School Accountability, Score Manipulation and Economic Geography*

Erich Battistin†

Queen Mary University of London, CEPR, IZA and IRVAPP

Lorenzo Neri‡

Queen Mary University of London

March 2017

Abstract

We document how grading standards for exams at the end of primary education in England have triggered inflation of school quality indicators in national league tables. The cumulated effects over time resulted in significant differences in the quality signalled to parents for otherwise identical primary schools of the country. Institutional features ensure that these differences are as good as random, and reveal that inflation followed from discretion in grading of randomly assigned external markers. We use census data and administrative records on standardized tests, residential sales and business activities to show that this quasi experimental variation reflected in inequality of house prices and land use, influencing local development and urban sprawl. An instrumental variables strategy yields significant house price gains for increased perception of school quality, and lower deprivation in school neighborhoods. Our approach ensures improved external validity with respect to boundary discontinuity strategies.

JEL Classification: C26, C31, I2

Keywords: School quality, House prices, Score manipulation

---

*Preliminary and incomplete version. Special thanks go to Barbara Donahue, Rebekah Edgar, Rebecca Evison, Martin Harris and Tim Leunig for advice and guidance in interpreting results from the National Pupil Database. Our thanks to Ghazala Azmat, Francesca Cornaglia, Francesco Fasani, Stephen Gibbons, Ellen Greaves, Marco Manacorda, Barbara Petrongolo, Steve Pischke, Enrico Rettore, Olmo Silva, Andrea Tesei and participants at the 2016 Workshop on Labour and Family Economics (WOLFE) for helpful discussions and comments. The views expressed here are those of the authors alone.

†School of Economics and Finance, Queen Mary University of London, Mile End Road, London E1 4NS, UK. Contact: e.battistin@qmul.ac.uk. +44 20 7882 3997.

‡School of Economics and Finance, Queen Mary University of London, Mile End Road, London E1 4NS, UK. Contact: l.neri@qmul.ac.uk. +44 075 19622673.

1 Introduction

Understanding household preferences for neighbourhood attributes is of fundamental importance. For one thing, a conspicuous part of household expenditure is devoted to housing. Differences in house prices and economic activities across locations shape residential choices, influencing urban and suburban sprawl and the social consequences associated with such expansion. Among a number of local amenities, there are good reasons to value proximity to good schools in particular. In many countries home residence is the way most children gain access to public schooling. In England for example, the context considered here, boroughs require evidence of tax payment and electoral-roll registration as proof of address. As a consequence, parents may be prepared to pay a substantial premium to secure an address within the desirable school catchment. Mobility motivated by school quality results in residential sorting into communities with similar taste for area characteristics (such as open space and child friendly amenities). Parents may also move because they value socio-economic composition of their peers, affecting neighbourhood characteristics and, eventually, school quality.

Disentangling area composition from school quality effects on house prices and local development is the subject of a sizeable empirical literature (see Black and Machin, 2011, and Duranton and Puga, 2014, for reviews). The relationship between school quality and house prices is typically investigated in the context of hedonic regressions. The evidence available suggests important effects on the willingness to pay for a good school: the review in Machin (2011) documents a 3% house price premium following from one standard deviation increase in test scores. The identifying variation in most empirical strategies stems from price differences across school admission boundaries. Regression discontinuity style strategies have been used to identify the effects of school quality in many countries, including the US (Black, 1999, Bayer et al., 2007, Tannenbaum, 2015), France (Fack and Grenet, 2010) and the UK (Gibbons and Machin, 2003, Gibbons et al., 2013). Alternative quasi-experimental designs exploit school openings and closures or changes to a school's catchment boundaries (Ries and Somerville, 2010, Tannenbaum, 2015).

We study how quality of primary schools affects residential sorting, house prices and land use across narrowly defined neighbourhoods in school districts in England. We focus on public schools, as these enrol over 95% of students in the country (Department for Education, 2012).

Standardized testing for evaluation purposes has been in place since the early 1990s, and provides the yardstick by which school quality is assessed and compared. Examination results are used by the Department for Education to form Performance Tables, an accessible source of comparative data to which is given considerable attention by media and local authorities. Tables are updated every year and contain indicators of student performance along with contextual data on school environment, staff and finances. Household preferences recovered from survey data in England confirm that, among school attributes, parents strongly value academic performance in determining choices (Burgess et al., 2015). Performance Tables report the fraction of students scoring above subject-specific national targets at the end of primary school, as we explain in the next section. Differences across schools in students attaining at these targets is our measure of school quality.[1]

Our identification strategy relies on institutional features regulating the marking scheme of standardized tests. Exams are proctored locally and marked externally by an agency appointed to maintain and develop the national school curriculum and the educational assessments.[2] Exams are randomly assigned to markers, grading is blind and score thresholds used to award national targets are disclosed by the Department for Education only at the end of the process. Until 2007, students who narrowly missed the target had their exam reviewed by the same marker, but those who barely scrape over the borderline kept their result without additional scrutiny. The re-scoring was limited to exams within three marks from the proficiency cutoff, leaving the rest of the distribution unaffected (a similar procedure is implemented in New York's Regents exams, as explained in Dee et al., 2016).

This practice, originally introduced to avoid unfair denial of levels because of low marking quality, was dismissed in 2007 because deemed responsible of boosting results for thousands of students and overstating school standards for over two decades (see Statistical First Release 32/2009, Department for Children, Schools and Families). Evidence of score inflation around achievement thresholds is shown in Figure 1, where presented are language score distributions

[1]Tables layout and information reported underwent minor changes over the years considered in our analysis. A detailed description of content and changes to the information published can be found at http://www.education.gov.uk/schools/performance/archive/index.shtml.

[2]The agency in charge changed over time, but after the period relevant to our analysis. The National Assessment Agency (NAA) was in charge until 2007, when the Qualifications and Curriculum Development Agency (QCDA) took over. The testing program is described below and in https://www.gov.uk/government/collections/national-curriculum-assessments-key-stage-2-tests, accessed on 03 October 2016.

computed from national tests in selected years before and after the removal of borderlining.[3] Continuous lines are obtained from a local-linear fit estimated excluding scores three points around Level 3 ("working towards expected level"), Level 4 ("expected level") and Level 5 ("exceeded expected level") thresholds, which are denoted by vertical lines. Dots represent the percentage of exams by value of the score at the national test. Bunching is evident on the right of critical pass-marks, with no student downgraded. Our estimates below show that, on average, about 18% of exams re-scored below Level 5 were eventually inflated; this figure is 16% for exams below Level 4. Manipulation fades away after 2007, when the re-scoring of exams was abolished.

We argue here that score manipulation reflects marker behavior - specifically, leniency in grading for borderline students. Blind marking yields inflation independent of student and school demographics, adding noise to the fraction achieving at the critical level and changing the perception of school quality. The randomness in the signalling value of test scores is used for identification.[4] The thought experiment sets out the comparison of schools with the same counterfactual score distribution (i.e., the distribution without inflation) but different ranks in Performance Tables because of manipulation. The identifying variability follows from random assignment of exams to markers, year-to-year variation in a school's number of students scoring below critical thresholds, and changes of achievement targets over time and across school subjects which can't be anticipated by markers. Our data show enough noise cumulated in the decade 1998-2007, with sizeable variation across schools. The width of the manipulation region here is known *ex ante* and follows from features of the grading protocol, simplifying estimation with respect to other empirical research on bunching (see Diamond and Persson, 2016)

We use information from a number of administrative sources over two decades, and exploit the multi-layer geographic hierarchy developed by the Office for National Statistics (territorial

[3]KS2 data used here are described in the next section, and in documents and links at https://www.gov.uk/government/collections/national-pupil-database, accessed on 03 October 2016. Figure C.1 and Figure C.2 of Appendix C present distributions for math and science scores, and convey a similar message.

[4]There is evidence that the housing market responds to information about school quality beyond the signalling value of standardized testing. For example Figlio and Lucas (2004) find that school report cards, the listing and evaluation of school performance issued by the education department, have an effect on local house prices over and above that of test scores. In constrast, the variation to the signalling value of quality we consider here originates from random noise. Fuzziness in the quality of schools signaled by matriculation rates in Israel is considered in Lavy (2009).

units, boundaries and maps) for the production of their official statistics. This allows us to measure house prices and socio-economic characteristics of the population across narrowly defined areas of the same residential neighbourhood, and to measure the quality of accessible schools on a block-by-block basis. The analysis controls for unobserved attributes of the local neighbourhood (as in Bayer et al., 2007, and Caetano, 2015) and yields causal estimates with improved external validity compared to studies that live off boundary discontinuities. Indicators of school quality and the geocoding of residential locations are derived from the National Pupil Database, a rich education datasets with standardised scores for all children in England. Residential property sales in census blocks surrounding all schools are obtained from the Land Registry. In addition, we use local development indicators from multiple censuses, and administrative data with location and economic activity of all business organisations in England.

Our findings point to a significant effect of school quality on house prices. One standard deviation increase in school quality yields a 7% increase in the price of houses in the surrounding blocks, corresponding roughly to £14,170 on average. In addition, we document effects on the composition of households in the school neighborhood. We find that one standard deviation increase in school quality raises by about 1% the percentage of professionals and by about 3% the percentage of people with high qualifications in the sorrounding blocks, and lowers unemployment by about 0.5%.

The remainder of the paper is organized as follows. The next section presents the institutional background on schools and tests in England. Section 3 describes our data and the sample selection criteria. Following a graphical analysis, Section 4 documents the effects of 'borderlining' across schools. Section 5 shows the identifying variability and the empirical specifications and presents results for house prices, school composition and local development. Conclusions and directions for further work are in Section 6.

2 Background and Context

The National Curriculum Assessments in England

School age in England begins the term following a child's fifth birthday, and education is compulsory until age 16. Primary education consists of two blocks of years: Key Stage 1

(KS1; ages 5 to 7) and Key Stage 2 (KS2; up to 11). The former phase runs from reception year, which is delivered as pre-school, and two years of formal education known as Year 1 and Year 2. KS2 runs from Year 3 to Year 6. The National Curriculum, introduced by the Education Reform Act in 1988, sets out standardized programmes of core knowledge and attainment targets for all subjects at both cycles.

Our analysis considers only public schools, for which coordination and financial support typically lies with the Local Authority (LA). Community schools, by far the most common, are established and fully funded by LAs.[5] Faith schools, originally established by voluntary or religious bodies (e.g., churches), are more independent but still largely funded by LAs. Among the remaining state-funded schools, foundation schools and academies have governing bodies with the greatest freedom in management. In particular, academies (akin to charter schools in the US, and rare in primary education) are independent of LA control and don't have to follow the National Curriculum. With this exception, all remaining public schools must follow precise guidelines for core subjects. Our working sample retains community, faith, and foundation schools, and excludes a limited number of institutions providing education to children with special needs. Importantly, the number of school openings and closings in the time window considered in the analysis was negligible, as all major changes following the introduction of academies happened from the late 2000s (see Eyles and Machin, 2015)

Criteria for school entry are regulated by LAs. Priority is given to children with special education needs, or with siblings at the school. Faith schools are also allowed to enrol students on grounds of religiosity of parents. Other than this, the most common way of prioritising applications is closeness to school. Catchment areas are by and large within the LA. In our data, for example, only 3.6% of students don't meet this condition. Although there are no legal restrictions on the school choice, applications outside the LA are burdensome and LAs don't have the statutory requirement to find a school for children from a different district (see Burgess et al., 2015, and Gibbons et al., 2013, for institutional details).

Academic assessment is statutory at the end of each stage of education, and attainment targets are set using six progressive levels of learning (from Level 1 to Level 6). Level 2 is expected by the end of KS1, and Level 4 is expected at KS2. Important changes to the

[5]In England the majority of students attend state-funded schools: among students aged 5 to 10 only about 5% of them attend private schools (Department for Education, 2015)

measurement tools used to assess progress were made in the past two decades. Nationwide standardized testing at KS1 was phased out in 2004 in favour of decentralized assessments from teachers. The Standards and Testing Agency (STA), a government body charged with educational assessment, provides teachers with standards a child should be assessed against at the end of KS1. Still teachers are free to make judgements based on their knowledge of the student.

On the other hand, standardized testing at KS2 has been conducted continuously in the three core subjects (English, mathematics and science). Because of this, KS2 results are key for accountability purposes. Minimum levels of quality, or "floor standards", are regularly set by the Government using KS2 scores to hold schools responsible for their performance. In addition to a number of contextual indicators, Performance Tables report every year for all state-funded schools percentage of students at or above Level 4 and Level 5, and an overall score obtained combining these percentages across subjects as explained below.[6] It is well documented that high academic standards are considered the most important school attribute by parents, followed by socio-economic composition and proximity (for examples see Hastings and Weinstein, 2008, and, for England, Burgess et al., 2015).

Grading Protocols

KS2 tests are proctored locally and marked externally by an agency appointed by the Department for Education. LAs or STA can make unannounced visits on the test day to ensure that test protocols are implemented correctly. Markers have no relationship with the school and receive training on the marking scheme they must follow. Grading is blind and carried out without knowing the thresholds required to award achievement levels.[7] Mark boundaries are set by senior examiners and made public only at the end of the process. Thresholds change every year and across subjects, and official documents as well as our data offer no

[6] KS1 results are not reported directly, although in some academic years were used to publish student value added. The computation of this quantity, however, did not follow a consistent methohology in the time window considered in our analysis. A measure of value-added has been reported since 2002. However, the results in Wilson et al. (2006) show that this indicator is not valued by parents in determining residential choices.

[7] Burgess and Greaves (2013) use KS2 test scores as the "true" assessment as opposed to teacher assessment. They define KS2 tests grading as "quasi-blind" since markers are able to see the name of the pupil on the scripts and thus can infer their ethnicity. In our data this does not seem to be the case as we do not observe any discontinuity around thresholds when we consider student ethnicity (see Figure C.3).

evidence that they can be predicted. Once marking is concluded, schools can request a review of their scripts for a fee, which is refunded in case of successful appeal. Tests are a combination of multiple choice questions and open-response items, for which a more intense grading effort is required. This opens the door to interpretation and opinion (we show examples of items in Appendix C). Scoring materials and instructions are provided to markers to enforce consistent grading, including examples and precise guidelines on how to interpret possibly ambiguous answers.

Since tests were instigated in the 1990s, to avoid students being unfairly denied a level all exams falling *three* points or less below the pass-mark were revisited by the original marker; exams falling above were not. This procedure, known as "borderlining", was abolished in 2007. It has been estimated that between 1996 and 2007 borderlining led to 300,000 pupils being upgraded, with more pronounced effects at Level 4 and Level 5 cutoffs.[8] In Figure 1, for example, the fraction of students scoring above Level 5 in 2007 exceeds by about 3% the value extrapolated through the continuous line. One year later, this same quantity is below 1%. The remaining discontinuities are the result of school appeals, which increased substantially after 2008. This can be seen from Figure C.4 of Appendix C, which reports the percentage of exams for which schools appealed (dotted line) and the percentage of successful appeals (continuous line).

This evidence suggests that the abolition of borderlining made schools more liable, at least in their perception, for correcting errors around thresholds. At the same time, the limited effect on score distributions of the increased number of appeals suggests that borderlining is the prime suspect for discontinuities around achievement thresholds until 2007. As Performance Tables report the fraction of students attaining at each target, not the value of their scores, there may be a large signal change in perceived school quality caused by such discontinuities.

[8]See http://www.standard.co.uk/news/marking-fiddle-has-boosted-sats-results-6918127.html and Statistical First Release 32/2009, Department for Children, Schools and Families. Results available on request show that the size of discontinuities in Figure 1 is not differential by school type (community, faith, and foundation).

3 Data

Geographic Hierarchies and Sample Selection

We use administrative records from the National Pupil Database (NPD) on primary school students in England (about 600k per year). Data include scores and progression (i.e., attainment level awarded) though key stages, along with school and student characteristics such as gender, ethnicity, first language, eligibility for free school meals and special educational needs. The first wave of the NPD started in 2002 by linking national tests to the school census, although scores in English, mathematics and science are collected from 1998.[9] The availability of students' residence and school postcodes allows the linkage with small area statistics (e.g., crime rates, social homogeneity, labour market participation and land use) produced by the Office for National Statistics (ONS) using the 2001 and 2011 censuses.

The geography considered is very fine and consists of areas of compact shape, fitted within LA boundaries, with a target population of 400 households. Given this size, we will conventionally call these areas "blocks". We use variability within homogenous neighbourhoods consisting, on average, of an aggregation of 5 adjacent blocks.[10] The right hand side panel of Figure 3 presents an example of the geographic hierarchy for the borough of Tower Hamlets, to the East of the City of London and including the redeveloped Docklands region. The borough is organized into 31 neighbourhoods and 130 blocks with a population of 254,100 (listed in the 2011 census). Blocks have an average size of just above 0.05 squared miles, and neighbourhoods are akin to squares with each side 0.5 mile long.

We keep all neighbourhoods in metropolitan areas, and urban neighbourhoods in non-metropolitan areas of England. Our primary sample consists of all blocks with at least one school of the LA within a 0.6-mile radius of the block's centroid. A similar geographic width was used in other studies (see Machin, 2011, Gibbons et al., 2013 and Burgess et al., 2015), and represents the 60th percentile of the student-school distance distribution in the NPD. Our primary sample consists of 5,187,610 students in 12,481 schools, across 27,414 blocks

[9] Science tests were discontinued in 2010. English scores are aggregated from separate reading and writing tests.

[10] Our definition of neighbourhoods uses Middle Layer Super Output Areas (MSOAs) as defined by the Office for National Statistics. There are 6,781 of such neighbourhoods in England, aligning to LA boundaries, with a population size between 5,000 and 15,000, and on average 3,000 households. What we call blocks are instead Lower Layer Super Output Areas (LSOAs), a set of 32,482 narrowly defined areas across England used by the ONS for the computation of small-area statistics. See Appendix A for additional details.

of 6,104 neighbourhoods in England (see the left hand side panel of Figure 3). We however check the robustness of our conclusions considering two alternative samples defined from 0.4-mile and 0.8-mile radiuses centred on block centroids. Descriptive statistics for school and demographic composition across blocks are presenetd in Table 1.

Residential sales and data on businesses

We use administrative records from the Land Registry with all residential sales between 1995 and 2011. Each transaction reports the sale price, date of transfer, property type (detached, semi-detached, terraced, flats/maisonettes) and property age (newly built property or established residential building). Addresses are geocoded (within blocks), and linked to statistics on the area where each sale lays. In particular, number of dwellings by council tax bands is published yearly by the ONS among their neighborhood statistics.

Business data from 1997 to 2015 have been collected by the Office for National Statistics and HM Revenue and Customs and are available in the Business Structure Database (BSD). Businesses listed are obtained from the Inter-Departmental Business Register (IDBR) and account for about 99% of the UK economic activity. Each business is divided between the "enterprise", which represents the overall business organization, and "local units" (e.g. stores, bank branches). For each business industrial classification, birth and termination year, among other information, are reported. Each unit is precisely geocoded within blocks.[11]

4 Graphical Analysis

School Quality Effects of Borderlining

We begin with non-parametric plots quantifying bunching in score densities near achievement cutoffs pooling data from 1998 (first available year of data from the NPD) to 2007 (when borderlining was abolished). Figure 2 is obtained considering scores in the $[-8, 7]$ window centered at relevant cutoff. We compute f_{scjt}, the percentage of students scoring $s \in [-8, 7]$ around cutoff c (Level 3, Level 4 and Level 5) for subject j (English, mathematics and science) in year t (between 1998 and 2007). Plotted are residuals from separate regressions of f_{scjt} on

[11] A description of the Business Structure Database can be found at https://discover.ukdataservice.ac.uk/catalogue?sn=6697.

a full set of subject and time dummies for the three achievement thresholds. Continuous lines are fitted values generated by local linear regressions (LLR), and the smoother uses data on one side of the cutoff only with a normal kernel. Information is presented by attainment level. Consistent with expectations LLR fits show discontinuities around cutoffs, the sharpness of the break varying with attainment level. Regressions show a drop in score densities from three points below cutoffs, which is compensated by bunching above achievement thesholds for $s \leq 1$. It is clear that a number of students are moved from below to just above thresholds, with otherwise smooth score distributions away from these critical points.

The effect of borderlining is obtained by contrasting f_{scjt} to the distribution $\hat{f}_{scjt}$ that would have been observed in the absence of notches and bunching around proficiency cutoffs. Such counterfactual distribution is retrieved borrowing from the literature on bunching (see Kleven, 2016, for a review). The idea is to fit a flexible polynomial to the observed score distribution excluding data in a window around thresholds. Knowing how borderlining is implemented and using non-parametric results from LLR, we consider the area from three marks below to two marks above thresholds ([-3,1]) as excess bunching fades out after this point.[12] The following equation is estimated separately by attainment threshold (the index c on parameters is omitted):

$$f_{scjt} = \alpha(j,t) + \sum_{i=0}^{2} \beta_i s^i + \sum_{i=-3}^{1} \gamma_i 1(s = i) + \varepsilon_{scjt}, \quad (1)$$

where $\alpha(j,t)$ is shorthand for a full set of subject and time effects centred at zero, a second order polynomial in s is used to approximate counterfactual densities, and the γ_i's represent score specific effects of notches or bunching (below and above thresholds, respectively). The equation is estimated imposing that the "missing mass" equals the "bunching mass", which implies a linear restriction of the estimated γ_i's. Dashed lines in Figure 2 represent predicted values of $\hat{f}_{scjt}$ implied by the equation above.

The size of this drop is shown in Panel A of Table 2, which reports in columns (1) to (5) estimates of γ_i's by attainment threshold. The value $\sum_{i=-3}^{-1} \gamma_i$ is in column (6), and represents our estimate of the notch induced by bordelining. Consistent with Figure 1 the notch in score densities at Level 5 is much larger and estimated at about 1.5%, twice as much

[12] Our conclusions are robust to the choice of this interval, and to the order of the polynomial used in equation (1), below.

that at Level 4. The discontinuity at Level 3 is instead negligible. Importantly, equation (1) doesn't detect discontinuities away from scores relevant for borderlining. This can be seen from Panel B of Table 2, where reported are estimates using a $[-8, 7]$ window centred ten points below the critical thresholds (see Appendix B for details). All γ_i's and the missing mass are precise zeros, across the three attainment levels.

The Anatomy of Discontinuities

A closer look at score distributions before 2008 reveals that discontinuities are not the result of adjustments to random errors in marking. Random errors are (arguably) symmetrically distributed, and would result in some scores being adjusted downwards. Figure 1 weighs against this hypothesys, showing that the density of marks more than four points from achievement levels is not affected. At the same time, the drop in score distributions becomes more evident near relevant thresholds, again suggesting that random errors are not likely to be the explanation. A visual inspection of Table 2 reveals this gradient, with values for γ_{-1} larger (in absolute terms) than for γ_{-3}.

The simplest story seems most likely: markers manipulate scores moved by a "genuine" willingness to help, and students falling just below an important grade boundary may benefit from having their score manipulated upwards. The fact that score densities are smooth across achievement thresholds after 2008 suggests that manipulation is unrelated to accountability incentives, as these were not changed by the abolition of borderlining. Cultural norms arising from the grading protocol have been found to be the driver of manipulation in other contexts (see Angrist et al., 2016, for Italy, Diamond and Persson, 2016, for Sweden, and Dee et al., 2016, for the United States).

The theoretical case for manipulation can be fleshed out using a stylized model of grading behaviour for exams reviewed with borderlining. Assume that the utility of external markers is linear in number of exams upgraded. The cost of upgrading increases with distance between a student's mark before re-scoring and the threshold. It follows that students upgraded will have their mark moved to the threshold, implying a spike in the score distribution at that value. Upgrading should be more likely for marks that were originally closer to thresholds, implying more pronounced drops in the distribution near critical values. These are the empirical regularities observed in Figure 1 (indeed, in all years until 2007).

Table 3 shows that manipulation is across the board and doens't target exams selectively.[13] We first collapse data to the score-cutoff-subject-year level, and then estimate equation (1) using on the left hand side a number of student, school and area characteristics at Level 4 and Level 5 thresholds. We consider ethnicity, gender, eligibility for free school meals and language ability of students from NPD data. In addition, we use characteristis of the area where the student lives from the 2001 census. Estimates of the γ_i's by attainment threshold are precise zeros, suggesting that students upgraded are not selectively different in terms of family background and type of school attended. The statistical analysis in Table 3 mirrors conclusions from a visual check for discontinuities in Figure C.3 of Appendix C. These findings are consistent with the fact that exams are graded anonimously, implying that markers are not under pressure to improve the standing of schools in Performance Tables or the reputation of students as demonstrated in other institutional contexts.

5 First Stage and Empirical Specifications

First Stage

Our statistical analysis models outcomes in block b of neighbourhood n as function of school quality in that same block, q_{bn}. The latter quantity is proxied by two indicators constructed as in Performance Tables. We first consider the percent of students scoring "above expectations" (i.e., at or above Level 5). The percent of students above Level 5 is a measure of school excellence, as Level 4 is the expected level at KS2. In addition, we derive a summary score calculated by assigning points to each student's results using an equivalence scale provided by the Department for Education.[14] We construct both indicators of school quality pooling

[13] Dishonesty and shirking, as in Jacob and Levitt (2003) or Angrist et al. (2016), are not the only forces driving manipulation. Studies have documented that discretion in grading may favour students with certain characteristics. Dee et al. (2016) argue that score manipulation in New York Regents Examinations is motivated by altruism towards students by local proctors. Diamond and Persson (2016) show that teachers in Sweden adjust marks of students who had a bad test day. Lavy (2008) documents the existence of gender bias (against male students) for matriculation exams in Israel. Neal and Schanzenbach (2010) show that the establishment of proficiency levels, used for school accountability, leads schools to focus more on the "marginal student". Additional examples of students' discrimination by teachers were documented for England (Burgess and Greaves, 2013), France (Terrier, 2016) Israel (Lavy and Sand, 2015) and India (Hanna and Linden, 2012).

[14] In Performance Tables for 2003, for example, each student is awarded a number of points depending on the achievement level attained (15 points at Level 2; 21 points at Level 3; 27 points at Level 4; 33 points at Level 5) and the average school score is calculated by adding the total points across subjects. A score of 30 would therefore mean that, on average, students achieved more than Level 4 but less than Level 5.

exams from 1998 to 2007 in schools within a fixed distance of a block's centroid, mirroring the procedure discussed in the data section. It follows that q_{bn} represents the average quality of schools around block b in the years before the abolition of borderlining (see Panel C of Table 1 for summary statistics).

A measure of how borderlining affects the perception of schools in block b of neighbourhood n, z_{bn}, is constructed in a similar manner. We start by computing f_{scjt} using all schools within a fixed distance of a block's centroid. We then estimate equation (1) by block and attainment threshold to obtain values of the missing mass $\sum_{i=-3}^{-1} \gamma_i$.[15] By iterating over blocks in the sample, this procedure yields our proxy for the noise in the quality of schools around blocks. The quantity z_{bn} is defined as:

$$\frac{\sum_{i=-3}^{-1} \gamma_i}{\sum_{s=-3}^{-1} \sum_{i=0}^{2} \beta_i s^i}, \tag{2}$$

which represents the percent noise arising from borderlining relative to $\sum_{s=-3}^{-1} \sum_{i=0}^{2} \beta_i s^i$, the counterfactual fraction of students falling within three marks from thresholds. Figure 4 demonstrates the randomness of z_{bn}, on the vertical axis, conditional on $\sum_{s=-3}^{-1} \sum_{i=0}^{2} \beta_i s^i$. The correlation coefficients between these two measures are 0.05 and 0.004 for Level 5 and Level 4 respectively. Panel D of Table 1 shows that the average of z_{bn} is about 19%, and that "noise to signal" ratios (missing mass divided by percentage of students attaining at Level 4 or Level 5) are about 6%.

Figure 5 documents the relationship between school quality measured in Performance Tables, q_{bn}, and the effects of borderlining, z_{bn}, in the main sample. Panel A considers the composite score derived by the Department for Education, while percent of students scoring "above expectations" in math is presented in Panel B. The vertical axis in both graphs reports residuals from a regression of q_{bn} on the same controls included in equation (3), below, offering a visual interpretation of the first stage. The associated first stage estimates, reported in Table 4, show that one standard deviation change in the value of z_{bn} (about 6.5% in the main sample) increases school quality with a coefficient around 0.08 standard deviations (hereafter, σ).

[15] In the interest of precision, in our preferred specification we pool scores from 1998 to 2007, use collapsed data to the score, cutoff and subject level and estimate a version of equation (1) without the index t. Our results are not sensitive to this choice. Appendix B offers an in depth analysis of this generated regressor, showing that the block-specific effects of borderlining are precisely estimated. In particular, we perform placebo tests to show that estimation of the missing mass yields precise zeros away from proficiency cutoffs.

House Prices

We regress the logarithm of house prices pooling residential sales from 1995 to 2011 on a full set of time and LA dummies, LA-specific quadratic trends (meant to capture general deterioration, or beautification, of school districts) and house characteristics. Prices in block b of neighbourhood n, y_{bn}, are obtained as block average of residuals from this regression, as this is the aggregation level at which the regressor of interest, q_{bn}, varies. We assume a one-year lag before information published in Performance Tables reflects in house prices and compute block averages, in our primary sample, using 2,036,105 observations from 2008 to 2011. Summary statistics for y_{bn} and the outcomes used in what follows are reported in Table 1.

We consider parametric models that exploit discontinuities in score distributions arising from borderlining. The following equation is estimated:

$$y_{bn} = \tau_0(n) + \tau_1 q_{bn} + \tau_2 x_{bn} + u_{bn}, \tag{3}$$

where $\tau_0(n)$ is shorthand for a full set of neighbourhood effects meant to describe unobserved quality at that level, and x_{bn} is a vector of block-averaged covariates. All regressions control for a quadratic polynomial in distance from the closest school, number of schools within the radius and their size between 1998 and 2007, and population density (Panels B and C of Table 1 present descriptives). The instrument used for reduced form and 2SLS estimation of equation (3) is z_{bn}. Standard errors are clustered on LA.

A simple OLS estimation, shown as a benchmark, suggests a premium of approximately 4% for one standard deviation increase in the value of q_{bn}. This can be seen from columns (1) and (4) of Table 5 for the two indicators of school quality considered. Estimates from the primary sample are reported in the central panel and are robust to the choice of distance from block centroids.

2SLS estimates of school quality on house prices are in columns (2) and (5) of Table 5. One standard deviation change in q_{bn} increases school quality between 6.5% and 8%, depending on the indicator considered. A mild gradient in distance with the expected sign emerges from the table, although estimates across panels are not statistically different. The premium is larger than that estimated from plain OLS. When used to quantify the monetary equivalents, our estimates suggest a willingness to pay for better quality of £13,157-16,194 (at the mean of £202,425 in 2008-11 prices for our main sample).

Schools and Their Catchments

We estimate equation (3) using, on the left hand side, indicators of school and area composition obtained by collapsing data to the block-year cell. Our analysis starts by using block-level statistics from the 2011 census. Four indicators of socio-economic composition are considered: percentage of professionals, percentage holding a degree, percentage of unemployed and percentage of people in good health.[16] Table 6 shows that one standard deviation change in school quality has positive effects on the socio-economic status of residents, and yields lower unemployment. Effects here picture a gradient with distance somewhat stronger than for house prices, and get larger as the school catchment shrinks. These results suggest that school quality is a driver of house prices, and induces substantial residential sorting in a school's catchment.

The use of distance as tie-breaker to determine school admission implies that poorer families are "priced out" and more likely to live far from high-achieving schools. We should therefore expect that changes in socio-economic indicators of the area reflect into changes in composition of students. This expectation is borne out by Table 7, which reports estimates of equation (3) using a number of school ouctomes from 2011 NPD data.

6 Summary and Directions for Further Work

Score manipulation in Key Stage 2 exams is used to study how households respond to available information on school quality in England. An IV strategy that exploits quasi-experimental variation arising from the marking process unveils strong preference of parents for performance of accessible state-funded primary schools. The willingness to locate close to good schools triggers surge pricing, yielding a £13,157-16,194 premium (at the mean of £202,425 in 2008-11 prices) for one additional standard deviation of (perceived) quality. Areas with good schools experience lower unemployment and new homeowners from higher socio-economic

[16]Following the Standard Occupational Classification (SOC2010), the census defines as professionals managers, employees in professional occupations (e.g., science research, teaching) and associate professionals working in technical occupations (e.g., engineers, business). Individuals with high qualifications are those with university degree or holding UK national vocational qualifications (NVQ) at Level 4-5. For details see https://www.nomisweb.co.uk/census/2011/qs501ew.pdf.

backgrounds, causing substantial sorting across school districts and influencing urban and suburban sprawl.

Why is the relationship between manipulation of Key Stage 2 exams and the economic geography of English neighborhoods of general interest? Most research on the effects of school quality on residential sorting and house prices relies on boundary discontinuities (see Black and Machin, 2011). By contrast our strategy uncovers a new substantive identification strategy with improved external validity, and has potential for replicability in different institutional settings. As in the case considered here, manipulation of New York's Regent's exams appears to be unrelated to NCLB-style accountability pressure (Dee et al., 2016). Similar conclusions hold for Sweden (Diamond and Persson, 2016).

Our findings raise a number of additional questions, including those of why manipulation not motivated by accountability concerns is so prevalent and what can be done to enhance accurate assessment in England and elsewhere. One important policy question is what happens to children of households "priced out" of good areas. It is also worth asking whether score manipulation has longer terms effects on student outcomes (see Ebenstein et al., 2016, in addition to references above). We hope to address these questions in future work.

Table 1: Descriptive Statistics

	Within 0.4 mile		Within 0.6 mile		Within 0.8 mile	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. House characteristics						
Prices (2008-11)	198,280.50	144,578.00	202,425.40	143,988.00	204,798.40	145,542.70
Log prices (2008-11)	10.8929	0.4710	10.9194	0.4716	10.9309	0.4735
Percent detached (2008-11)	0.1516	0.1831	0.1685	0.1940	0.1766	0.1992
Percent semi-detached (2008-11)	0.3174	0.2204	0.3187	0.2171	0.3178	0.2155
Percent flats (2008-11)	0.1829	0.2412	0.1803	0.2375	0.1784	0.2359
Percent in medium/high tax bands (2011)	0.1404	0.2002	0.1535	0.2094	0.1593	0.2128
Panel B. Area characteristics						
Unemployment rate (2011)	0.0486	0.0252	0.0470	0.0248	0.0464	0.0247
Percent highly qualified (2011)	0.2556	0.1289	0.2601	0.1281	0.2616	0.1276
Percent managers (2011)	0.0960	0.0382	0.0988	0.0394	0.1001	0.0401
Percent in good health (2011)	0.8065	0.0545	0.8082	0.0547	0.8088	0.0547
Percent white (2011)	0.8718	0.1853	0.8790	0.1790	0.8818	0.1768
Number of schools around block	1.8257	1.0859	2.9731	1.9612	4.5019	3.0720
Panel C. School characteristics						
Scores (1998-2007)	27.3536	1.2821	27.3971	1.1568	27.4149	1.0505
Percent attaining at math L5 (1998-2007)	0.2666	0.1039	0.2698	0.0944	0.2709	0.0861
Scores (2011)	27.7487	1.1813	27.7824	1.0543	27.7971	0.9518
Percent attaining at math L5 (2011)	0.3348	0.1161	0.3379	0.1041	0.3390	0.0941
Percent white students (2002-2007)	0.8095	0.2396	0.8200	0.2243	0.8238	0.2159
Percent male students (2002-2007)	0.5084	0.0305	0.5086	0.0247	0.5087	0.0208
Percent students on free school meals (2002-2007)	0.2008	0.1491	0.1895	0.1353	0.1840	0.1265
KS2 enrolment (2002-2007)	47.5165	24.1886	48.5633	25.4531	48.7008	26.3776
Panel D. Instruments						
Percent noise at L5	0.1890	0.0781	0.1953	0.0642	0.1987	0.0541
Percent noise at L4	0.1654	0.1049	0.1727	0.0890	0.1774	0.0769
Percent noise at L3	0.0603	0.3334	0.0721	0.2781	0.0816	0.2851
Noise to signal at L5	0.0578	0.0310	0.0574	0.0242	0.0570	0.0199
Noise to signal at L4	0.0108	0.0082	0.0105	0.0067	0.0104	0.0056
Noise to signal at L3	0.0010	0.0028	0.0010	0.0022	0.0010	0.0018
Counterfactual density at L5	0.0750	0.0141	0.0741	0.0121	0.0736	0.0110
Counterfactual density at L4	0.0450	0.0137	0.0428	0.0115	0.0415	0.0102
Counterfactual density at L3	0.0149	0.0092	0.0129	0.0074	0.0118	0.0064
Number of blocks	23260		27414		28621	

Note. Means and standard deviations across blocks. Only blocks with at least one school within a 0.4-mile radius are retained in columns (1) to (2). Alternative samples using 0.6-mile and 0.8-mile radiuses are considered in columns (3) to (4) and (5) to (6), respectively. Numbers in brackets next to variable names represent the period over which means and standard deviations are computed.

Table 2: Discontinuities at achievement theresholds

	deviations from thresholds:					missing mass
	-3	-2	-1	0	1	
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A. Real thresholds			
Level 3	-0.0001	-0.0002	-0.0007***	0.0008***	0.0002	0.0010***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)
Level 4	-0.0018***	-0.0019***	-0.0036***	0.0062***	0.0011*	0.0073***
	(0.0005)	(0.0005)	(0.0006)	(0.0007)	(0.0006)	(0.0008)
Level 5	-0.0046***	-0.0036***	-0.0067***	0.0137***	0.0012***	0.0149***
	(0.0006)	(0.0005)	(0.0006)	(0.0010)	(0.0004)	(0.0009)
			Panel B. Placebo thresholds			
Level 3	0.0000	-0.0000	0.0000	-0.0000	0.0000	0.0001
	(0.0001)	(0.0001)	(0.0000)	(0.0001)	(0.0001)	(0.0001)
Level 4	0.0000	-0.0000	-0.0000	-0.0000	0.0000	0.0000
	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
Level 5	-0.0001	-0.0001	-0.0000	-0.0000	0.0002	0.0002
	(0.0004)	(0.0003)	(0.0003)	(0.0002)	(0.0002)	(0.0003)

Note. Differences between observed and counterfactual score densities from equation (1). Scores range from -3 (three points below threshold, in column (1)) to 1 (one point above threshold, in column (5)). Column (6) reports estimates of the notch induced by bordelining. Panel A is obtained using thresholds (at Level 3, Level 4 and Level 5) used by external markers. Panel B considers placebo thresholds centred ten points below critical scores (see Section 4 for details). Standard errors are shown in brackets. * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Discontinuities at achievement theresholds

	deviations from Level 4 threshold:					missing mass	deviations from Level 5 threshold:					missing mass
	-3	-2	-1	0	1		-3	-2	-1	0	1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A. Student characteristics (NPD data)												
Male	0.0011	-0.0003	-0.0002	-0.0008	0.0002	0.0006	-0.0001	-0.0012	-0.0008	0.0009	0.0012	0.0021
	(0.0022)	(0.0026)	(0.0028)	(0.0019)	(0.0019)	(0.0027)	(0.0026)	(0.0027)	(0.0020)	(0.0017)	(0.0019)	(0.0026)
White	0.0007	0.0011	-0.0016	-0.0018	0.0016	0.0002	0.0002	0.0010	0.0001	-0.0003	-0.0010	0.0013
	(0.0013)	(0.0014)	(0.0014)	(0.0012)	(0.0011)	(0.0015)	(0.0007)	(0.0008)	(0.0009)	(0.0008)	(0.0009)	(0.0011)
Free school meals	-0.0011	0.0020	-0.0002	-0.0000	-0.0008	0.0008	-0.0009	-0.0012	0.0011	0.0009	0.0001	0.0010
	(0.0019)	(0.0019)	(0.0012)	(0.0014)	(0.0015)	(0.0019)	(0.0011)	(0.0010)	(0.0009)	(0.0007)	(0.0008)	(0.0011)
English speaking	0.0004	0.0005	-0.0009	-0.0009	0.0009	0.0000	0.0003	0.0003	0.0004	-0.0003	-0.0007	0.0010
	(0.0011)	(0.0013)	(0.0010)	(0.0009)	(0.0009)	(0.0012)	(0.0005)	(0.0006)	(0.0006)	(0.0006)	(0.0008)	(0.0008)
Panel B. School characteristics (NPD data)												
Independent	-0.0026***	-0.0005	0.0022**	0.0007	0.0002	0.0009	0.0018*	-0.0021*	0.0012	-0.0006	-0.0003	0.0009
	(0.0010)	(0.0010)	(0.0011)	(0.0009)	(0.0011)	(0.0013)	(0.0009)	(0.0012)	(0.0008)	(0.0007)	(0.0010)	(0.0011)
KS2 grade enrolment	0.0303	-0.1147	-0.1774	0.2168**	0.0450	0.2618*	0.0607	-0.0992	-0.2805*	0.2732***	0.0459	0.3190***
	(0.1302)	(0.1255)	(0.1538)	(0.0962)	(0.0911)	(0.1361)	(0.0751)	(0.1084)	(0.1432)	(0.0854)	(0.0705)	(0.1128)
Independent KS2 grade enrolment	0.0504	0.1187	-0.3429	0.1357	0.0382	0.1739	0.0302	-0.1063	-0.3893**	0.4223***	0.0431	0.4654***
	(0.2844)	(0.2083)	(0.3363)	(0.2164)	(0.2093)	(0.2936)	(0.1504)	(0.1544)	(0.1808)	(0.1286)	(0.1366)	(0.1767)
Panel C. Area characteristics (2001 census)												
Pct white in the area where student lives	0.0009	-0.0000	-0.0009	-0.0006	0.0006	0.0000	0.0001	0.0003	0.0003	-0.0001	-0.0006	0.0007
	(0.0006)	(0.0006)	(0.0007)	(0.0006)	(0.0005)	(0.0007)	(0.0003)	(0.0004)	(0.0003)	(0.0003)	(0.0004)	(0.0005)
Pct high qualified in the area where student lives	-0.0001	-0.0002	-0.0001	0.0001	0.0003	0.0000	0.0000	-0.0001	0.0002	0.0000	-0.0001	0.0004
	(0.0004)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0004)	(0.0004)	(0.0002)	(0.0004)	(0.0003)	(0.0003)	(0.0004)
Pct unemployed in the area where student lives	-0.0002	0.0001	-0.0000	0.0003**	-0.0003	0.0000	-0.0001	0.0000	0.0000	-0.0000	0.0001	0.0000
	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0000)

Note. Differences between observed and counterfactual densities from equation (1) using student, school and area characteristics (see Section 4 for details). Level 4 and Level 5 thresholds are in columns (1) to (6) and (7) to (12), respectively. Scores range from -3 (three points below threshold, in columns (1) and (7)) to 1 (one point above threshold, in columns (5) and (11)). Columns (6) and (12) report estimates of the notch induced by bordelining. Panel A considers student characteristics from NPD data. Panel B considers school characteristics from NPD data. Panel C considers area characteristics from the 2001 census. Standard errors are shown in brackets. * p<0.10, ** p<0.05, *** p<0.01.

Table 4: First Stage regressions

	Average Point Score			Math Level 5	
	(1)	(2)	(3)	(4)	(5)
			Panel A. Schools within 0.4 mile		
Percent noise at L5	0.115***	0.083***	0.082***	0.094***	0.066***
	(0.012)	(0.008)	(0.008)	(0.012)	(0.010)
Percent noise at L4			0.011		
			(0.009)		
Number of blocks	23,260	23,260	23,260	23,260	23,260
			Panel B. Schools within 0.6 mile		
Percent noise at L5	0.109***	0.075***	0.074***	0.091***	0.061***
	(0.014)	(0.008)	(0.008)	(0.014)	(0.009)
Percent noise at L4			0.015*		
			(0.009)		
Number of blocks	27,414	27,414	27,414	27,414	27,414
			Panel C. Schools within 0.8 mile		
Percent noise at L5	0.082***	0.058***	0.057***	0.066***	0.046***
	(0.013)	(0.009)	(0.009)	(0.013)	(0.011)
Percent noise at L4			0.021**		
			(0.008)		
Number of blocks	28,621	28,621	28,621	28,621	28,621
MSOA fixed effects	✗	✗	✗	✗	✗
Controls		✗	✗		✗
Over-identified model			✗		

Note. OLS regressions of average point score (columns (1) to (3)) and percentage of students attaining at level 5 in math (columns (4) to (5)). Instruments and outcomes are standardised to have zero mean and unit variance. Blocks are the statistical units of analysis. Panel A presents results using all schools within a 0.4-mile radius from a block's centroid; Panel B and Panel C consider a 0.6-mile and 0.8-mile radius, respectively. All columns control for a full set of neighbourhood (MSOA) fixed effects. Columns (2), (3) and (5) add a quadratic polynomial in the running variable, a quadratic polynomial in distance from the closest school, number of schools within the radius and their size, and population density. All columns use the instrument at the L5 threshold. Column (3) shows results from an over-identified model in which L4 and L5 instruments are used jointly. Standard errors, shown in brackets, are clustered on Local Authority. * p<0.10, ** p<0.05, *** p<0.01.

Table 5: OLS and IV regressions of school quality on house prices

	Average Point Score			Math Level 5	
	(1)	(2)	(3)	(4)	(5)
	OLS	2SLS	2SLS	OLS	2SLS
			Panel A. Schools within 0.4 mile		
Log prices 2008-2011	0.042***	0.063***	0.061***	0.035***	0.079***
	(0.003)	(0.014)	(0.014)	(0.003)	(0.018)
Property tax (2011)	0.032***	0.024*	0.020	0.031***	0.030*
	(0.004)	(0.014)	(0.014)	(0.004)	(0.018)
Number of blocks	23,260	23,260	23,260	23,260	23,260
			Panel B. Schools within 0.6 mile		
Log prices 2008-2011	0.041***	0.065***	0.059***	0.037***	0.080***
	(0.004)	(0.019)	(0.019)	(0.003)	(0.024)
Property tax (2011)	0.032***	0.042**	0.034*	0.033***	0.052**
	(0.004)	(0.018)	(0.018)	(0.003)	(0.022)
Number of blocks	27,414	27,414	27,414	27,414	27,414
			Panel C. Schools within 0.8 mile		
Log prices 2008-2011	0.038***	0.059**	0.021	0.035***	0.075**
	(0.005)	(0.029)	(0.025)	(0.004)	(0.038)
Property tax (2011)	0.026***	0.025	0.000	0.028***	0.032
	(0.003)	(0.023)	(0.020)	(0.003)	(0.030)
Number of blocks	28,621	28,621	28,621	28,621	28,621
MSOA fixed effects	✗	✗	✗	✗	✗
Controls	✗	✗	✗	✗	✗
Over-identified model			✗		

Note. OLS and IV regressions of outcomes on standardized average point score (columns (1) to (3)) and standardized percentage of students attaining at level 5 in math (columns (4) to (5)). Blocks are the statistical units of analysis. Panel A presents results using all schools within a 0.4-mile radius from a block's centroid; Panel B and Panel C consider a 0.6-mile and 0.8-mile radius, respectively. All columns control for a full set of neighbourhood (MSOA) fixed effects, a quadratic polynomial in the running variable, a quadratic polynomial in distance from the closest school, number of schools within the radius and their size, and population density. All columns use the instrument at the L5 threshold. Column (3) shows results from an over-identified model in which L4 and L5 instruments are used jointly. Standard errors, shown in brackets, are clustered on Local Authority. * p<0.10, ** p<0.05, *** p<0.01.

Table 6: OLS and IV regressions for area characteristics

	Average Point Score			Math level 5	
	(1)	(2)	(3)	(4)	(5)
	OLS	2SLS	2SLS	OLS	2SLS
	Panel A. Schools within 0.4 mile				
Pct unemployed	-0.005***	-0.006***	-0.006***	-0.003***	-0.008***
	(0.000)	(0.002)	(0.002)	(0.000)	(0.002)
Pct high qualified	0.021***	0.027***	0.026***	0.018***	0.034***
	(0.002)	(0.006)	(0.006)	(0.002)	(0.007)
Pct managers	0.006***	0.008***	0.007***	0.005***	0.010***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.003)
Pct in good health	0.005***	0.014***	0.013***	0.005***	0.018***
	(0.001)	(0.004)	(0.004)	(0.001)	(0.005)
Number of blocks	22,687	22,687	22,687	22,687	22,687
	Panel B. Schools within 0.6 mile				
Pct unemployed	-0.003***	-0.004**	-0.004**	-0.002***	-0.005**
	(0.000)	(0.002)	(0.002)	(0.000)	(0.002)
Pct high qualified	0.022***	0.032***	0.029***	0.019***	0.039***
	(0.002)	(0.009)	(0.009)	(0.001)	(0.010)
Pct managers	0.006***	0.008***	0.006**	0.005***	0.009***
	(0.001)	(0.003)	(0.003)	(0.001)	(0.004)
Pct in good health	0.006***	0.009	0.008	0.005***	0.010
	(0.001)	(0.006)	(0.006)	(0.001)	(0.007)
Number of blocks	26,735	26,735	26,735	26,735	26,735
	Panel C. Schools within 0.8 mile				
Pct unemployed	-0.003***	-0.003	-0.002	-0.002***	-0.004
	(0.000)	(0.002)	(0.002)	(0.000)	(0.003)
Pct high qualified	0.020***	0.034***	0.019*	0.018***	0.043***
	(0.002)	(0.012)	(0.010)	(0.001)	(0.016)
Pct managers	0.005***	0.007	0.001	0.005***	0.009*
	(0.001)	(0.004)	(0.004)	(0.001)	(0.005)
Pct in good health	0.005***	0.011	0.006	0.005***	0.013
	(0.001)	(0.007)	(0.007)	(0.001)	(0.009)
Number of blocks	27,901	27,901	27,901	27,901	27,901
MSOA fixed effects	✗	✗	✗	✗	✗
Controls	✗	✗	✗	✗	✗
Over-identified model			✗		

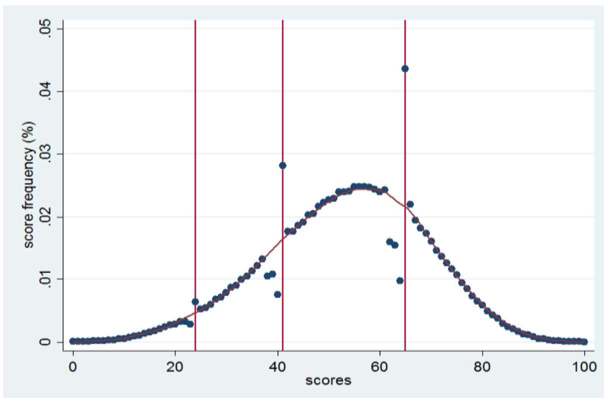
Note. OLS and IV regressions of outcomes on standardized average point score (columns (1) to (3)) and standardized percentage of students attaining at level 5 in math (columns (4) to (5)). Blocks are the statistical units of analysis. Panel A presents results using all schools within a 0.4-mile radius from a block's centroid; Panel B and Panel C consider a 0.6-mile and 0.8-mile radius, respectively. All columns control for a full set of neighbourhood (MSOA) fixed effects, a quadratic polynomial in the running variable, a quadratic polynomial in distance from the closest school, number of schools within the radius and their size, and population density. All columns use the instrument at the L5 threshold. Column (3) shows results from an over-identified model in which L4 and L5 instruments are used jointly. Standard errors, shown in brackets, are clustered on Local Authority. * p<0.10, ** p<0.05, *** p<0.01.

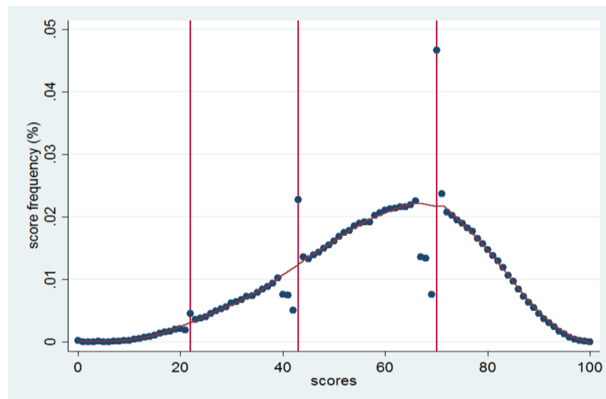
Table 7: OLS and IV regressions for school outcomes

	Average Point Score			Math level 5	
	(1)	(2)	(3)	(4)	(5)
	OLS	2SLS	2SLS	OLS	2SLS
	Panel A. Schools within 0.4 mile				
Pct Level 5 math	0.092***	0.071***	0.067***	0.076***	0.085***
	(0.003)	(0.016)	(0.016)	(0.002)	(0.019)
Pct on Free School Meals	-0.056***	-0.058***	-0.054**	-0.038***	-0.070***
	(0.004)	(0.021)	(0.021)	(0.003)	(0.026)
Pct black	-0.005**	0.000	0.001	-0.006***	0.000
	(0.002)	(0.009)	(0.009)	(0.002)	(0.011)
Pct with special needs	-0.040***	-0.061**	-0.060**	-0.028***	-0.074**
	(0.003)	(0.026)	(0.026)	(0.002)	(0.032)
Number of blocks	22,351	22,351	22,351	22,351	22,351
	Panel B. Schools within 0.6 mile				
Pct Level 5 math	0.084***	0.076***	0.072***	0.070***	0.087***
	(0.003)	(0.014)	(0.014)	(0.002)	(0.018)
Pct on Free School Meals	-0.045***	-0.058***	-0.053***	-0.032***	-0.067***
	(0.003)	(0.017)	(0.017)	(0.002)	(0.021)
Pct black	-0.005***	-0.006	-0.007	-0.005***	-0.007
	(0.002)	(0.010)	(0.011)	(0.001)	(0.012)
Pct with special needs	-0.032***	-0.051***	-0.051***	-0.025***	-0.058***
	(0.003)	(0.019)	(0.019)	(0.002)	(0.022)
Number of blocks	26,936	26,936	26,936	26,936	26,936
	Panel C. Schools within 0.8 mile				
Pct Level 5 math	0.077***	0.058***	0.049***	0.064***	0.070***
	(0.003)	(0.016)	(0.016)	(0.002)	(0.022)
Pct on Free School Meals	-0.039***	-0.047**	-0.038*	-0.030***	-0.057**
	(0.003)	(0.020)	(0.020)	(0.002)	(0.025)
Pct black	-0.004***	-0.001	-0.001	-0.003***	-0.001
	(0.001)	(0.006)	(0.006)	(0.001)	(0.008)
Pct with special needs	-0.030***	-0.063***	-0.063***	-0.024***	-0.076**
	(0.002)	(0.023)	(0.021)	(0.002)	(0.031)
Number of blocks	28,380	28,380	28,380	28,380	28,380
MSOA fixed effects	✗	✗	✗	✗	✗
Controls	✗	✗	✗	✗	✗
Over-identified model			✗		

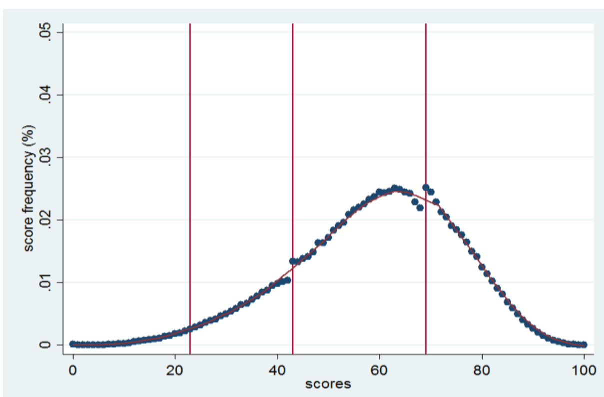
Note. OLS and IV regressions of outcomes on standardized average point score (columns (1) to (3)) and standardized percentage of students attaining at level 5 in math (columns (4) to (5)). Blocks are the statistical units of analysis. Panel A presents results using all schools within a 0.4-mile radius from a block's centroid; Panel B and Panel C consider a 0.6-mile and 0.8-mile radius, respectively. All columns control for a full set of neighbourhood (MSOA) fixed effects, a quadratic polynomial in the running variable, a quadratic polynomial in distance from the closest school, number of schools within the radius and their size, and population density. All columns use the instrument at the L5 threshold. Column (3) shows results from an over-identified model in which L4 and L5 instruments are used jointly. Standard errors, shown in brackets, are clustered on Local Authority. * p<0.10, ** p<0.05, *** p<0.01.

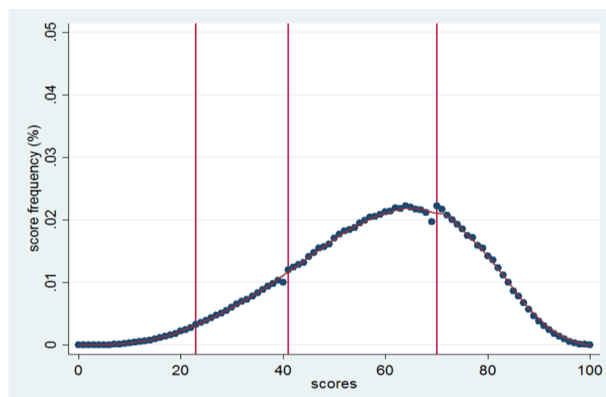
Figure 1: Score manipulation over time: English





Panel A: 2004 **Panel B: 2007**

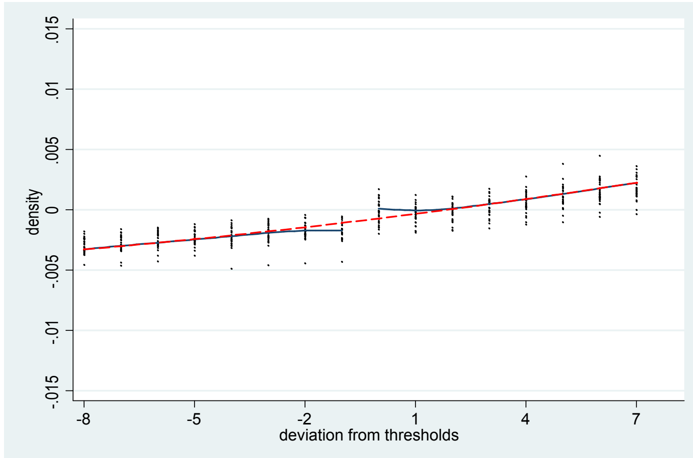




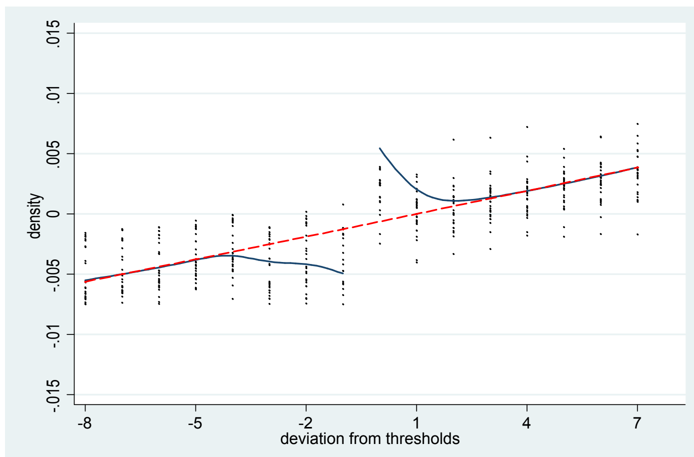
Panel C: 2008 **Panel D: 2011**

Note. This figure shows Key Stage 2 score distributions for English in selected years before (2004 and 2007) and after (2008 and 2011) the removal of borderlining. In each panel, the vertical lines are critical thresholds set in that year. The continuous line is a local linear regression fit obtained excluding observations in the [-3,2] windows around thresholds.

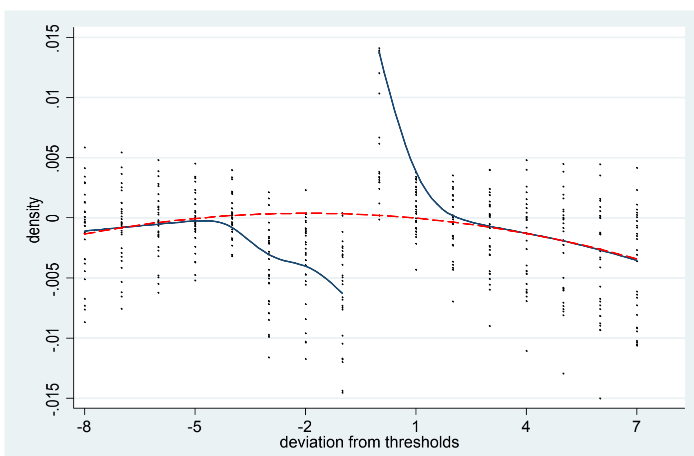
Figure 2: Bunching around achievement thresholds



Panel A: Level 3



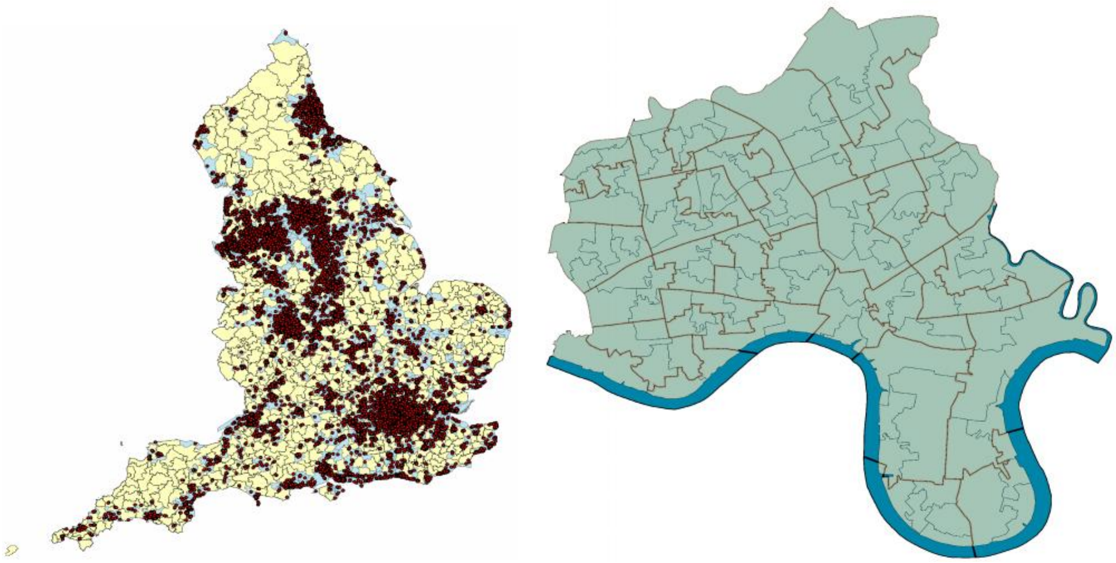
Panel B: Level 4



Panel C: Level 5

Note. This figure plots residuals of test score frequencies around Level 3 threshold (Panel A), Level 4 threshold (Panel B) and Level 5 threshold (Panel C). Continuous lines are fitted values obtained with local linear regressions (LLR). The window around the cutoffs is chosen so that consecutive windows do not overlap. Dashed lines represent the test score distribution that would have been observed in the absence of bunching around achievement thresholds. See text for more details.

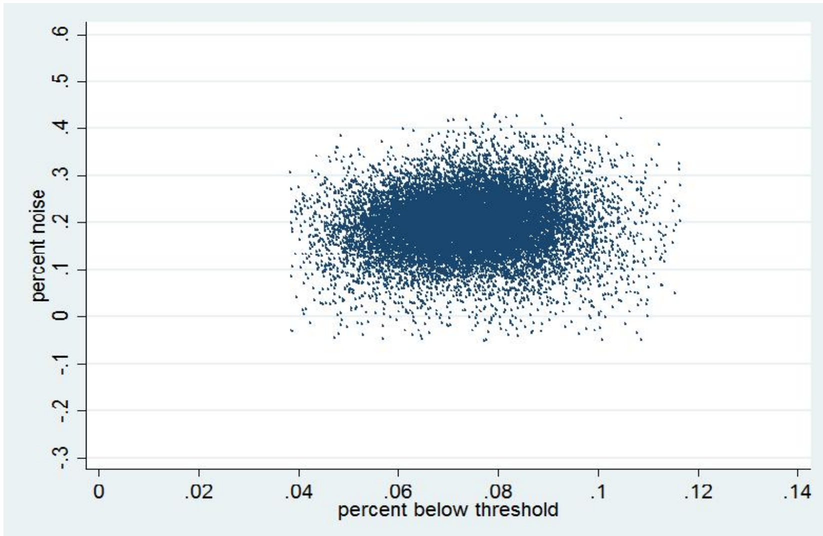
Figure 3: Geography and sample selection



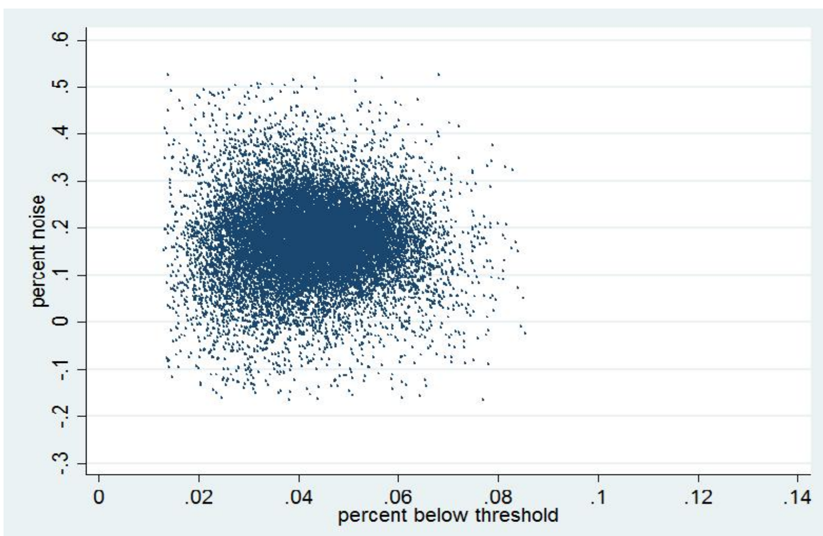
Panel A **Panel B**

Note. Panel A shows urban MSOAs in England (light blue areas) with superimposed the corresponding listing of schools (red dots). Panel B shows the geographic hierarchy defined by LSOAs and MSOAs for the London borough of Tower Hamlets, comprising 31 MSOAs (brown outline) and 130 LSOAs (grey outline).

Figure 4: Noise in school quality



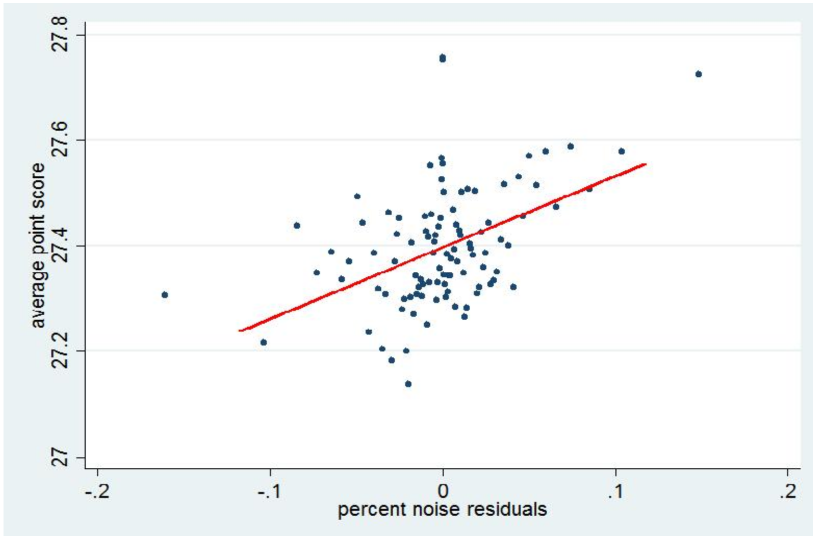
Panel A: Level 5



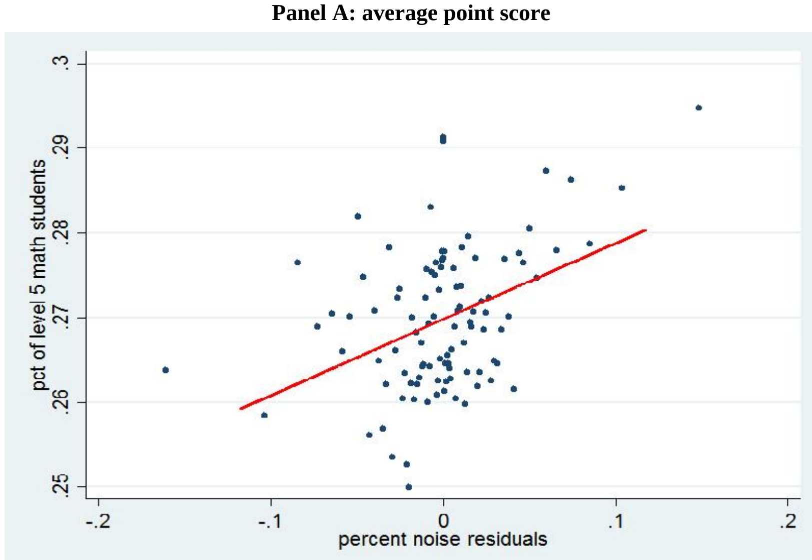
Panel B: Level 4

Note. The figure shows how noise in school quality (vertical axis) varies with percent of students below the critical threshold (horizontal axis). Figures are drawn considering a 0.6 mile of radius. Panel A and Panel B refer to Level 5 and Level 4 thresholds, respectively.

Figure 5: First stage graphs



Panel A: average point score



Panel B: pct of level 5 math students

Note. This figure shows visual interpretation of first stage regressions for average point score (Panel A) and percentage of students awarded level 5 in math (Panel B). Both panels consider the instrument defined using the L5 threshold and are drawn considering a 0.6 mile radius. Regressions control for a full set of neighbourhood (MSOA) fixed effects, a quadratic polynomial in the running variable, a quadratic polynomial in distance from the closest school, number of schools within the radius and their size, and population density, as detailed in the text.

References

Angrist, J. D., Battistin, E., and Vuri, D. (2016). In a small moment: class size and moral hazard in the mezzogiorno. *American Economic Journal: Applied Economics*.

Bayer, P., Ferreira, F., and McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4):588–638.

Black, S. E. (1999). Do better schools matter? parental valuation of elementary education. *The Quarterly Journal of Economics*, 114(2):577–599.

Black, S. E. and Machin, S. (2011). Housing valuations of school performance. *Handbook of the Economics of Education*, 3:485–519.

Burgess, S. and Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31(3):535–576.

Burgess, S., Greaves, E., Vignoles, A., and Wilson, D. (2015). What parents want: school preferences and school choice. *The Economic Journal*, 125(587):1262–1289.

Caetano, G. (2015). Neighborhood sorting and the valuation of public school quality. Working paper, University of Rochester.

DCSF (2009, Revised). National curriculum assessments at key stage 2 in england. *Statistical First Release 32*.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. (2016). The causes and consequences of test score manipulation: evidence from the new york regents examinations. NBER WP 22165.

DfE (2012). Schools, pupils and their characteristics, january 2012. *Statistical First Release 10*.

DfE (2015). Schools, pupils and their characteristics, january 2015. *Statistical First Release 16*.

Diamond, R. and Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. SIEPR Discussion Paper No. 16-003.

Duranton, G. and Puga, D. (2015). Urban land use. *Handbook of Regional and Urban Economics*, 5:467–560.

Ebenstein, A., Lavy, V., and Roth, S. (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.

Eyles, A. and Machin, S. (2015). The introduction of academy schools to england's education. *CEP Discussion Paper No. 1368*.

Fack, G. and Grenet, J. (2010). When do better schools raise housing prices? evidence from paris public and private schools. *Journal of Public Economics*, 94:59–77.

Figlio, D. N. and Lucas, M. E. (2004). What's in a grade? school report cards and the housing market. *American Economic Review*, pages 591–604.

Gibbons, S. and Machin, S. (2003). Valuing english primary schools. *Journal of Urban Economics*, 53:197–219.

Gibbons, S., Machin, S., and Silva, O. (2013). Valuing school quality using boundary discontinuities. *Journal of Urban Economics*, 75.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–168.

Hastings, J. S. and Weinstein, J. M. (2008). Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics*, 123(4):1373–1414.

Jacob, B. A. and Levitt, S. D. (2003). Rotten apples: an investigation of the prevalence and predictors of teacher cheating.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Public Economics*, 92:2083–2105.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review*, 99(5):1979–2011.

Lavy, V. and Sand, E. (2015). On the origins of gender human capital gaps: short and long term consequences of teachers' stereotypical biases. NBER Working Paper 20909.

Machin, S. (2011). Houses and schools: valuation of school quality through the housing market. *Labour Economics*, 18:723–729.

Neal, D. and Schanzenbach, D. W. (2010). Left behind by design: proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2):263–283.

Ries, J. and Somerville, T. (2010). School quality and residential property values: evidence from vancouver rezoning. *The Review of Economics and Statistics*, 92(4):928–944.

Tannenbaum, D. I. (2015). Does school quality affect neighborhood development? evidence from a redistricting reform. Working paper.

Terrier, C. (2016). Boys lag behind: How teachers' gender biases affect student achievement. *IZA Discussion Paper No. 10343*.

Wilson, D., Croxson, B., and Atkinson, A. (2006). What gets measured gets done. head teachers' responses to the english secondary school performance management system. *Policy Studies*, 27(2):153–71.

Appendix A. Sample Selection

Geographic hierarchies and area selection

The analysis is limited to Middle Layer Super Output Areas (MSOAs) located in metropolitan counties and Greater London, and urban MSOAs in non-metropolitan counties. MSOAs are a geographic hierarchy developed by the Office for National Statistics (ONS) consisting of 7,194 homogenous areas (6,781 in England and 413 in Wales), with a minimum population of 5,000 (an average of 7,200) and a minimum resident household of 2,000 (an average of 3,000). Metropolitan and non-metropolitan counties, and the region of Greater London, are official administrative subdivisions in England. The final sample of areas consists of 6,133 MSOAs in England (90.44% of the total in the country). The listing of MSOAs is obtained according to the following steps.

- There are six metropolitan counties, typically with populations of 1.2 to 2.8 million: Greater Manchester, Merseyside (e.g., Liverpool), South Yorkshire (e.g., Sheffield), Tyne and Wear (e.g., Newcastle), West Midlands (e.g., Birmingham) and West Yorkshire (e.g., Leeds and Bradford). We keep all 1,499 MSOAs in this group of areas (24.44% of the final sample).

- Greater London is the region comprising districts around London, and its structure is similar to that of metropolitan counties. We keep all 983 MSOAs in this region (16.03% of the final sample).

- Non-metropolitan counties consists of areas not in the two points above. The definition of rurality follows from the official classification of Output Areas (OAs) published by the ONS, and based on land use. We classify as rural those MSOAs with the majority of OAs (above 50%) falling within the "Small Town and Fringe areas", "Village" or "Hamlet and Isolated Dwelling" categories, and whose surrounding areas are sparsely populated. This classification leaves us with 3,651 urban MSOAs in this group of areas (59.53% of the final sample).

Selection of schools and catchment areas

We consider community, faith and foundation public primary schools in MSOAs identified in the previous section. There are 12,973 of such schools in the register of educational establishments provided by the Department for Education in the years 1998 to 2007. This is the period relevant to the construction of the effects of borderlining on school quality. We drop a limited number of special schools (e.g., pupil referral units) specifically organized to provide education to children with special needs (excluded, sick, or unable to follow the mainstream curriculum). Number and composition of schools are substantially stable over the period considered, without major mergers or institutional changes (e.g., tranformation from community to autonomous, or into academies).[17] The left hand side panel of Figure 3 shows the areas of England selected, together with the listing of schools considered.

The neighbourhood definition used in the analysis is restricted to Lower Layer Super Output Areas (LSOAs) of school catchments. The right hand side panel of Figure 3 shows the ONS geography of LSOAs for the borough of Tower Hamlets in East London. We consider houses and amenities located in LSOAs selected by applying the following criteria.

- We compute distance from centroids of all LSOAs to the closest school, and keep LSOAs with at least one school of the same LA within 0.6 miles (the 60th percentile of the student-school distance in NPD data). This is the sample described in Section 3. The corresponding sample size gradient implied by the various selection steps is in Table A.1.

- The same procedure is replicated using a .4-mile and .8-mile radius (50th and 75th percentiles from the NPD student-school distance distribution, respectively). The sensitivity of our findinds in Section 5 is investigated considering these samples, which consist of 23,260 and 28,621 LSOAs, respectively.

Our sample cut implies that the same LSOA might belong to the catchment area of multiple schools. This is important for the computation of schools z_{bn}, as discussed in Appendix B.

[17] The Edubase database, which can be accessed at http://www.education.gov.uk/edubase/home.xhtml, provides additional information on these institutional changes. There were about 203 academies in England until 2010, mainly at secondary school, and their number grew in the last five years. Currently, 2,075 out of 3,381 secondary schools are academies, while 2,440 of 16,766 primary schools have academy status.

Table A.1: Sample selection criteria

	Census blocks (1)	Schools (2)	Students (3)
In England	32,482	17,961	5,874,230
Drop private schools and school with special status		15,711	5,732,873
Schools with at least one student around thresholds		15,704	5,732,818
Schools with non-missing geographical data		15,600	5,702,603
Census blocks:			
- in metropolitan counties	7,183		
- in the Greater London region	4,765		
- in urban non-metropolitan counties	17,676		
In the three areas above	29,624	12,973	5,241,482
With at least one school within:			
- 0.8 miles	28,621	12,694	5,226,434
- 0.6 miles	27,414	12,481	5,187,610
- 0.4 miles	23,260	12,147	5,099,800

Note. This table shows the selection criteria applied to define the working sample. Column (1) shows the number of Census blocks left after each step; column (2) shows the number of schools left; column (3) shows the number of students left.

Appendix B. Effects of borderlining on school quality

The variable z_{bn} is defined to proxy the effects of borderlining on test scores and school quality measurements published in Performance tables. The variable is constructed by estimating equation (1) for each LSOA, pooling tests for all schools having that LSOA in their catchment (as explained in Appendix A). To gain precision, we estimate a LSOA's missing mass using percentage of students f_{scj} scoring $s \in [-8, 7]$ around cutoff c (Level 3, Level 4 and Level 5) for subject j (English, math and science) pooling (a) tests from 1998 to 2007, and (b) all schools associated to the LSOA within a certain radius (see Appendix A). The following equation is estimated for each LSOA:

$$f_{scj} = \alpha(j) + \sum_{i=0}^{2} \beta_i s^i + \sum_{i=-3}^{1} \gamma_i 1(s = i) + \varepsilon_{scj}, \tag{4}$$

by attainment cutoff. This is a variant to the specification of equation (1) discussed in the main text. The value $\sum_{i=-3}^{-1} \gamma_i$ is then calculated, which represents our estimate of the notch induced by bordelining for schools having the LSOA considered in their catchment. We get z_{bn} by iterating this procedure over all LSOAs selected in Appendix A.

Panel A of Table B.1 reports summary statistics for the value $\sum_{i=-3}^{-1} \gamma_i$ at Level 3, Level 4 and Level 5 - see columns (1), (3) and (5), respectively. We also present in columns (2), (4) and (6) summary statistics for p-values associated to these estimates. Using the same format, Panel B of the same table considers $\sum_{s=-3}^{-1} \sum_{i=0}^{2} \beta_i s^i$, the "running variable" as defined in section 5.

The effect of borderlining should be zero, by construction, at any point in the distribution three marks away from achievement thresholds. It follows that z_{bn} should be centered at zero when estimated from (4) away from critical points. We replicate the procedure above using percentage of students f_{scj} scoring $s \in [-8, 7]$ in a window centered at scores away from relevant cutoffs, and present summary statistics in Table B.2. Fictitious cutoff are located 10 scores below math level thresholds, and 9 scores below English and science level threshold. A juxtaposition with Table B.1 reveals the expected pattern, with notches estimated below fictitious cutoffs being centered at zero at Level 3, 4 and 5.[18]

[18] Fictitious cutoffs are chosen to avoid having the true thresholds in this window.

Table B.1: Summary statistics for bunching estimates

	Level 3		Level 4		Level 5	
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A			
Percentile	notch	p-value	notch	p-value	notch	p-value
1	-0.005	0.000	-0.004	0.000	-0.002	0.000
5	-0.003	0.004	0.000	0.000	0.004	0.000
10	-0.002	0.015	0.001	0.001	0.006	0.000
25	-0.001	0.090	0.004	0.005	0.010	0.000
50	0.001	0.328	0.007	0.032	0.014	0.002
75	0.002	0.642	0.010	0.169	0.018	0.025
90	0.004	0.853	0.014	0.514	0.023	0.148
99	0.009	0.986	0.022	0.948	0.033	0.842
Mean	0.001	0.382	0.007	0.146	0.014	0.056
			Panel B			
	rv	p-value	rv	p-value	rv	p-value
1	0.000	0.000	0.017	0.000	0.042	0.000
5	0.005	0.000	0.024	0.000	0.052	0.000
10	0.006	0.000	0.029	0.000	0.058	0.000
25	0.009	0.000	0.036	0.000	0.067	0.000
50	0.014	0.000	0.045	0.000	0.076	0.000
75	0.020	0.000	0.054	0.000	0.084	0.000
90	0.027	0.004	0.062	0.000	0.092	0.000
99	0.048	0.406	0.079	0.000	0.109	0.000
Mean	0.015	0.012	0.046	0.000	0.076	0.000

Note. This table shows estimates for the notch induced by the borderlining (Panel A) and the running variable (Panel B), as defined in the text. Columns (2), (4) and (6) present summary statistics for p-values associated to these estimates.

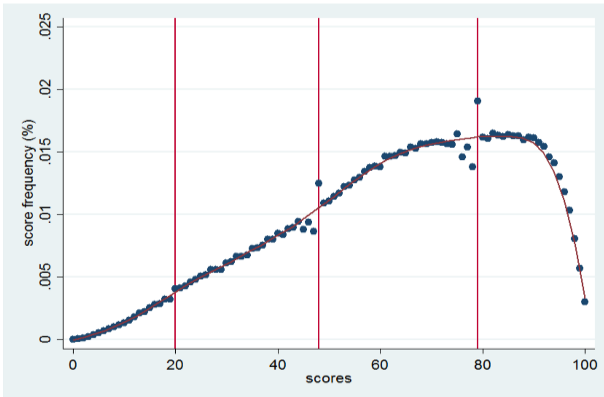
Table B.2: Summary statistics for bunching estimates with fictitious cutoffs

	Level 3		Level 4		Level 5	
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A			
Percentile	notch	p-value	notch	p-value	notch	p-value
1	-0.006	0.001	-0.008	0.001	-0.012	0.001
5	-0.003	0.011	-0.005	0.011	-0.007	0.013
10	-0.002	0.034	-0.004	0.032	-0.005	0.037
25	-0.001	0.148	-0.002	0.131	-0.002	0.149
50	0.000	0.401	0.000	0.372	0.000	0.402
75	0.000	0.689	0.002	0.679	0.003	0.689
90	0.002	0.880	0.004	0.874	0.006	0.876
99	0.006	0.992	0.009	0.989	0.014	0.987
Mean	0.000	0.431	0.000	0.415	0.000	0.428
			Panel B			
	rv	p-value	rv	p-value	rv	p-value
1	-0.006	0.000	0.009	0.000	0.052	0.000
5	0.000	0.000	0.014	0.000	0.059	0.000
10	0.000	0.000	0.016	0.000	0.061	0.000
25	0.003	0.000	0.022	0.000	0.066	0.000
50	0.006	0.002	0.030	0.000	0.070	0.000
75	0.010	0.043	0.039	0.000	0.075	0.000
90	0.016	0.319	0.047	0.000	0.080	0.000
99	0.076	0.917	0.067	0.004	0.096	0.000
Mean	0.009	0.086	0.031	0.001	0.071	0.000

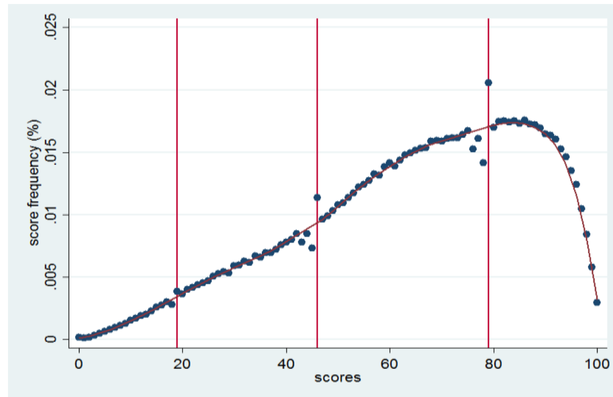
Note. This table shows estimates for the notch computed at fictitious thresholds. Fictitious thresholds are located by construction 10 scores below level thresholds for math, and 9 scores below level thresholds for English and science. Panel A (Columns (1), (3) and (5)) shows summary statistics for the notch computed considering scores in the window [-3,-1] below fictitious cutoff score. Panel B shows corresponding estimates for the running variable. Columns (2), (4) and (6) present summary statistics for p-values associated to these estimates.

Appendix C

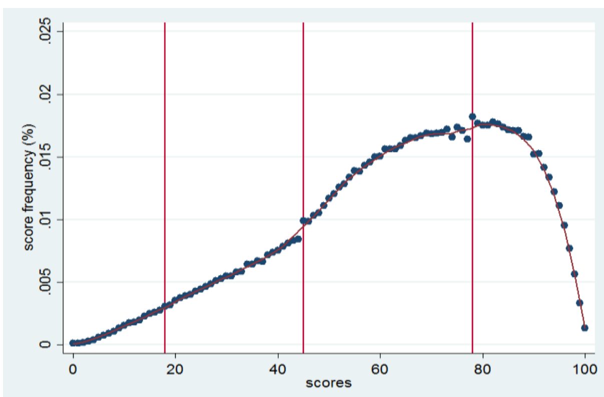
Figure C.1: Score manipulation over time: Math



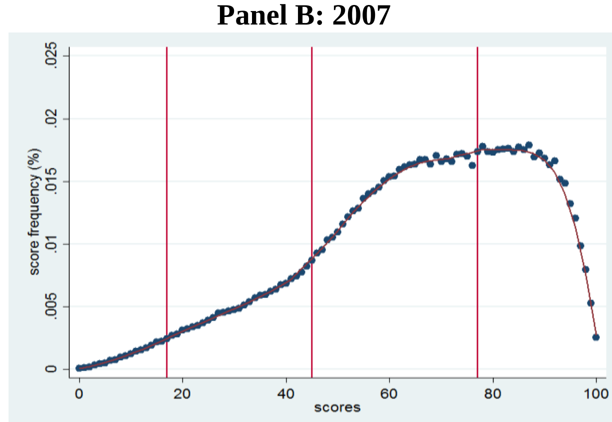
Panel A: 2004



Panel B: 2007



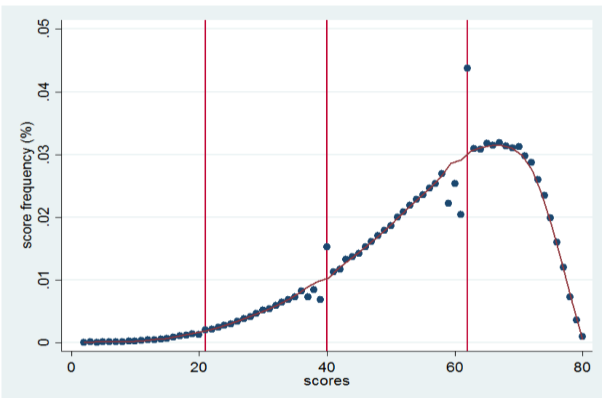
Panel C: 2008



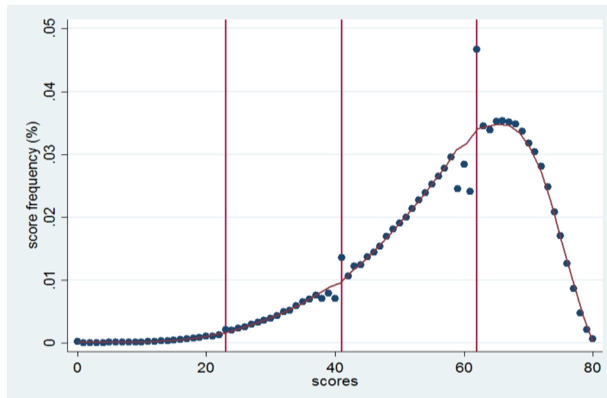
Panel D: 2011

Note. This figure shows Key Stage 2 score distributions for math in selected years before (2004 and 2007) and after (2008 and 2011) the removal of borderlining. In each panel, the vertical lines are critical thresholds set in that year. The continuous line is a local linear regression fit obtained excluding observations in the [-3,2] windows around thresholds.

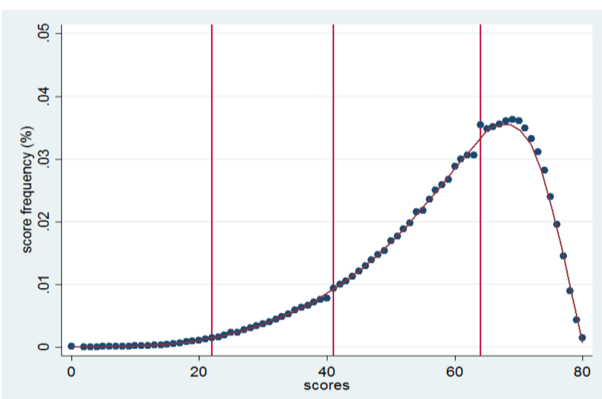
Figure C.2: Score manipulation over time: Science



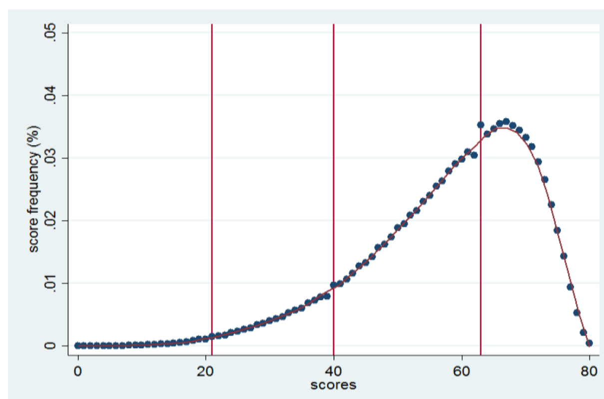
Panel A: 2006



Panel B: 2007



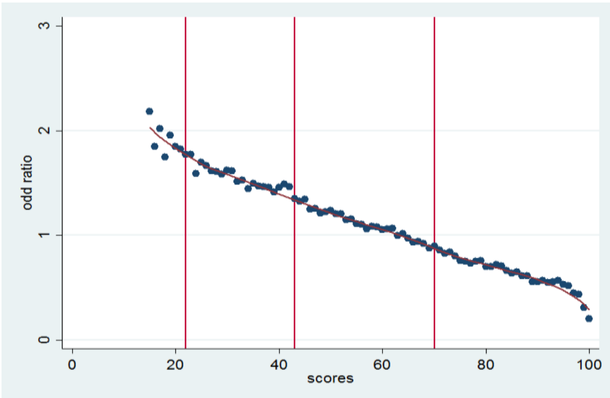
Panel C: 2008



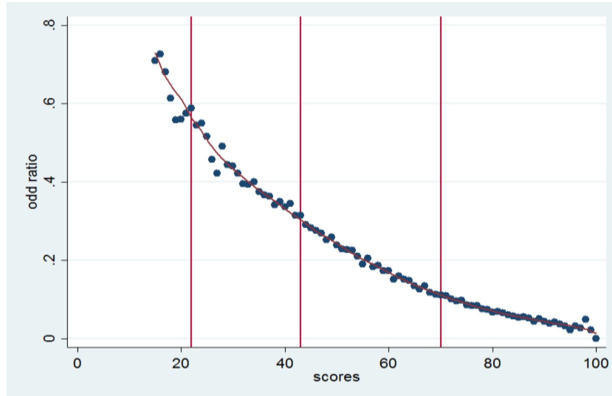
Panel D: 2009

Note. This figure shows Key Stage 2 score distributions for science in selected years before (2004 and 2007) and after (2008 and 2011) the removal of borderlining. In each panel, the vertical lines are critical thresholds set in that year. The continuous line is a local linear regression fit obtained excluding observations in the [-3,2] windows around thresholds. The science test was dismissed in 2010.

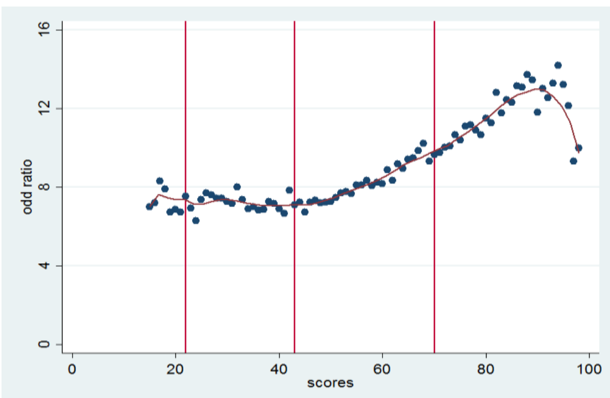
Figure C.3: Anatomy of manipulation



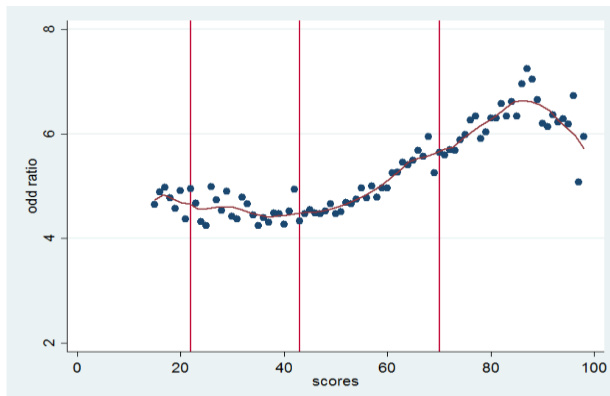
Panel A: Males



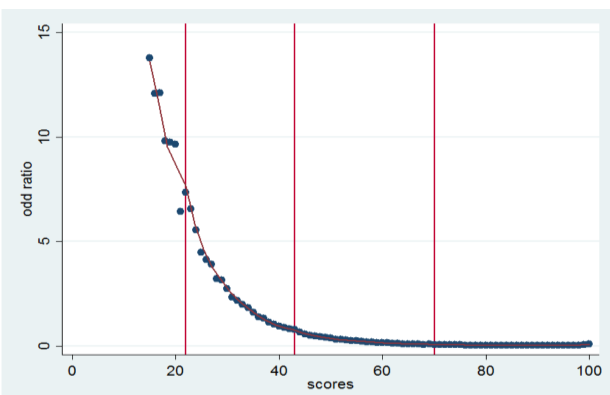
Panel B: Free School Meals



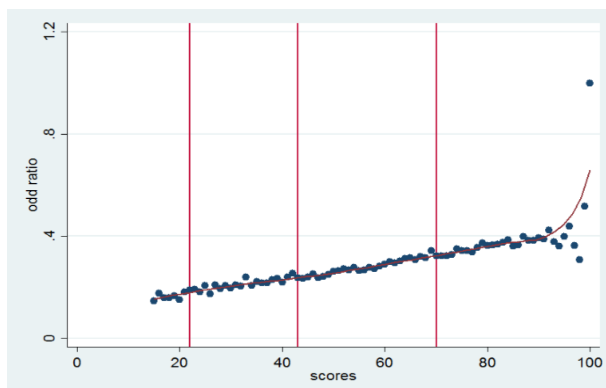
Panel C: English-speaking



Panel D: White



Panel E: Special Education Needs



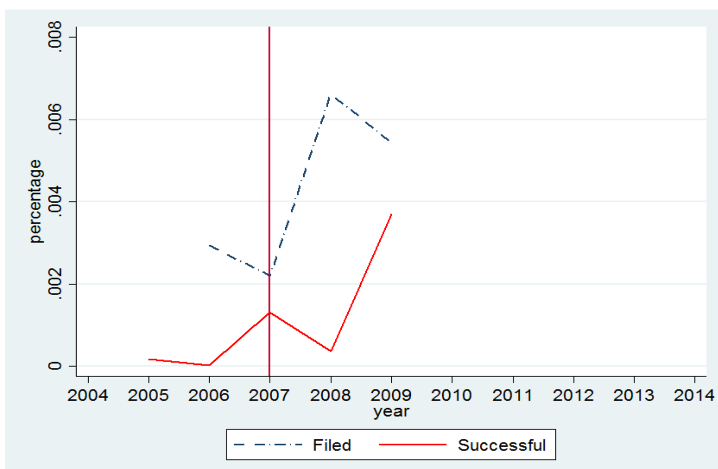
Panel F: Autonomous schools

Note. This figure shows odds ratios for selected school and student characteristics by values of the English score in 2007. Odds ratios are constructed as probability of an event (e.g., the student is male) divided by one minus this probability. The following variables are considered: gender of student (Panel A), student is on Free School Meals (Panel B), student is English-speaking (Panel C), ethnicity of student (Panel D), student with special education needs (Panel E) and school is autonomous (Panel F). In each panel, the vertical lines are critical thresholds set in 2007. The continuous line is a local linear regression fit obtained excluding observations in the [-3,2] windows around thresholds.

Figure C.4: Appeals filed by schools

Panel A: English

Panel B: Math



Panel C: Science

Note. The three panels show appeals filed by schools (dotted line) and successful appeals (continuous line) as fraction of number of scripts marked. Panel A refers to English scripts, Panel B refers to math scripts, Panel C refers to science scripts. The English test and science test were dismissed in 2010 and 2012 respectively. The vertical line denotes the last year before abolition of borderlining (2007).

Figure C.5: Example of open-ended question in reading test

13. Why do you think many people admire Evelyn Glennie?

13

3 marks

Figure C.6: Example of open-ended question in math test

15

Here is a number chart.
Every third number in the chart has a circle on it.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22			

The chart continues in the same way.
Here is another row in the chart.

Draw the missing circles.

71	72	73	74	75

15a
1 mark

Will the number **1003** have a circle on it?
Circle **Yes** or **No**.

Yes / No

Explain how you know.

15b
1 mark