as well as with 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). The Table 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 only have one such student. This rises to half of all schools in the sample of students within +/- 1 mark of the threshold.

The results are consistent with those shown for the larger sample and are qualitatively similar (especially when controlling for school fixed effects). They are generally statistically significant, except for when school fixed effects are included with smaller subsamples of students. The variable denoting enrolment in tertiary education is never statistically significant when school fixed effects are included but point estimates are always positive and slightly higher than for the global regressions reported in Table 4. The point estimates are usually higher when school fixed effects are not included and are highly consistent across specifications with a different number of students (i.e. columns 6 to 10). 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.[28]

5. 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 threshold for grade C. This results in significant bunching of students near this threshold in the (post-appeal)

[28] Donut estimates (see Barreca, Guldi, Lindo and Waddell, 2011) excluding observations that are very close to the C threshold produce very similar results. In addition, 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 in GCSE Mathematics. We run this placebo exercise and confirm this. The results of these two additional robustness checks are available upon request.

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. The mechanisms are likely to be related to the education system itself (i.e. lack of opportunity for those not meeting the C threshold and insufficient support for those who just miss it) and potentially the psychological effect that perceived failure can have on students' self-evaluation of their 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.

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).

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 exam taken at the end of compulsory schooling.

References

Altonji J., E. Blom E. 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.R. Sass. (2016). Do the Cheated Ever Prosper? The Long-Run Effects of Test-Score Manipulation by Teachers on Student Outcomes. CALDER Working Paper No. 155. National Center for Analysis of Longitudinal Data in Education Research. US.

Avery, C., O. Gurantz, M. Hurwitz, and J. Smith. (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. (2016). In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno. American Economic Journal: Applied Economics.

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

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

Calonico, S., M. Cattaneo and R. Titiunik (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. Econometrica, 82, 2295-2326.

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. (with Robert Cassen and Anna Vignoles). Routledge.

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

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

Ebenstein, A., 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. (In Press)

Feng. A, and G. Graetz (2013). A Question of Degree: The Effects of Degree Class on Labour Market Outcomes. Centre for Economic Performance. Discussion Paper No. 1221. London School of Economics.

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.

Imbens, G. and K. Kalyanaraman (2011). Optimal bandwidth choice for the regression discontinuity estimator. The Review of Economic Studies, 79, 933-959.

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

Kuczera, M., S. Field, and H.C. 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.

Office of Qualifications and Examinations Regulation (2013). GCSE and A level Enquiries about Results: Subject level analyses. Summer 2015 exam series. Ofqual/16/6007.

Papay, J. P., R.J. Murnane and J.B. 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 Distribution of Marks (Higher Tier)

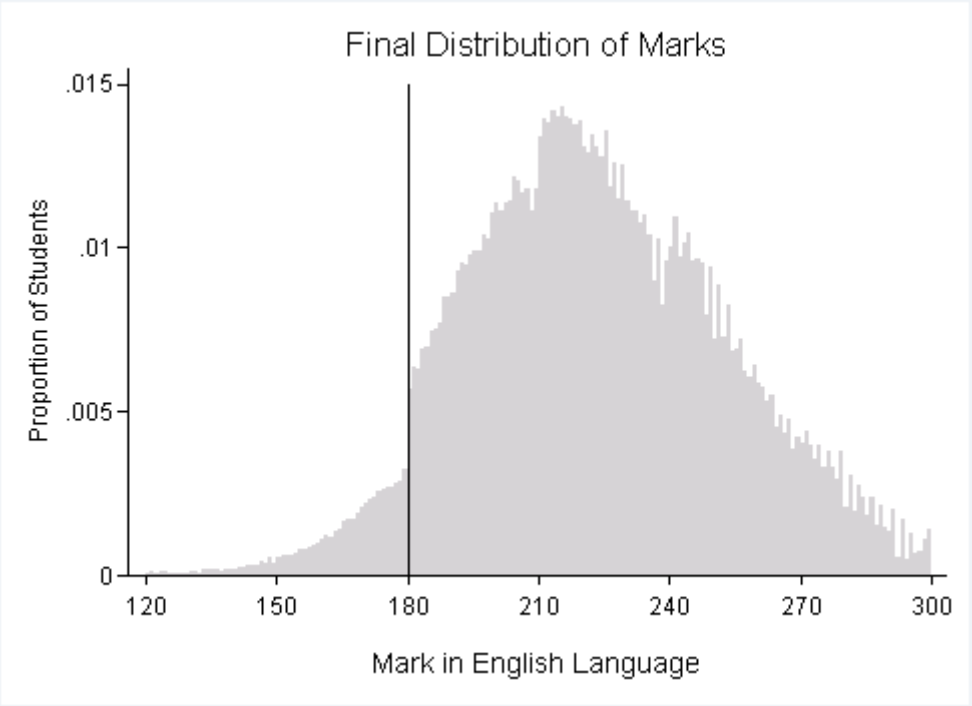


Figure 2. Final and Original Distribution of Marks (Higher Tier)

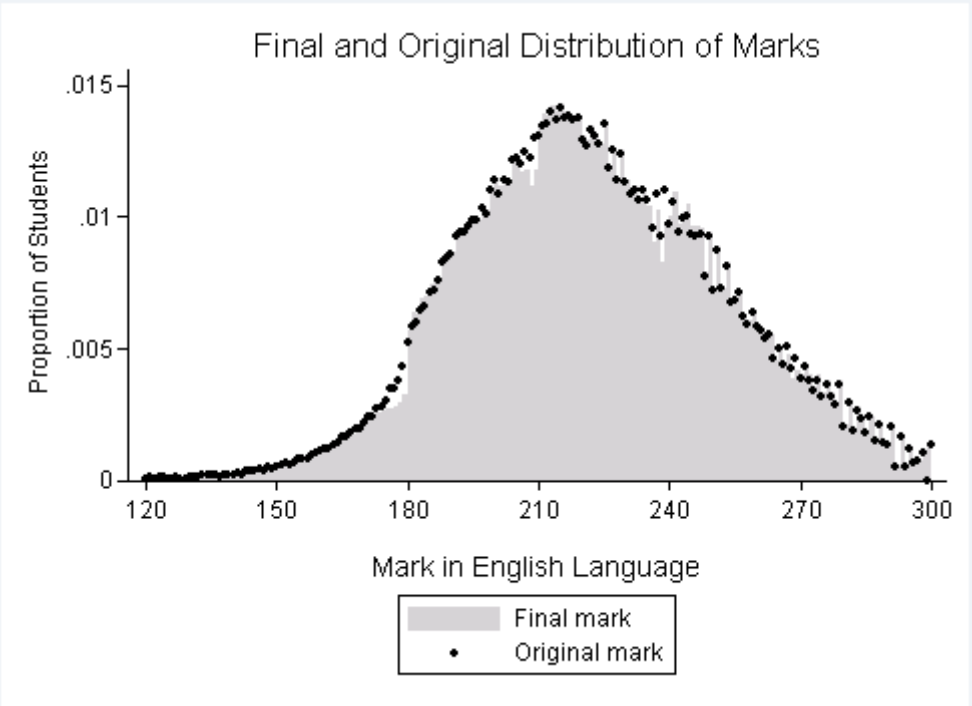


Figure 3. Proportion of Students Asking for a Review and Getting Upgraded, by Original Mark

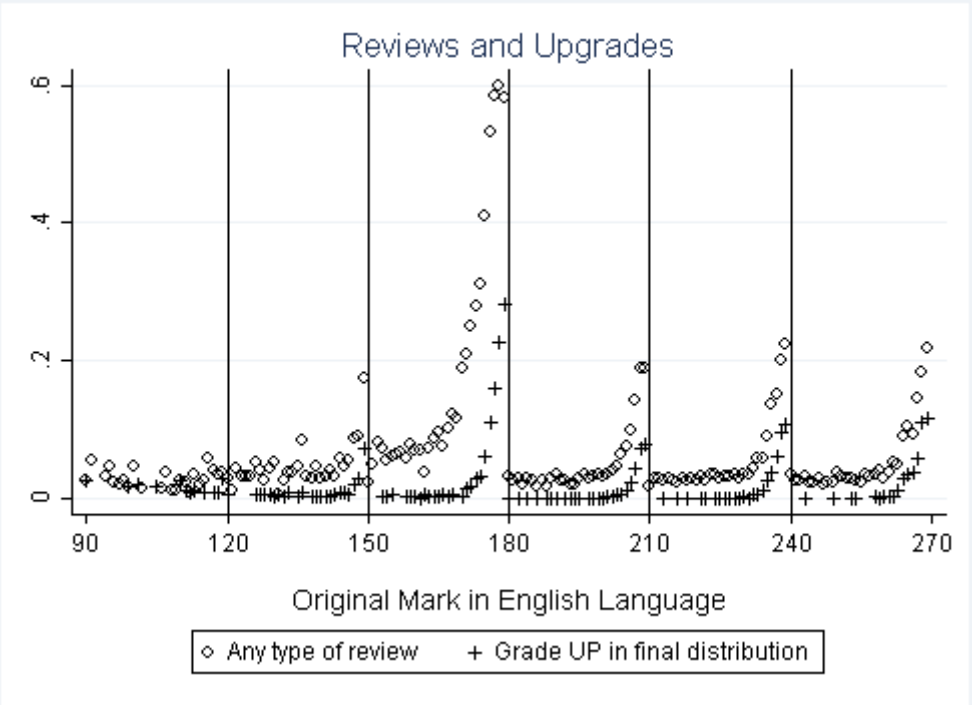


Figure 4. First Stage

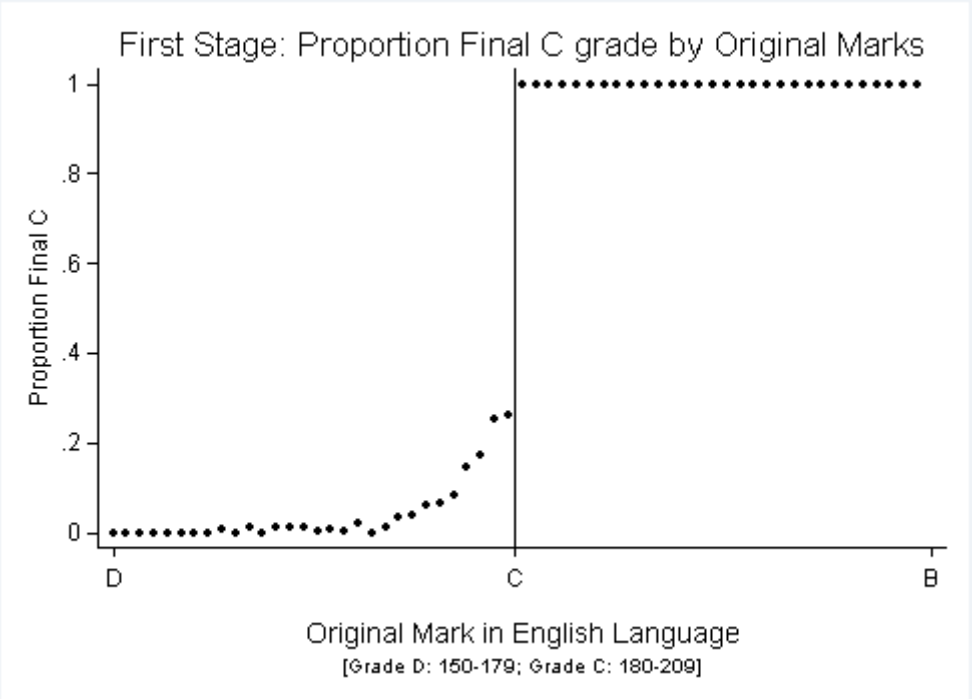
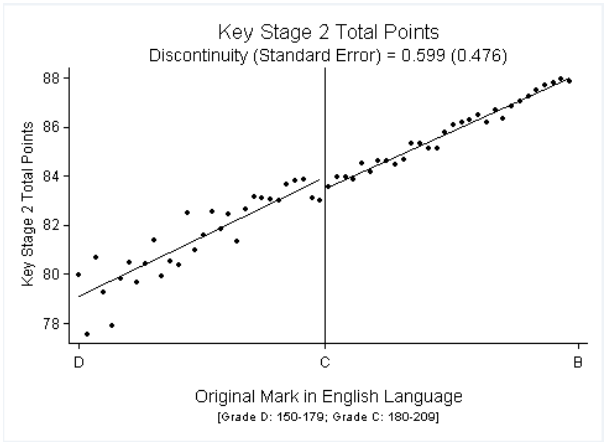
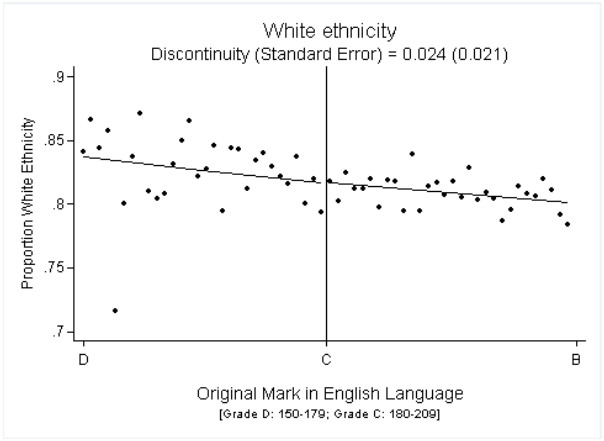
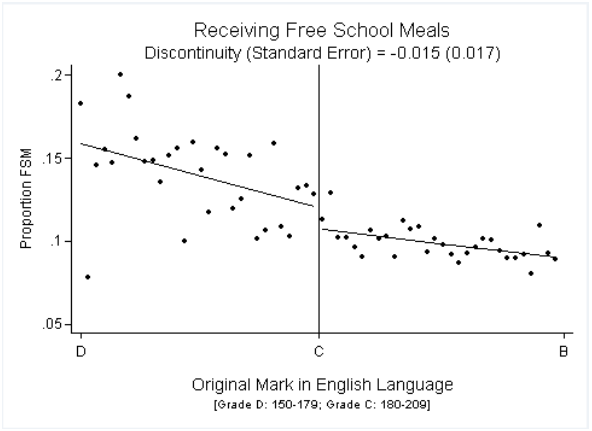
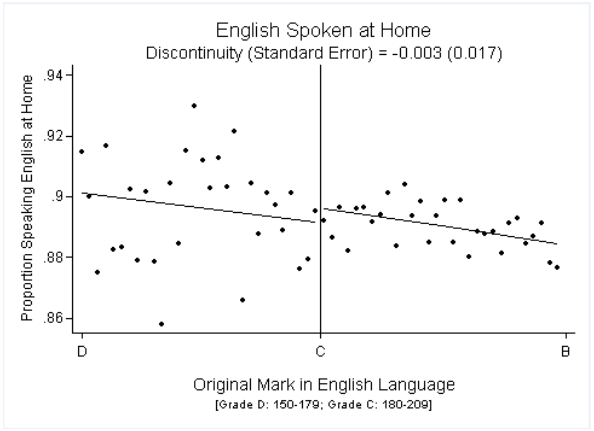


Figure 5. Baseline Characteristics by Forcing Variable









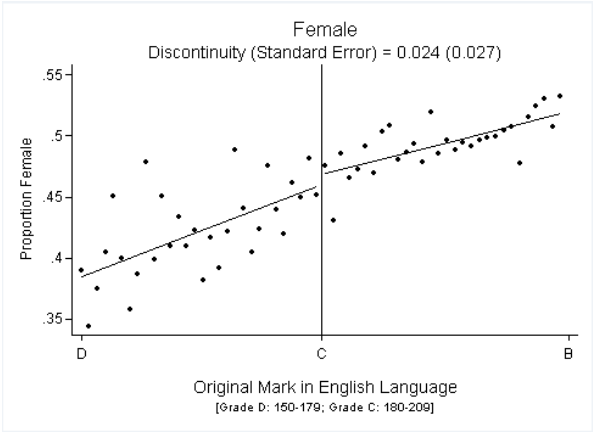


Figure 6. Not Observed in Education at Age 18 by Forcing Variable

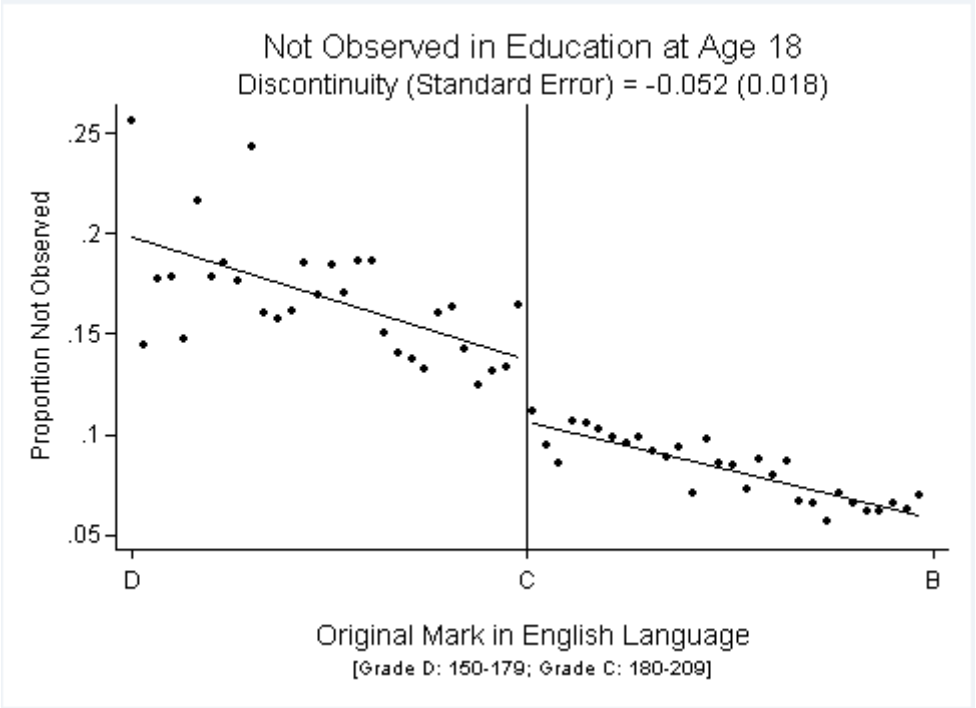


Figure 7. NEET at Age 18 by Forcing Variable

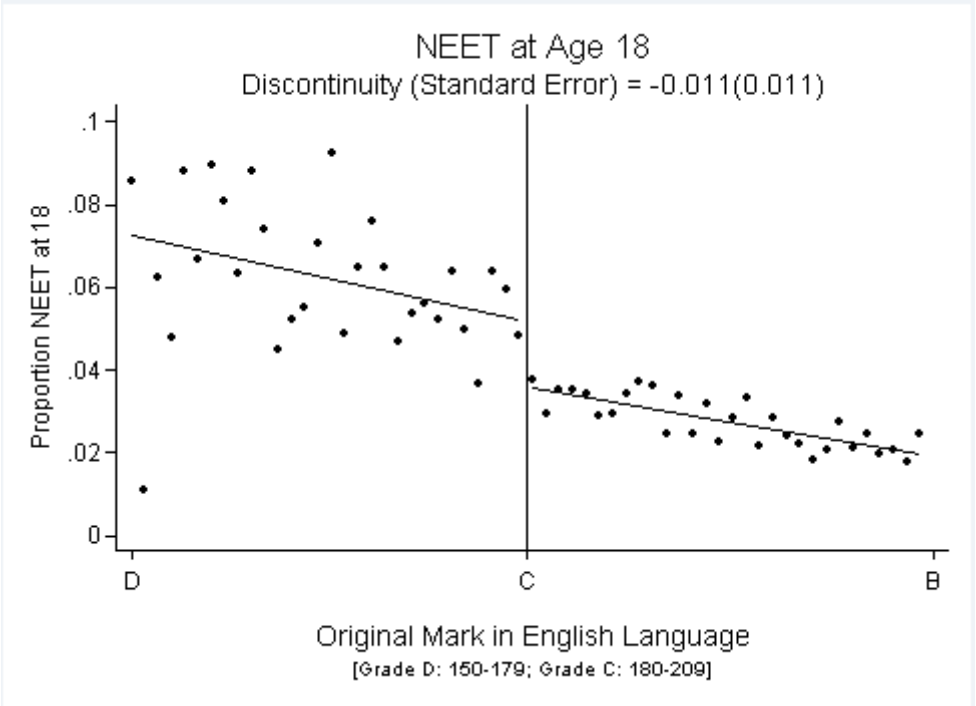


Figure 8. Enrolled in a Level 3 Qualification by Age 19 by Forcing Variable

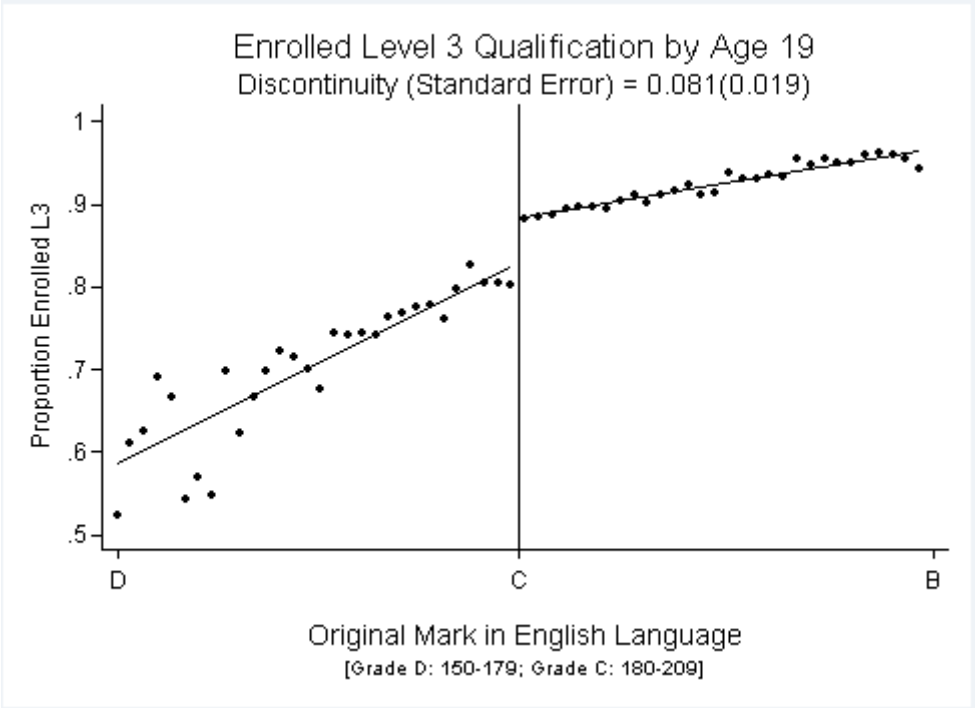


Figure 9. Achieved a Full Level 3 Qualification by Age 19 by Forcing Variable

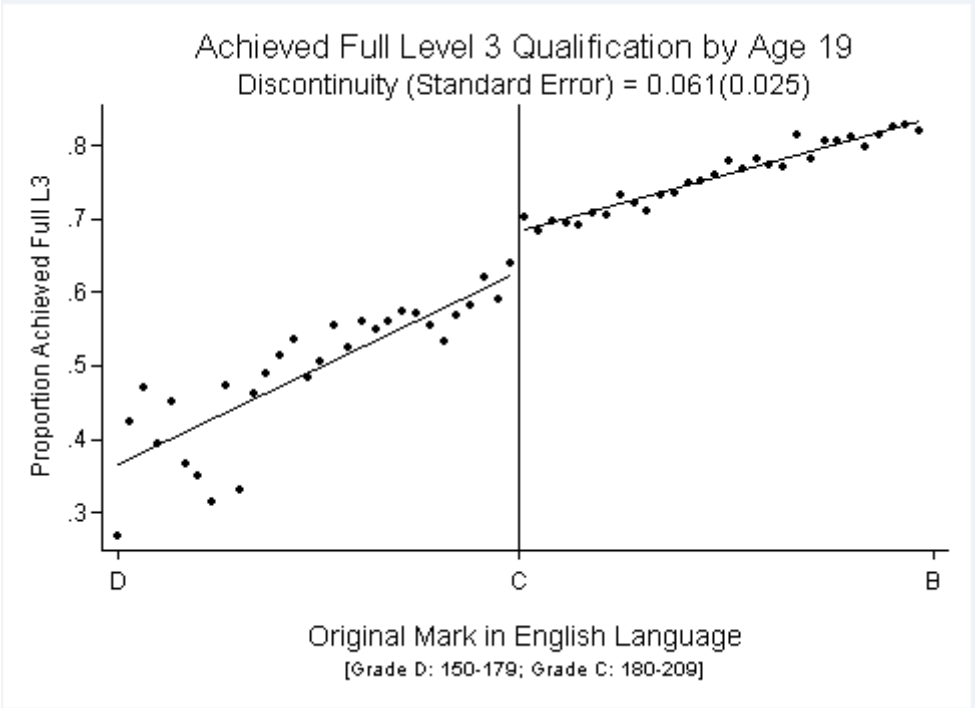


Figure 10. Enrolled in Tertiary Education by Age 19 by Forcing Variable

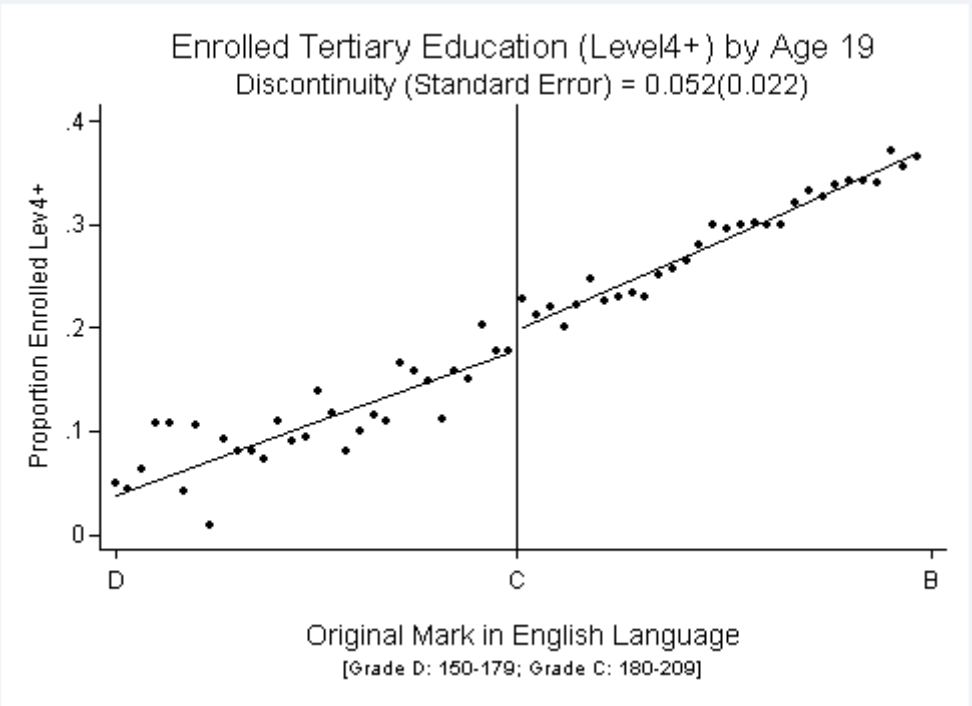


Table 1. Descriptive Statistics

	(1) 2013 cohort	(2) AQA English Language sample	(3) AQA English Language C&D sample - Higher Tier
Achieved C or above (Level 2) in GCSE English (%)	69.1	83.8	85.2
Panel A. Outcomes (%)			
Not observed in Education at Age 18	11.9	7.9	9.2
Not observed in Education, Employment or Training (NEET) at Age 18	5.8	3.0	3.2
Enrolled in a Level 3 Qualification (no matter the size)	78.2	89.0	90.0
Achieved a Full Level 3 Qualification	65.1	77.4	73.2
Enrolled in any Level4+ qualification	29.1	38.6	26.9
Panel B. Predetermined characteristics and prior Key Stage 2 performance			
White ethnicity (%)	81.8	79.9	81.1
Eligible for Free School Meal (%)	14.8	10.3	10.3
English spoken at home (%)	89.1	88.2	89.0
Female (%)	49.2	53.7	48.7
KS2 Total Points	82.9	87.2	85.5
Number of Pupils	544707	189485	49231

Notes: 2013 cohort are those students in the KS4 Candidate 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). Enrolled in a Level 3 qualification means that the student is observed taking a Level 3 subject, independently of the size. A Full Level 3 qualification is equivalent to 2 A-levels (or equivalent qualifications).

Table 2. Determinants of Asking for a Review and Getting an Upgrade

	(1)	(2)	(3)
Dependent variable:	Any review	Any review	Grade up after reviews
White	0.000	-0.003	-0.000
	(0.007)	(0.004)	(0.017)
FSM	-0.003	-0.001	-0.012
	(0.007)	(0.004)	(0.017)
English Language	0.003	-0.001	0.021
	(0.007)	(0.004)	(0.020)
Female	-0.011***	-0.006**	-0.003
	(0.004)	(0.002)	(0.010)
KS2 total points (std)	0.016***	0.005***	0.007
	(0.003)	(0.002)	(0.007)
Original marks	-0.004***	-0.003***	-0.004***
	(0.000)	(0.000)	(0.000)
Mean dependent variable (%)	9.6	9.6	11.2
Sample size	49231	49231	4714
Sample	All higher tier (C&D)	All higher tier (C&D)	Students involved in any kind of review (C&D)
Estimates	OLS estimates	Within school estimates	Within school estimates

Notes: The dependent variables in all regressions are dummy variables. In the first 2 columns, the dependent variable is equal to 1 if any of the units contributing to the final mark was subject to any kind of review. 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. Columns 2 and 3 include school fixed effects. Marginal effects coming from probit estimates are almost identical to the coefficients shown in this table.

Table 3. Exploring Discontinuities Around the C Threshold in Baseline Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
	Local regressions		Global regressions			
	Window: +/-5 points	Window: +/-1 point	Linear polynomial	Quadratic polynomial	Cubic polynomial	Fourth order polynomial
A: Dependent variable: KS2 Total Points						
Grade C	0.382	0.635	-0.292	0.018	0.440	0.440
	(0.462)	(0.732)	(0.185)	(0.290)	(0.386)	(0.386)
B. Dependent variable: White ethnicity dummy						
Grade C	-0.000	0.001	0.004	0.011	0.027**	0.027**
	(0.016)	(0.025)	(0.007)	(0.010)	(0.014)	(0.014)
C: Dependent variable: FSM dummy						
Grade C	0.003	0.004	-0.012*	-0.010	-0.018	-0.018
	(0.016)	(0.023)	(0.007)	(0.010)	(0.014)	(0.014)
D: Dependent variable: English spoken at home dummy						
Grade C	-0.004	0.003	0.007	0.007	0.009	0.009
	(0.013)	(0.019)	(0.005)	(0.007)	(0.011)	(0.011)
E: Dependent variable: Whether student is a female						
Grade C	-0.011	0.037	0.013	0.007	-0.011	-0.011
	(0.025)	(0.035)	(0.011)	(0.015)	(0.021)	(0.021)
Sample size	7082	1409	49231	49231	49231	49231

Notes: All regressions include KS4 school fixed effects (but results are very similar when we do not include them). Standard errors are clustered at the KS4 school level.

Table 4. Global Fuzzy RD Results

	(1)	(2)	(3)	(4)	(5)	(6)
	All C&D higher tier sample			(+/-10 points)		
	Linear FV	Linear FV + endogenous linear interaction	Linear FV + exogenous linear interaction	Linear FV	Linear FV + endogenous linear interaction	Linear FV + exogenous linear interaction
Panel A. Outcome variable: Not Observed in Education at Age 18						
Grade C	-0.039***	-0.037***	-0.037***	-0.047***	-0.056***	-0.052***
	(0.007)	(0.010)	(0.009)	(0.015)	(0.021)	(0.017)
Mean dependent variable (%)		9.2			11.3	
Panel B. Outcome variable: NEET at age 18						
Grade C	-0.020***	-0.021***	-0.021***	-0.026***	-0.029**	-0.028**
	(0.004)	(0.006)	(0.006)	(0.009)	(0.013)	(0.011)
Mean dependent variable (%)		3.2			4.0	
Panel C. Outcome variable: Enrolled in any Level 3 (upper secondary) qualification by age 19						
Grade C	0.104***	0.064***	0.068***	0.092***	0.082***	0.087***
	(0.008)	(0.011)	(0.010)	(0.015)	(0.022)	(0.018)
Mean dependent variable (%)		90.0			86.4	
Panel D. Outcome variable: Achieved a Full Level 3 qualification by age 19						
Grade C	0.093***	0.069***	0.072***	0.096***	0.082***	0.089***
	(0.010)	(0.013)	(0.013)	(0.021)	(0.028)	(0.024)
Mean dependent variable (%)		73.2			66.9	
Panel E. Outcome variable: Enrolled in tertiary education (Level 4 or above) by age 19						
Grade C	0.014	0.026***	0.025**	0.039**	0.024	0.031
	(0.009)	(0.010)	(0.010)	(0.017)	(0.023)	(0.019)
Mean dependent variable (%)		26.9			20.5	
Panel F. Summary Main First Stage: Obtaining a C grade after the appeal process						
Original C grade	0.887***	0.828***	0.828***	0.781***	0.726***	0.726***
	(0.006)	(0.008)	(0.008)	(0.011)	(0.013)	(0.013)
Endogenous interaction		-0.009***			-0.030***	
		(0.000)			(0.002)	
Sample size		49231			14597	
Number of schools		1638			1445	

Notes: Standard errors clustered at the KS4 school level. All regressions control for the following variables: student's ethnicity, gender, language spoken at home, whether receiving free school meals, Key Stage 2 Total Points and KS4 school fixed effects.

Table 5. Local Fuzzy RD Results

	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
	With School Fixed Effects					***Without School Fixed Effects***				
	+/-5 points	+/-4 points	+/-3 points	+/-2 points	+/-1 points	+/-5 points	+/-4 points	+/-3 points	+/-2 points	+/-1 points
Panel A. Outcome variable: Not Observed in Education at Age 18										
Grade C	-0.058**	-0.073***	-0.065*	-0.067	-0.030	-0.071***	-0.077***	-0.063**	-0.076**	-0.073***
	(0.024)	(0.028)	(0.034)	(0.046)	(0.034)	(0.023)	(0.026)	(0.030)	(0.039)	(0.025)
Mean dependent variable (%)	11.6	11.5	11.6	12.3	13.6	11.6	11.5	11.6	12.3	13.6
Panel B. Outcome variable: NEET at 18										
Grade C	-0.027*	-0.035**	-0.025	-0.004	0.016	-0.028**	-0.030*	-0.016	-0.003	-0.014
	(0.015)	(0.017)	(0.020)	(0.026)	(0.019)	(0.014)	(0.015)	(0.019)	(0.024)	(0.015)
Mean dependent variable (%)	4.1	4.1	4.3	4.2	4.3	4.1	4.1	4.3	4.2	4.3
Panel C. Outcome variable: Enrolled in any Level 3 (upper secondary) qualification by age 19										
Grade C	0.087***	0.094***	0.113***	0.110**	0.104***	0.101***	0.111***	0.108***	0.113***	0.110***
	(0.025)	(0.030)	(0.038)	(0.049)	(0.037)	(0.024)	(0.027)	(0.032)	(0.040)	(0.026)
Mean dependent variable (%)	85.8	85.6	85.2	84.9	84.6	85.8	85.6	85.2	84.9	84.6
Panel D. Outcome variable: Achieved a Full Level 3 qualification by age 19										
Grade C	0.067**	0.052	0.067	0.041	0.035	0.090***	0.088**	0.089**	0.075	0.085**
	(0.034)	(0.039)	(0.048)	(0.066)	(0.047)	(0.032)	(0.036)	(0.041)	(0.055)	(0.034)
Mean dependent variable (%)	66.0	66.1	66.4	66.1	67.4	66.0	66.1	66.4	66.1	67.4
Panel E. Outcome variable: Enrolled in tertiary education (Level 4 or above) by age 19										
Grade C	0.037	0.043	0.055	0.054	0.022	0.056**	0.071**	0.080**	0.089**	0.070**
	(0.028)	(0.033)	(0.040)	(0.052)	(0.037)	(0.026)	(0.030)	(0.036)	(0.044)	(0.028)
Mean dependent variable (%)	20.1	20.0	20.6	20.2	20.5	20.1	20.0	20.6	20.2	20.5
Panel F. First Stage										
Coefficient instrument First Stage	0.731***	0.728***	0.730***	0.750***	0.776***	0.724***	0.719***	0.715***	0.734***	0.737***
	(0.016)	(0.018)	(0.021)	(0.028)	(0.023)	(0.015)	(0.017)	(0.020)	(0.025)	(0.018)
Sample size	7082	5671	4212	2817	1409	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					

Notes: Standard errors clustered at the KS4 school level. All regressions control for the following variables: student's ethnicity, gender, language spoken at home, whether receiving free school meals, Key Stage 2 Total Points

Appendix

Figure A1. Not Observed in Education at Age 18 (B, A and A* thresholds)

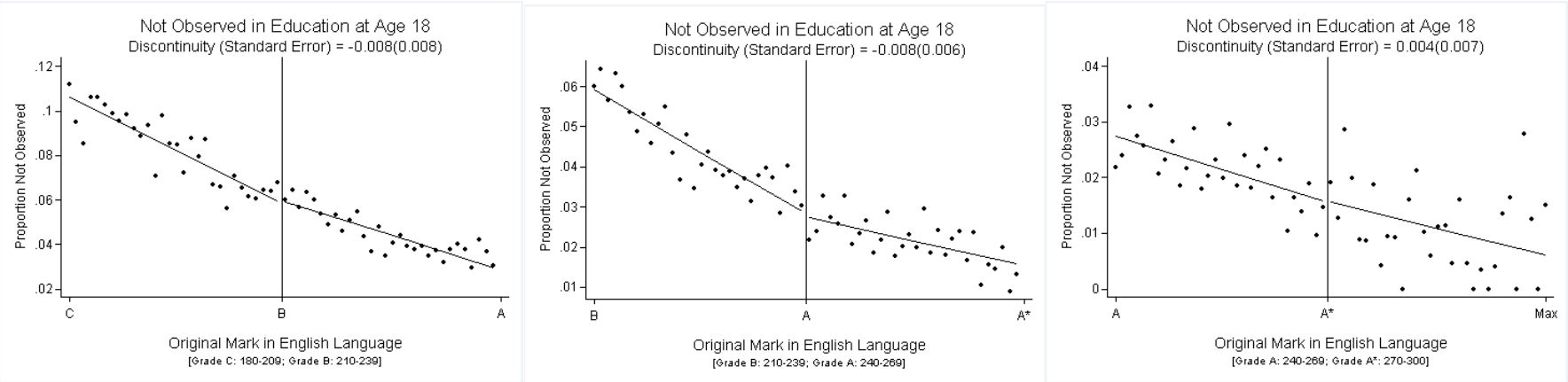


Figure A2. NEET at Age 18 (B, A and A* thresholds)

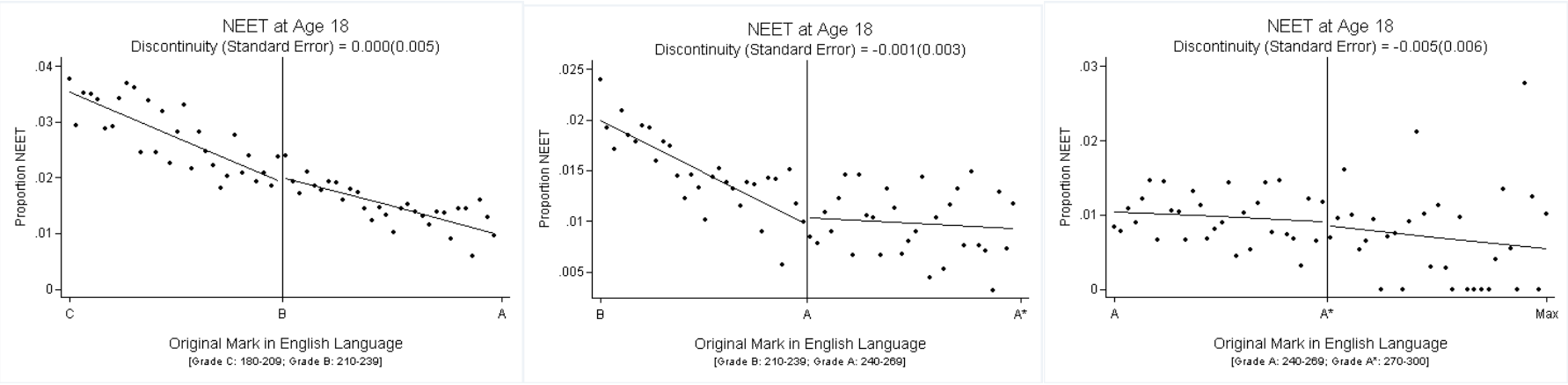


Figure A3. Enrolled in a Level 3 Qualification by Age 19 (B, A and A* thresholds)

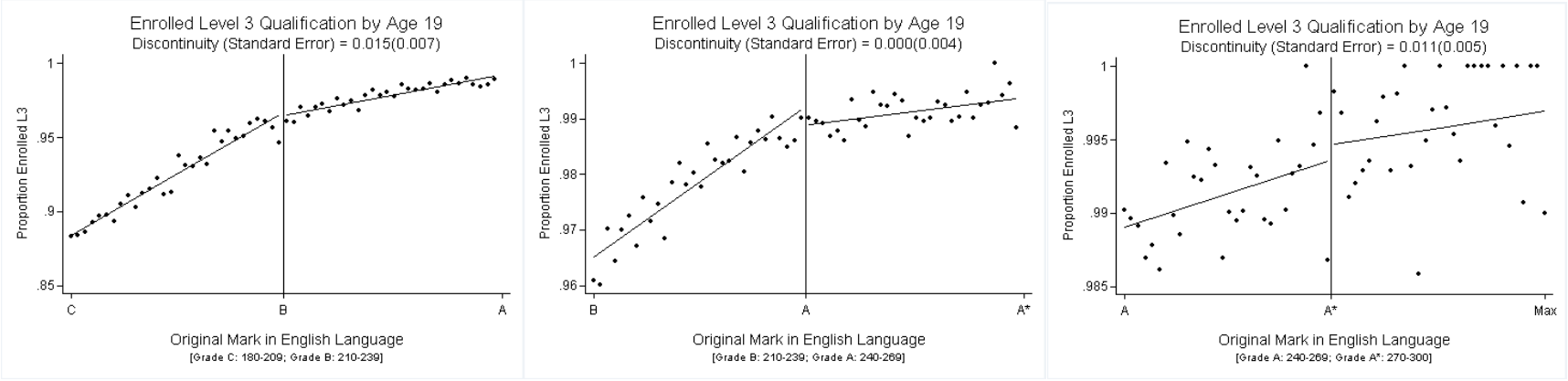


Figure A4. Achieved a Full a Level 3 Qualification by Age 19 (B, A and A* thresholds)

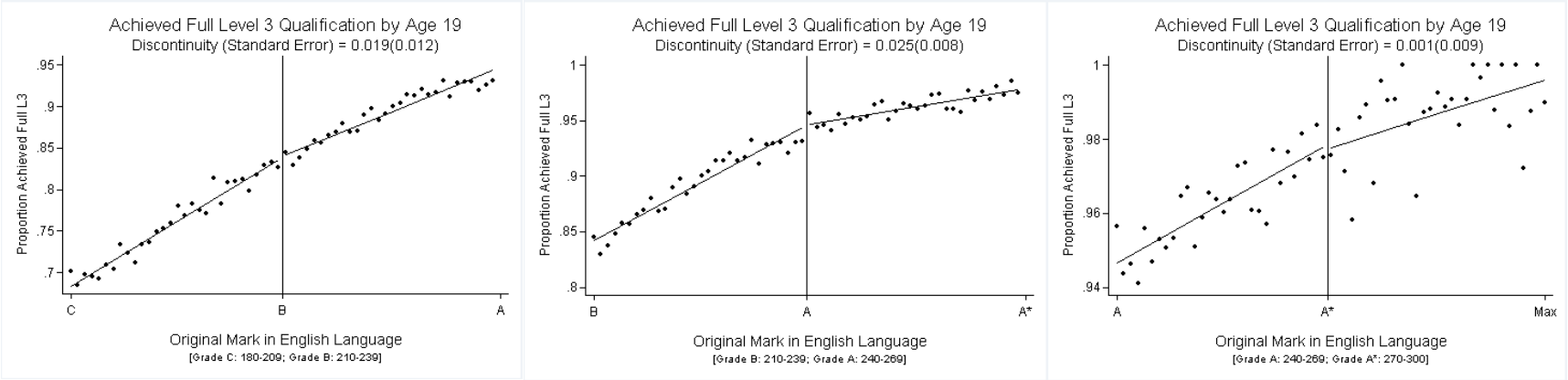


Figure A5. Enrolled in Tertiary Education by Age 19 (B, A and A* thresholds)

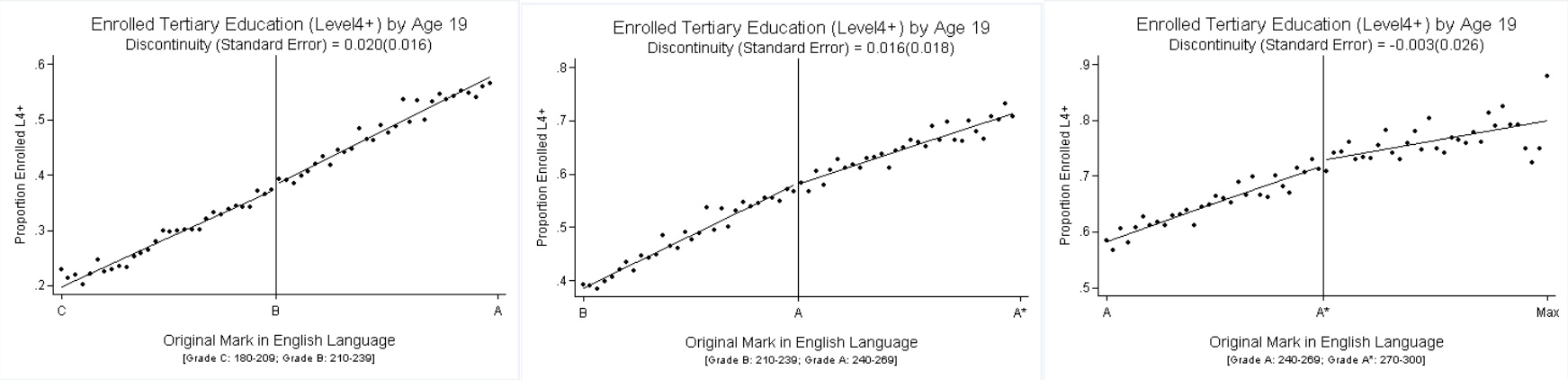


Table A1. Global Fuzzy RD Results – Higher Order Polynomials for Forcing Variable

	(1) Linear FV	(2) Quadratic FV	(3) Cubic FV	(4) Quartic FV
Panel A. Outcome variable: Not Observed in Education at Age 18				
Grade C	-0.039***	-0.037***	-0.035***	-0.038***
	(0.007)	(0.009)	(0.010)	(0.013)
Mean dependent variable (%)		9.2		
Panel B. Outcome variable: NEET at 18				
Grade C	-0.020***	-0.021***	-0.018***	-0.016**
	(0.004)	(0.006)	(0.006)	(0.008)
Mean dependent variable (%)		3.2		
Panel C. Outcome variable: Enrolled in any Level 3 (upper secondary) qualification by age 19				
Grade C	0.104***	0.064***	0.072***	0.078***
	(0.008)	(0.011)	(0.011)	(0.013)
Mean dependent variable (%)		90.0		
Panel D. Outcome variable: Achieved a Full Level 3 qualification by age 19				
Grade C	0.093***	0.067***	0.070***	0.080***
	(0.010)	(0.013)	(0.015)	(0.017)
Mean dependent variable (%)		73.2		
Panel E. Outcome variable: Enrolled in tertiary education (Level 4 or above) by age 19				
Grade C	0.014	0.026***	0.020*	0.022*
	(0.009)	(0.010)	(0.012)	(0.013)
Mean dependent variable (%)		26.9		
Panel F. Summary First Stages				
Coeff First stage	0.887***	0.835***	0.829***	0.783***
	(0.006)	(0.008)	(0.008)	(0.010)
Sample size		49231		
Number of schools		1638		

Notes: FV (forcing variable). Standard errors clustered at the KS4 school level. All regressions control for the following variables: student's ethnicity, gender, language spoken at home, whether receiving free school meals, Key Stage 2 Total Points and KS4 school fixed effects.