How Can We Evaluate the Effectiveness of Grammar Schools in England? A Regression Discontinuity Approach

Binwei Lu
ORCID: 0000-0002-3396-7635

School of Education, Durham University, Durham, the United Kingdom
binwei.lu@durham.ac.uk

How Can We Evaluate the Effectiveness of Grammar Schools in England? A Regression Discontinuity Approach

Previous research evaluating grammar school effectiveness has generally relied on snapshot or longitudinal regression models to deal with pre-existing differences between grammar school pupils and those in non-selective schools. These passive designs are based on correlations, and cannot demonstrate clear positive causal relationships between grammar school attendance and subsequent attainment. After accounting for the variables available for the analysis, pupils in different schools might still have distinct and unmeasured characteristics which threaten the validity of any conclusions drawn. Given that a randomised trial is not feasible under current conditions, this study addresses the limitations of previous research, using a Regression Discontinuity Design (RDD) approach. This is the first use of RDD to attempt to make a robust causal inference about the effectiveness of grammar schools in one local authority in England. However, the authority, the Department for Education and the schools would not provide the individual data on pupils' 11+ entry test scores, and the scores obtained could not be uniquely matched to Key Stage 4 outcomes. While the model presented suggests that there is an advantage to grammar school attendance, the incomplete data means the study is more a feasibility trial of this strong design than any kind of definitive test intended to settle the debate on grammar school effectiveness. Conducting this design with national data on grammar school selection would create the most powerful evidence so far. To promote an effective and equitable education system for generations, those advocating the expansion of grammar schools should make the responsible decision to disclose all grammar school selection data for the purposes of research.

Keywords: regression discontinuity design; grammar schools; school effectiveness; educational equity

Introduction

Grammar schools are probably the most controversial schools in England as they uniquely retain attainment-based selection in the comprehensive system. It is believed by the government that grammar school pupils outperform their counterparts in non-selective schools even after accounting for prior attainment (DfE, 2016). However, despite a plethora of evidence from both the government and researchers, grammar schools' effectiveness remains unclear due to research relying on passive designs which are not conducive to causal inference. Most existing research has used statistical models to control for pre-existing differences between grammar school pupils and others. However, these estimations become biased whenever influential baseline variables between pupil groups are neglected, unavailable, or unmeasurable. When differences in later attainment emerge, it is unclear whether this is due to the school attended, or imperfections in the modelling process. To conduct a robust evaluation on the effectiveness of today's grammar schools in England, this study applies a Regression Discontinuity Design (RDD) to compare the academic performance of grammar school pupils with those in non-selective mainstream state schools. An RDD approach is a strong alternative to a Randomised Control Trial (RCT) in making causal statements, and it does not rely on baseline variables for an accurate estimate (Lee & Lemieux, 2009). Thus, the RDD approach avoids the limitations of the regression models used in previous research.

The remainder of the paper starts with the recent history grammar schools and previous studies. The methods section explains how secondary data were obtained, cleared and matched, and presents the process of identifying the treatment effect in one chosen local authority (LA). The findings present the estimation results in the RDD. Lastly, implications are given based on the estimation results, the feasibility of applying the RDD, and the tough process of accessing grammar school data.

Political attempts at grammar school expansion

Despite Labour's 1998 ban on establishing new grammar schools, the effort to revive grammar schools has continued. In 2015, the Weald of Kent was allowed to establish an annexe 10 miles away, which is regarded as the first 'new' grammar school in 50 years (Coughlan, 2016). Since 2016, subsequent policies such as encouraging existing grammar schools to expand and allowing the conversion of non-selective schools into selective ones, has been released by the government to emphasise grammar schools' role in raising national academic performance and helping the poor (DfE, 2016). However, the expansion of grammar schools has received resistance most noticeably from campaign groups (e.g. Comprehensive Future and the Kent Education Network) and academics due to concerns such as low academic benefits and impediments for disadvantaged pupils (Allen, Bartley & Nye, 2017). But the government's effort to expand grammar schools has persisted (Harding, 2017). In May 2018, the Department for Education announced a £50m expansion fund to create more grammar school places, and 4,000 new school places were planned (Sellgren, 2018). While the number of grammar schools supported by the fund at this stage is small, the message emphasises the importance of grammar schools and the continued possibility of returning to the selective system.

However, it remains unclear whether there is any benefit to the expansion, as well as the extent of such benefit if it exists. The government claims that expanding grammar schools has two major benefits—enhancing academic standards nationally and promoting social mobility. Both reasons are based on the perception that grammar schools are more effective than other state schools. Therefore, before addressing the ideological debate over who should be allocated more resources within the state system (whether to give more rewards for high-performers to maximise excellence or to offer extra help to less advantaged pupils to protect equity), the most fundamental question is whether grammar schools are more effective than other mainstream state schools.

The value-added approach and its limits

Measuring school effectiveness is not an easy task. Schools have different intakes, and pupil characterises influence later learning outcomes. Although grammar schools usually have better test scores than other state schools, this may be a result of their advantaged intakes (Gorard & See, 2013). The search for a fair method to calculate a school's net effect is ongoing. The most influential innovation was the Value-Added (VA) approach. Controlling for prior attainment, VA calculates pupils' relative progress within a fixed duration (Leckie & Goldstein, 2017). Following this principle came a more complex approach, the Contextual Value-Added (CVA). CVA further controls for pupil's contextual backgrounds, thus capturing factors such as socioeconomic status, gender and ethnicity, as these factors also influence later performance but are beyond schools' control (Perry, 2016). The CVA approach is regarded as isolating the net effect of schools more thoroughly than VA, which alleviates the bias of unfairly assessing schools that serve disadvantaged pupils.

Despite the robust logic of the VA/CVA approach, there are several challenges to its validity (Harker & Tymms, 2004). Firstly, the calculation is largely limited by the availability of background variables and the proportion of missing data. Even in high-quality databases such as the National Pupil Database (NPD), only around 70%-85% of pupils have complete data on attainment and contextual factors (Gorard & See, 2013). Secondly, the estimated VA/CVA scores are very unstable across time, and the correlation after 4 years may drop below 0.5 (Gorard, 2010; Leckie & Goldstein, 2017). Thirdly, even a very moderate rate of measurement error (e.g. 10%) in test scores would be accumulated to large errors, which could be 40 times larger than the estimation results (Gorard, 2010). Therefore, the bias of unmeasured pre-existing differences between pupils and measurement errors in baseline variables is considerable (Perry, 2016). Regardless of threats to the validity of VA/CVA, this approach has been widely used in the UK since 2002 (Leckie & Goldstein, 2017).

Previous research on grammar school effectiveness applying the VA/CVA approach

Following the principle of VA/CVA, researchers in England have evaluated the effectiveness of grammar schools. So far, the evidence is mixed, but some systematic patterns emerged. A common practice in each of the studies discussed below is to control for both pupil-level and school-level variables. The major differences have been found between those studies which account for school-level prior attainment, and those which do not. While the latter usually finds a positive grammar school effect, the former tends to report no difference between grammar schools and others (see Appendix 1 for details of these studies).

Studies without controlling for school-level prior attainment

For studies without school-level prior attainment, the general pattern of the estimated grammar school effect is positive. For example, Schagen and Schagen (2003) found larger progress in grammar schools from Key Stage 2 (KS2) to Key Stage 3 (KS3), and a small positive advantage from KS3 to Key Stage 4 (KS4) for average pupils. But for high-performing pupils at KS3 (level 7 or above), there is no difference in progress associated with school type. A later study by Atkinson, Gregg, and McConnell (2006) concluded that grammar school pupils achieved four grades higher on capped GCSE than equivalent pupils in non-selective areas. Similarly, Levačić and Marsh (2007) noticed a six-grade advantage on total GCSE/GNVQ associated with grammar school attendance, but again the effect drops among high-attaining groups. Focusing on Buckinghamshire only, a positive result was found by Harris and Rose (2013) that grammar school pupils are 10% more likely to achieve 5 A*-C on GCSE/GNVQ than equivalent pupils in non-selective schools. The differences in statistical models, outcome variables, and geographical areas between these studies complicate the comparison of results. However, in all of these studies, grammar school pupils made more progress than their counterparts in comprehensive schools and secondary modern schools.

One exception is Gorard and Siddiqui's (2018) study, which did not account for school-level prior attainment, but still found no positive grammar school effect. They noticed that adding 'whether a pupil went to a grammar school or not' did not improve the fitness of the model for predicting capped GCSE results. This implies that grammar schools are no better than other schools. Unlike previous research using binary indicators of Free School Meals (FSM) eligibility, they calculated the total years of being eligible for FSM. Since this new variable carries more information than traditional ones, it may be more powerful in removing unmeasured differences between grammar school pupils and others, thus reducing the estimated effect of grammar schools.

Studies controlling for school-level prior attainment

The most extensive attempt at evaluating grammar schools' effectiveness is that of Coe et al. (2008). Based on both a systematic review of previous research and their own analysis, they concluded that pupils in grammar school might have an advantage of 0 to 3/4 of a GCSE grade per subject. The wide variance in the estimated effect is primarily the result of the choice of baseline variables. If regression models only control for pupil-level variables, the estimates are substantially positive (reaching 0.75 grade per subject at most). However, once school-level variables, which include prior attainment, proportion of FSM, the Income Deprivation Affecting Children Index (IDACI), and single-sex status, are also controlled for, the difference between grammar schools and others drops to around 0. Coe and his colleagues (2008) believed that the lower end of the grammar school effect is more reliable than substantially positive results, since grammar school pupils have already progressed more than their peers, prior to attending secondary schools. The conclusion thus presents salient differences from most studies discussed in the previous section which also controlled for pupil-level and some school-level variables, but omitted school-level prior attainment.

The choice of controlling for school-level prior attainment or not

The substantially decreased grammar school effect when school-level prior attainment is added is not surprising. One of the most distinct features of grammar schools is their high school-level prior attainment by design. Therefore, it is possible that accounting for school-level prior attainment removes genuine difference between schools. The issue of over-correction appears whenever beneficial school actions correlate with advantaged intakes (Visscher, 2001). For example, schools with better high average prior attainment may have more aspirational norms and high academic expectations, and they are also more likely to attract high-qualified teachers and managers. Since these positive factors all contribute to academic success, accounting for school-level prior attainment is likely to underestimates the real effect of grammar schools. However, there are also supporting claims that controlling for school-level prior attainment creates more accurate results. It is possible that adding this variable removes more unmeasured differences between pupil groups, differences which are not sufficiently accounted for by pupil-level surface variables or other school-level compositional variables (Coe et al., 2008). More importantly, adding school-level prior attainment is believed to correct measurement errors in baseline variables which otherwise upwardly bias the effectiveness of more advantaged schools (Perry, 2019). Based on this finding, Perry (2019) concluded that models without accounting for school-level prior attainment present a 'phantom' grammar school effect. Overall, these contradictory consequences of controlling for school-level prior attainment can be correct to some extent simultaneously. Since we don't know the exact reason why the estimated school effectiveness is different before and after school-level prior attainment is added, it is hard to tell which model creates more accurate results.

Due to the correlational nature of the VA/CVA approach, the choice of baseline variables is always difficult and there is no perfect solution (Visscher, 2001). However, although previous studies do not reach a consensus on the effectiveness of grammar schools, their results provide the lower and upper bounds of the potential real effect. The mixed evidence suggests that grammar schools may perform better than other state schools, yet we are unsure to what extent.

The RDD approach and its application in the effectiveness of grammar schools

The most robust approach to make a causal inference is to conduct an RCT (Shadish, Cook, & Campbell, 2002). However, in school effectiveness research, pupils cannot be allocated randomly to different pathways. As a strong alternative to an RCT, the function of an RDD is similar. In an RDD, participants are allocated to either the treatment or the control group according to the cut-off point of a continuous assignment variable. Only those who reach the cut-off point are given the treatment. If participants' assignment variables could not be manipulated precisely, their chances of just making it or just missing it can be regarded as locally random (Lee & Lemieux, 2009). This process solves the problem of pre-existing differences between the treatment and control group.

While there have been fruitful applications of RDDs for school effectiveness (e.g. Gibbons, Machin, & Silva, 2013; Luyten, Tymms, & Jones, 2009), only one study has used this design to evaluate grammar schools in England. Clark (2010) focused on grammar schools in East Riding and detected a small grammar school effect in Year 9 test scores, which is 7% higher than pupils just below the cut-off point. Meanwhile, there is also a positive impact on Higher Education participation, the rate of which is 6% (Clark, 2010). However, this study applied data in the late 1960s, when the transformation of comprehensive schooling was prevalent. Data collected decades ago raises doubts on the external validity of this research in the present (Coe et al., 2008).

Generalisability of the RDD

The generalisability of the effect of grammar schools in the RDD approach is important both in terms of methods and policy. While the RDD approach is considered as a strong alternative to an RCT, the localised nature of this design makes its generalisability from the cut-off point to the whole data range a concern (Bloom & Porter, 2012). This is especially relevant to grammar schools as the treatment effect may be heterogeneous at different data points. Previous studies have demonstrated that grammar school attendance is more beneficial for pupils with lower attainment, presenting differential school effectiveness (Atkinson, Gregg, & McConnell, 2006; Levačić & Marsh, 2007). This means the treatment effect of grammar schools may be inconsistent across performance levels. Thus, we might expect the result at the cut-off point to be larger than at higher points. However, even if the estimate is only relevant to borderline pupils, the result is still meaningful because the expansion of grammar schools would also have the most influence on borderline pupils.

While attending grammar schools might be beneficial (e.g. highly-qualified teachers and better resources), there are also concerns such the negative self-perception resulting from the big-fish-little-pond effect (Marsh, 1987). Due to the mixed evidence in previous studies of grammar schools' effectiveness, an estimation based on borderline pupils is thus important for its own sake.

Despite the potential limits of the RDD approach, researchers have also found evidence contrary to the pessimistic perception of the low generalisability of RDD (Bloom & Porter, 2012). According to Lee and Lemieux (2009), the discontinuity at the cut-off is a weighted average effect across all observations, and the weight calculates an individual's probability of being located near the cut-off point. Therefore, the estimate is relevant to all the observations, and the strength of relevance is influenced by the rate of noise in the assignment variable. Larger errors in the assignment variable create a more heterogeneous pupil group near the cut-off point, increasing generalisability (Jacob et al., 2012). There is no public information on the quality of the 11+ test, but previous research in Northern Ireland has pointed out the low reliability of the 11+ (Gardner & Cowan, 2000). The larger error contained in the test score means that pupils with different aptitudes may score equally-well on the 11+, and thus have a similar probability of being located near the cut-off point. In this case, the results at the cut-off point would be closer to the overall average treatment effect and more relevant to pupils at higher points.

The external validity of the study also warrants discussion. The proportion of grammar school places varies across LAs, ranging from about 1% to over 30%. The unbalanced chances to attend grammar school leads to variation in selection difficulty across local areas (Coe et al., 2008). Additionally, grammar schools in different LAs also vary in ethnic mix and the underrepresentation rate of disadvantaged pupils. Therefore, the results of the chosen LA in this study might have limited similarity to other LAs, and the findings should not be regarded as the effect of grammar schools in the national scope. Beyond geographic differences, caution is also needed when generalising the findings to grammar schools in other historical periods, because the nature of grammar schools has changed over the years.

In summary, while the generalisability of the RDD has some constraints, it does not effect its strong ability in making causal inferences. Based on the limitation of regression models controlling for baseline variables in evaluating grammar school effectiveness, the impossibility of applying an RCT, and the lack of evidence from research applying designs which are strong enough to make a causal inference, this new study solves these limitations by using an RDD.

Methods

The process of selecting pupils into grammar schools

England has no national system of selecting pupils into grammar schools, and each area has its own selection process. The analysis in this study only focuses on one LA in England, where the overall proportion of grammar school places is high. In this LA, eligibility for grammar school attendance is primarily decided by the 11+. In 2011, the 11+ in this LA included three subjects. In order to be eligible to attend grammar school, pupils not only had to cross the threshold in total score (360/420), but also the minimum requirement for each individual subject. Apart from the formal test, head teachers of primary schools can appeal if they are not satisfied with their pupils' test results. In this case, extra supporting materials are evaluated.

The theoretical framework of RDD design

The basis of a valid RDD is that the allocation of participants into the treatment or the control group is decided by the cut-off point of an assignment variable. While participants who made the threshold are given the treatment, those who just missed the cut-off are not. As the values of the assignment variable are similar among participants in the neighbourhood of the cut-off point, a comparison of the outcome variable can attribute any discontinuity at the cut-off point to the treatment (Lee & Lemieux, 2009).

In the ideal 'sharp' RDD, all individuals who have passed the cut-off point would get the treatment and those who have missed it would not. However, in reality, it is more common to encounter programmes with imperfect compliance and programmes in which the eligibility to get the treatment is not decided by one assignment factor alone. This means an individual who reaches the threshold may not get the treatment ('no shows'), while one who does not reach the threshold may in fact get it ('crossovers'). For example, there might be some pupils who did not achieve a passing score on the 11+, but still attended grammar schools. Inversely, it is also reasonable that not all pupils who passed the selection attended grammar schools. These situations are categorised as 'fuzzy' RDDs, which are similar to an RCT with imperfect compliance (Lee & Lemieux, 2009, 23).

The selection process for grammar schools in our sample LA is a typical 'fuzzy' RDD, in which the total score on the 11+ is not the only factor deciding grammar schools' eligibility. Pupils' test scores on the three individual subjects of the 11+, and the Head Teacher Panel, also influence eligibility. Therefore, a 'fuzzy' RDD approach is applied to estimate the treatment effect.

Empirical strategy

According to the definition of the treatment effect in a 'fuzzy' RDD (Jacob et al., 2012; Lee & Lemieux, 2009), the estimation can be written as:

$$Y_i = \alpha + \beta T_i + f(X_i) + u_i, \qquad (1)$$

$$T_i = \gamma + \delta D_i + g(X_i) + v_i, \qquad (2)$$

where Y_i is the outcome measure for each individual i; T_i is the treatment dummy; X_i is the assignment variable (X_i=0 is the cut-off point); D_i is the binary indicator of whether individual i reached the cut-off point (D_i=1 if $X_i \geq 0$); u_i and v_i are the random error for each individual. The effect of attending grammar schools which needs to be estimated equals β. To make it easier to understand, these two equations can be simplified as:

KS4 performance = grammar school effect * grammar school or not +
the effect of prior attainment,
Grammar school or not = compliance rate * passed threshold or not +
the effect of prior attainment

The treatment effect in the 'fuzzy' RDD revealed in equation (1) and (2) is consistent with a standard instrumental variable setting, and thus it can be estimated using a Two-Stage Least Squares (2SLS) model (Hahn, Todd, & van der Klaauw, 2001). The parametric approach involves finding appropriate regression lines to fit data points. A correct estimation thus requires accurately modelling the relationship between KS4 performance and prior attainment ($f(X_i)$), and the relationship between the 'grammar school or not' and prior attainment ($g(X_i)$). For example, if the relationship between prior and later attainment can be graphically presented as a straight line, then a linear functional form can be used as $f(X_i)$. However, although it is a widespread practice to use a linear function to depict the relationship between prior and later attainment, the actual relationship between these two variables may be a curve line, as it could be harder to make equivalent progress at a high level than at lower ones. Therefore, quadratic function forms are also fitted in this study to avoid misspecification. Meanwhile, the slopes of the regression lines are also allowed to vary on two sides of the cut-off point. To comply with the calculation rule of the 2SLS analysis, the same type of regression line is used for both equations (Lee & Lemieux, 2009). More details of identifying the treatment effect are attached in Appendix 2.

Apart from the parametric approach, which finds regression lines to fit data, the estimation of the treatment effect can also be realised through the non-parametric approach, which selects data to fit regression lines. The non-parametric approach applies local linear regression to depict the relationship between explanatory and outcome variables. While the overall pattern between these variables may not be linear, if we only select data points within a small range, it is likely to see a linear relationship (Hahn, Todd, & van der Klaauw, 2001). Therefore, unlike the parametric approach which makes an estimation based on all the data, the non-parametric approach only uses data within a limited range. Since the estimation of interest in this study is at the cut-off point, data should also be selected on both sides of the cut-off point. The range of selected data on each side of the cut-off point is also referred to as a ‘bandwidth’, and an accurate estimation heavily depends on choosing a right bandwidth. Instead of using visual inspection, the optimal data bandwidth is calculated according to the data-driven algorithm proposed by Imbens and Kalyanaraman (2012). To avoid redundancy, the same bandwidth is used on both sides of the cut-off point, and in equations (1) and (2) (Imbens & Lemieux, 2008). Despite the different calculation processes, the non-parametric estimate should be similar to the estimate in the parametric approach. It is thus used as the complimentary approach to the parametric estimation.

In this study, the treatment effect of interest is the effectiveness of grammar schools compared with non-selective mainstream state secondary schools. The outcome variable used as the indicator of school effectiveness is the capped GCSE point score at KS4. According to the Secondary Accountability Measures (DfE, 2018), the highest point score for a GCSE subject in 2017 was 8.5 (A*). The interval between each grade is 1.5 point score for A*-C grades and 1 point score for C-E grades. This means the highest possible capped GCSE point score for each individual is 68. However, there were 53 pupils in the sample who achieved GCSE results higher than this. A similar situation has been encountered by Coe et al. (2008), who noted that comparisons between schools would not be affected since the point score scale is consistent for all (p. 200). As the total test score on the 11+ is the major factor deciding pupils’ eligibility to attend grammar schools, it is centered at the lowest passing score and set as the assignment variable (point 0 is the cut-off point). The value of the assignment variable ranges from -140 to 60, but there are only about 10% of pupils scored lower than -60. An important premise of a valid RD design is participants’ inability to precisely control the assignment variable (Lee & Lemieux, 2009). This condition can be easily met on the 11+ as the passing score of the 11+ may change each year.

Based on the principle of the RDD, baseline covariates are believed to be randomly distributed in the treatment and control group near the cut-off point. Thus, there is no need to control for these variables. However, the regression estimates between models with and without baseline variables are still compared to evaluate the internal validity of the design, as theoretically both types of models should yield similar results. A robustness check is also conducted by trimming the 10% outermost observations at both ends of the assignment variable. For privacy reasons, data points representing fewer than 5 cases are not presented in all the figures.

The data set

Absent from the NPD and all the major databases in England, the result of the 11+ is not publicly available. The 11+ data used in this study is provided by a local group, which includes 7,917 local pupils who sat the 11+ in this LA in 2011 (2011/2012 KS2 cohort). This file contains the 11+ test data, including test score for each subject, whether a pupil has been entered in the Head Teacher Panel, and the result of the selection. It also keeps a record of pupils' backgrounds, including FSM status, ethnicity, IDACI, and KS2 average point score. While the 11+ file tells whether a pupil passed the selection, it provides no information on actual attendance. A comparison between the NPD and the 11+ file demonstrates that the total number of local pupils in grammar schools in the NPD is close to the number of local pupils in this LA who passed the selection as recorded in the 11+ file, with an attrition rate below 3%. This is a small number compared to the overall effect size as shown below. Therefore, the selection result in the 11+ file is used as the indicator of actual participation in grammar schools.

Lacking any record of academic performance at later stages, the 11+ file is linked to the NPD data of the same cohort for the 2016/2017 GCSE results. However, since the 11+ data is anonymous (without any form of identifier), the 11+ file and the NPD data extract are matched through family backgrounds and KS2 attainment. While FSM status, ethnicity, KS2 point score and school types can be exactly matched between the two files, IDACI scores are slightly different in the 11+ file and the NPD, which is thus matched with a 0.01 tolerance rate. Another problem which occurred in the matching process is duplicate cases of pupils who share the same combinations of all the available demographic and attainment variables. In order to make one-to-one unique matches between the two files, these duplicate cases are deleted. This process excludes 52% (4,119) of the total samples in the 11+ file. While this process might threaten the representativeness of the sample, it is the best available option due to the limited information in the 11+ file (alternative sampling strategies are discussed in Appendix 3). After data clearing, 2,628 valid cases in the 11+ file are matched to their NPD records, and 2,541 cases in the mainstream state schools are kept for the RDD analysis.

It should be noted that typical RDDs do not involve matching. The complicated process of matching pupils' prior attainment with later performance in this study is a result of the limited 11+ data in England. Cases omitted during the matching process imply that the estimation is not definitive, and the results are more about the feasibility of the RDD approach in causally evaluating the effectiveness of grammar schools.

Descriptive results

Limited by the 11+ data, only one LA is included for analysis. This LA has a large proportion of grammar school places, higher than the average rate of selective LAs. The KS2 and KS4 results for grammar school pupils in this LA are lower than those of grammar schools in other LAs, but only slightly (about 3% lower). Despite this small difference, the overall proportion of FSM pupils in grammar schools in this LA is identical to the national mean of grammar schools, and the IDACI score is also similar to other selective LAs.

As mentioned in the methods section, only about half of the valid cases in the 11+ files were included for analysis. Therefore, the characteristics of the selected sample are first contrasted with the original cohort in the 11+ file. As presented in Table 1, the characteristics of the sample group are similar to the population data in the 11+ file. This is apparent in terms of academic performance, as the KS2 performance and the 11+ test scores are nearly identical between the two groups. However, the population in the 11+ file has a more advantaged average IDACI score and a lower proportion of FSM pupils. In order to minimise the influence of the difference in IDACI and FSM between the sample and the population, analysis has also been done to randomly delete cases from the sample to keep the average IDACI score and the proportion of FSM pupils consistent with the 11+ file. However, the two sample sets yield similar results and lead to the same conclusion and the sample group is not trimmed further.

Among the sample group, 40% (1,043) of the local pupils who sat the 11+ in 2011 were assessed as suitable to attend grammar schools. The difference between those who passed the selection and those who did not is clear, with the former having more advantageous results, both in terms of their academic performance at two key stages, and their demographic characteristics (Table 1). Despite the pre-existing differences between grammar school pupils and their counterparts, crossing the cut-off point does not cause simultaneous discontinuities in baseline variables. Figure 1 is an example of the similar demographic features of pupils just above and below the cut-off point. When the average IDACI score at each point of the assignment variable is plotted, neither the binned average value, nor the fitted regression lines, presents discontinuity at the cut-off point. This means there is no systematic difference in IDACI scores between pupils who just reached the threshold and those who just missed it. Other baseline variables, FSM and KS2 performance, are also similar to Figure 1. This proves the irrelevance of background variables in estimating the treatment effect at the cut-off point in this study.

In order to test the internal validity of the RDD further, the frequency of the assignment variable is also checked. If pupils can have accurate manipulation over the assignment variable, we could anticipate that the frequency just above the cut-off point would be higher than below the cut-off point. The frequency at each score is plotted in Figure 2. The graph shows a ceiling effect at the right end. Since 60 is the highest possible value in the assignment variable, it also contains pupils who might have achieved higher scores otherwise. Despite this outlier, the distribution of the assignment variable is smooth. There is no evidence that pupils can have full control over their test scores, with the frequencies just above and below the cut-off point being similar.

Although the assignment variable is not the only deciding factor of pupils’ eligibility to attend grammar school, the jump in the probability of treatment is still strong at the cut-off point. As presented in Figure 3, the probability of passing the selection is near zero before point -10. The rate grows from point -10 and increases to about 0.4 at the cut-off point. The probability reaches 1 at point 13 and stabilises after point 30. Overall, the probability of going to grammar schools increased from near 0 to 1 within the small interval of -10 to 13.

The relationship between the outcome variable and the assignment variable is also depicted. A visual inspection of the average GCSE results at each score of the assignment variable shows a positive correlation (Figure 4). The distribution on the left lower side is irregular, which is the result of the dearth of observations at these scores. Due to the concentrated points, a visual inspection of the raw data reveals little discontinuity in the outcome variable at the cut-off point. However, when observations are grouped into bins and the number of observations within each bin is represented by the size of the dot, a discontinuity at the cut-off point is revealed (Figure 5). According to Figure 5, pupils in grammar schools with low 11+ scores actually have similar GCSE results at KS4, despite differences in prior attainment. While this study is unable to reveal the exact reason, it is possible that some borderline grammar school pupils who performed less well on the 11+ day actually do better later than those with higher 11+ scores. The pattern might be related to the differential effectiveness of grammar schools as revealed in previous studies, which noted that grammar schools benefit borderline pupils the most (e.g. Atkinson, Gregg, & McConnell, 2006; Levačić & Marsh, 2007). Based on these two figures, the following sections test whether the graphic discontinuity can be regarded as the treatment effect of grammar schools.

Estimation results

The parametric approach of RDD

The result of the parametric approach is presented in Table 2. The first model calculates the treatment effect by fitting linear functions on both sides of the cut-off point, and in both stages of the regression in the 2SLS. The functional form used in the second model also includes interaction terms to allow slopes to vary on both side of the cut-off point (as described in the methods section). In the third and fourth models, quadratic and quadratic interaction functions are fitted.

As shown in Panel A of Table 2, the estimates of the treatment effect vary when different functional forms are used. Based on the calculation of the first three models, the treatment effect is small. It is in fact negative in Model 2 and Model 3. Meanwhile, Model 4 not only has a much larger effect size, it also reveals that attending grammar schools is beneficial, which is about four GCSE point scores. This is equivalent to 0.57 C-E grade or 0.38 A*-C grade per GCSE subject.

The real effect of grammar schools depends on which model presents a more convincing result. As shown in Panel A of Table 2 above, Model 4 has the highest R-square value, but the difference is subtle. However, when using the specification test suggested by Lee and Lemieux (2010), Model 4 is the only one that passes. This suggests that there is unexplained variability missing from Model 1-3. Meanwhile, Model 4 also yields the best result evaluated by the Akaike information criterion (AIC). While it is possible that when a functional form gives more parameters, the fitness of the model will inevitably increase, a graphic presentation also suggests the fitness of a quadratic interaction function (Figure 6). The estimate is supported by the robustness check and the non-parametric approach in later sections as well.

In Model 5-8 (Panel B of Table 2), pupils' FSM status, IDACI score and whether they speak English as an additional language (EAL) are included as baseline covariates. The distinct estimates between models are still clear when different functional forms are used, and including these baseline variables increases the effect size in each model without changing the direction. Overall, the results in Model 5-8 correspond to each of the models in Panel A. This means that whether or not the baseline variables in the RDD are included, the results are similar when the same functional form is applied.

Robustness check of the parametric approach of RDD

In order to assess whether the estimated treatment effect is sensitive to the changes in the data (especially cases with extremely high and low values in the assignment variable), a robustness check trimmed the 10% outmost data points at the two farthest ends of the assignment variable. The process thus excludes data points above 57 or below -77 in the assignment variable.

Overall, the treatment estimates experienced some changes when observations with the highest and lowest values in the assignment variable were excluded (Table 3). The estimated treatment effects increased to about 2 point scores in Model 1-3. This was much larger than in the original models in Table 2, and the direction of the coefficient was also altered in Models 2 and 3. The results in Model 4 also grew, but only to a mild extent. After trimming 10% of the cases, the treatment effect in Model 4 equals 5 point scores, which is close to the result (4.57) in the original model (Table 2). In Panel B, the robustness check was conducted on models with demographic variables. The estimated treatment effects remained close to Panel A, revealing again the irrelevance of including baseline variables in a valid RDD. However, the treatment effects in Model 5-7 are inconsistent with the original results in Table 2, and Model 8 is the only one remaining close to its untrimmed result. Therefore, the unstable results in Model 1-3/ Model 5-7 and the similar results in Model 4/ Model 8 before and after the data trimming reveals the better fit of a quadratic interaction function in depicting the sample data again. Among all the functional forms, it is the least sensitive to the changes in the data.

The non-parametric approach of RDD

To confirm whether the treatment effect in the parametric approach is convincing, a non-parametric approach using local data points within a bandwidth on both sides of the cut-off point was also applied. As mentioned previously, the optimal bandwidth was decided based on the calculation proposed by Imbens and Kalyanaraman (2012). According to this principle, the optimal bandwidth in this study is 41.3. Based on the non-parametric estimation, the treatment effect is 4.32, which is close to the parametric results (4.57) in Model 4 (Panel A of Table 2). When baseline covariates are included in the non-parametric approach, the estimated treatment effect grows slightly to 4.62. This is also similar to the corresponding parametric results (5.21) in Model 8 (Panel B of Table 2). Therefore, the non-parametric approach yields results similar to those of the parametric approach using quadratic interaction functional forms.

A sensitivity test of the non-parametric estimation was also conducted to assess how stable the results are when different bandwidths are used. As revealed in Figure 7, the non-parametric estimation is negative when the bandwidth is below 20, which is the interval where the probability of getting the treatment spikes. The estimated treatment effect grows as the bandwidth widens, and stabilises around 4 point scores within a bandwidth of 30 to 50. The result decreases slightly after bandwidth 50, but remains positive until 60. Results of bandwidth larger than 60 are not presented, as it is already the largest possible size (the same bandwidth is selected on both sides of the cut-off point, and 60 is the maximum assignment value on the right side). Overall, the treatment effect based on the given optimal bandwidth is stable within a large interval.

Generalisability of the estimated grammar school effect

As discussed in previous sections, the degree of generalisability of the result at the cut-off point is largely related to the quality of the assignment variable. While it is not possible to know how close the estimation is from the overall average treatment effect, it is practical to examine the characteristics of pupils in the neighbourhood of the cut-off point. If the composition near the cut-off point is heterogeneous, the subgroup will be more similar to the population in the full data range, and the result at the cut-off can be applied to a wider scope. In order to see whether this subgroup is similar to the whole population, the KS2 test results of two groups are presented. For most of the pupils who scored right at the cut-off point in this study, their average KS2 points were 30 (42%) and 33 (54%). Meanwhile, the proportion of pupils in this LA with these two KS2 points is 47%. For pupils whose scores in the assignment variable are within a ±10 interval of the cut-off point, the average KS2 points for the majority are 27 (10%), 30 (48%) and 33 (40%). This overlaps with the KS2 performance level of 77% of the pupils in this LA. While this may have been due to the low discriminative ability of KS2 points, the results are consistent when KS2 average marks are used for comparison. Therefore, the estimated treatment effect is at least relevant to pupils with the above mentioned KS2 academic levels. Meanwhile, if a differential grammar school effect exists, the grammar school effect could be less pronounced for pupils with high prior attainment (Atkinson, Gregg, & McConnell, 2006; Levačić & Marsh, 2007). In this case, the treatment effect estimated at the cut-off point in this study would be larger than the effect for pupils with higher KS2 attainment, especially for those who scored above 33.

Based on the evaluation of pupils' attainment and background characteristics, the intake of grammar schools in this LA is similar to other selective LAs. However, due to the imbalance in grammar school places and different selection processes in each area, the nature of grammar schools in this LA may still differ from others. Since the content and threshold of the selection test varies across LAs, the discontinuity gap in this LA would not be informative about treatment elsewhere if the selection aims to pick the highest-performing 5% or 10% of pupils in a given year group. Additionally, the broader social context of each area may also influence the effectiveness of grammar schools and non-selective alternatives for pupils not in grammar schools. This means that the evidence in this study may diverge from the national pattern.

Conclusion

According to the RDD approach, the estimated treatment effect of grammar schools in this LA is approximately 4.5 GCSE point scores, which is equivalent to half grade per subject (about 10% of the average attainment of pupils below the cut-off point). The results of this study partly overlap with the only available RDD research on the grammar school effect in England, which revealed that the treatment effect on the Year 9 test score is 7% of the average performance of pupils just below the cut-off point (Clark, 2010). The results of this study are also within the range of the national pattern of possible grammar school advantage presented by Coe et al. (2008).

While the findings partly correspond to previous studies, the estimated effectiveness of grammar schools is not a definitive answer. Firstly, the findings are substantially limited by the 11+ data, which lacks both information on pupils' later academic performance, and identifiers consistent with any major databases in the UK. Although every precaution was made to carefully deal with the 11+ data, many cases in the original 11+ file were omitted. This process may have negatively impacted the quality of the evidence which is difficult to be compensated for by research design. Therefore, the value of this study lies more in the feasibility of applying the RDD approach to generate robust causal evidence of grammar schools' effectiveness, than in the actual estimation results.

Secondly, this study only covers a single LA. Although the characteristics of grammar school pupils in this LA do not present obvious differences to other selective LAs, the unbalanced chances of attending grammar schools, the varied selection difficulty in each area and the broader social context may create dissimilarities in school effectiveness. Therefore, the effectiveness of grammar schools revealed in this study is most relevant to the pattern in this LA, and may diverge from the national pattern even when the same design is applied. According to traditional school effectiveness models such as OLS, the effectiveness of grammar schools in this LA is higher than the national average. Thus, it is reasonable to anticipate that the RDD estimate of the grammar school effect on the national level would also be less obvious than the effect in this LA. Additional RDD studies in other LAs are needed to present a complete picture.

Lastly, as the treatment effect in this study is yielded through the comparison between grammar schools and non-selective mainstream state schools, the estimated grammar school effect is not an absolute academic level; it is the benefit in relation to non-selective schools in the same LA. While the results may indicate the effectiveness of grammar schools in raising academic achievement, they may also be a signal of penalties for other schools in an LA with a high proportion of grammar school places. As this study doesn't reflect impacts on pupils outside grammar schools, future research on surrounding schools is needed.

The threats imposed by the availability of the 11+ data are not unique to this study. They are also relevant to all future research attempting to present an accurate evaluation of England's selective system. The results of this study reveal that a strong research design which bypasses previous limitations in grammar school evaluation is workable. However, a definitive answer to grammar school's effect cannot be reached without transparent disclosure of the national 11+ data linked to later achievement. Government policy requires the support of hard evidence, but the absence of data prevents accurate evaluation, even with the help of effective research methods. Based on the high costs of new grammar school places, the small number of the potential participants in grammar schools, and the concurrent need to invest in basic educational areas in England, grammar school selection data should be disclosed for the purposes of research.

References

Allen, R., Bartley, J., & Nye, P. (2017). *The 11-plus is a Loaded Dice: Analysis of Kent 11-plus Data* (London, Education Datalab).

Atkinson, A., Gregg, P. & McConnell, B. (2006) *The Result of 11 Plus Selection: An Investigation into Opportunities and Outcomes for Pupils in Selective LEAs* (Bristol, The Centre for Market and Public Organisation).

Bloom, H. S., & Porter, K. E. (2012). *Assessing the Generalizability of Estimates of Causal Effects from Regression Discontinuity Designs.* Society for Research on Educational Effectiveness. Available online at: https://files.eric.ed.gov/fulltext/ED530557.pdf (accessed: 6 July 2018).

Clark, D. (2010) Selective Schools and Academic Achievement, *The B.E. Journal of Economic Analysis & Policy,* 10(1), 1-40.

Coe, R., Jones, K., Searle, J., Kokotsaki, D., Kosnin, A. M. & Skinner, P. (2008) *Evidence on the effects of selective educational systems: A report for the Sutton Trust*. Available online at: https://www.suttontrust.com/wp-content/uploads/2008/10/SuttonTrustFullReportFinal-1.pdf. (accessed: 6 June 2018).

Coughlan, S. (2016) *The persistent appeal of grammar schools*. Available online at: http://www.bbc.co.uk/news/education-30483031 (accessed: 6 June 2018).

Department for Education (DfE). (2016) *Schools that work for everyone: Government consultation*. Available online at: https://consult.education.gov.uk/school-frameworks/schools-that-work-for-everyone/supporting_documents/SCHOOLS%20THAT%20WORK%20FOR%20EVERYONE%20%20FINAL.PDF. (accessed: September 12 2017).

Department for Education (DfE). (2018) Secondary accountability measures: Guide for maintained secondary schools, academies and free schools. Available online at:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/772801/Secondary_accountability_measures_guidance.pdf (accessed: 18 March 2018).

Gardner, J. & Cowan, P. (2000) *Testing the test: A study of the reliability and validity of the Northern Ireland transfer procedure test in enabling the selection of pupils for grammar school places*. Available online at: https://eric.ed.gov/?id=ED467432. (accessed: 6 June 2018).

Gibbons, S., Machin, S. & Silva, O. (2013) Valuing school quality using boundary discontinuities, *Journal of Urban Economics,* 75, 15-28.

Gorard, S. (2010). Serious doubts about school effectiveness. *British Educational Research Journal, 36*(5), 745-766. doi:10.1080/01411920903144251

Gorard, S. & See, B. H. (2013) *Overcoming Disadvantage in Education* (Abingdon, Routledge).

Gorard, S. & Siddiqui, N. (2018) Grammar schools in England: a new analysis of social segregation and academic outcomes, *British Journal of Sociology of Education,* 39(7), 909-924.

Hahn, J., Todd, P. & Van der Klaauw, W. (2001) Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design, *Econometrica*, 69(1), 201-209.

Harding, E. (2017) *We've not given up on grammars, says May: PM backs expansion of selective schools for first time since election*. Available online at: http://www.dailymail.co.uk/news/article-4990862/Theresa-backs-grammar-school-expansion.html. (accessed: 6 June 2018).

Harker, R. & Tymms, P. (2004) The Effects of Student Composition on School Outcomes, *School Effectiveness and School Improvement,* 15(2), 177-199.

Harris, R. & Rose, S. (2013) Who benefits from grammar schools? A case study of Buckinghamshire, England, *Oxford Review of Education,* 39(2), 151-171.

Imbens, G. & Kalyanaraman, K. (2012) Optimal Bandwidth Choice for the Regression Discontinuity Estimator, *Review of Economic Studies*, 79(3), 933-959.

Imbens, G. W. & Lemieux, T. (2008) Regression Discontinuity Designs: A Guide to Practice, *Journal of Econometrics,* 142(2), 615-635.

Jacob, R., Zhu, P., Somers, M. A. E. & Bloom, H. (2012) *A practical guide to regression discontinuity*. Available online at: https://www.mdrc.org/sites/default/files/RDD%20Guide_Full%20rev%202016_0.pdf. (accessed: 6 June 2018).

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'progress 8'. *British Educational Research Journal*, 43(2), 193-212. doi:10.1002/berj.3264

Lee, D. & Lemieux, T. (2009) *Regression discontinuity designs in economics.* Available online at: https://www.nber.org/papers/w14723.pdf. (accessed: 8 June 2018).

Levačić, R. & Marsh, A. J. (2007) Secondary modern schools: are their pupils disadvantaged?, *British Educational Research Journal,* 33(2), 155-178.

Luyten, H., Tymms, P. & Jones, P. (2009) Assessing school effects without controlling for prior achievement, *School Effectiveness and School Improvement,* 20(2), 145-165.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. Journal of educational psychology, 79(3), 280.

Perry, T. (2016). English Value-Added Measures: Examining the Limitations of School Performance Measurement. *British Educational Research Journal, 42*(6), 1056-1080. doi:10.1002/berj.3247

Perry, T. (2019) 'Phantom' compositional effects in English school value-added measures: the consequences of random baseline measurement error, *Research Papers in Education,* 34(2), 239-262.

Schagen, I. A. N. & Schagen, S. (2003) Analysis of National Value-added Datasets to Assess the Impact of Selection on Pupil Performance, *British Educational Research Journal,* 29(4), 561-582.

Sellgren, K. (2018) *The 16 grammars that have won funds to expand*. Available online at: https://www.bbc.com/news/education-46429040 (accessed: 8 Jan 2019).

Shadish, W., Cook, T. & Campbell, D. (2001) *Experimental and quasi-experimental designs for generliazed causal inference.* New York: Houghton Mifflin Company.

Visscher, A. J. (2001). Public school performance indicators: Problems and recommendations. *Studies in Educational Evaluation*, *27*(3), 199-214.

Tables:

	KS2 average result		Total score on the 11+		IDACI		FSM (%)	
	Grammar schools	Others	Grammar schools	Others	Grammar schools	Others	Grammar schools	Others
Selected samples for analysis	32.5	29	392	326	0.14	0.18	5	19
All the cases in the 11+ file	32.7	29.2	392	327	0.13	0.17	3	12

Table 1: Characteristics of pupils in grammar schools and non-selective schools

	R-square of the model	Treatment effect
Panel A: Models without baseline variables		
Model 1 (Linear)	0.363	0.538
Model 2 (Linear interaction)	0.365	-0.395
Model 3 (Quadratic)	0.363	-1.180
Model 4 (Quadratic interaction)	0.370	4.572
Panel B: Models with baseline variables (FSM, IDACI, and EAL)		
Model 5 (Linear)	0.377	0.595
Model 6 (Linear interaction)	0.378	-0.429
Model 7 (Quadratic)	0.376	-1.224
Model 8 (Quadratic interaction)	0.384	5.209

Table 2: The parametric estimation of the treatment effect

	R-square of the model	Treatment effect
Panel A: Models without baseline variables		
Model 1 (Linear)	0.202	2.847
Model 2 (Linear interaction)	0.203	2.684
Model 3 (Quadratic)	0.203	2.212
Model 4 (Quadratic interaction)	0.203	5.039
Panel B: Models with baseline variables (FSM, IDACI, and EAL)		
Model 5 (Linear)	0.218	3.022
Model 6 (Linear interaction)	0.219	2.824
Model 7 (Quadratic)	0.219	2.276
Model 8 (Quadratic interaction)	0.220	5.554

Table 3: The robustness check of the parametric estimation (trimmed the outermost 10% at both ends of the assignment variable)

Figures:

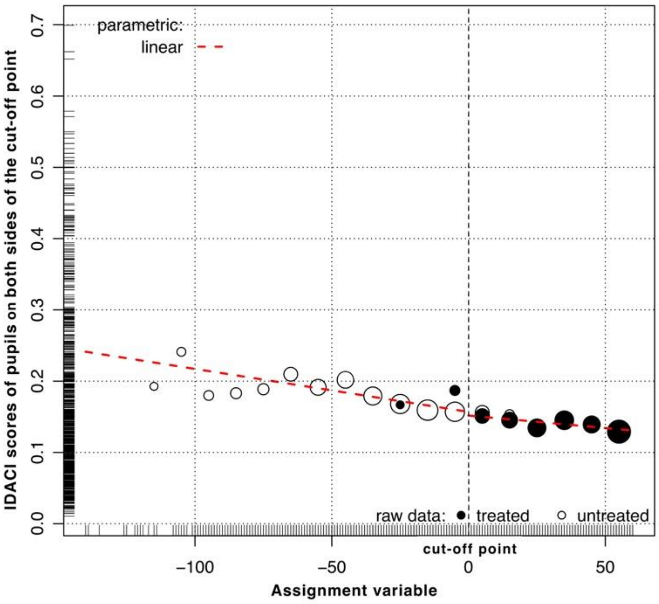

Figure 1: IDACI scores on both sides of the cut-off point

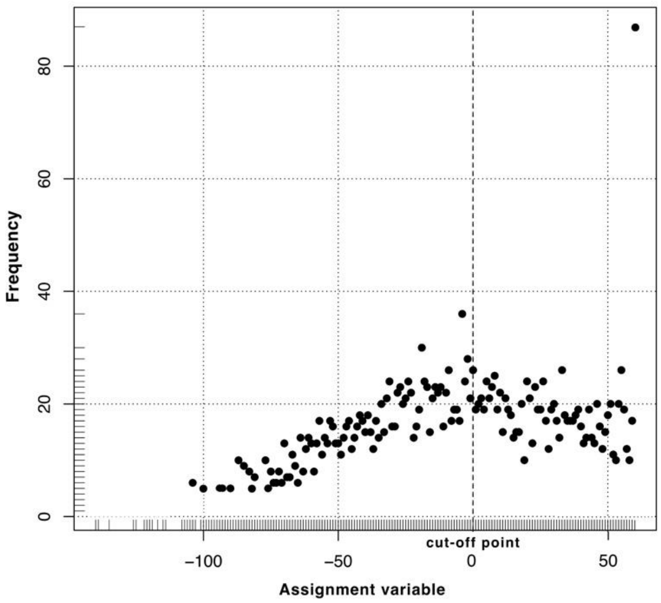

Figure 2: Frequency of the assignment variable

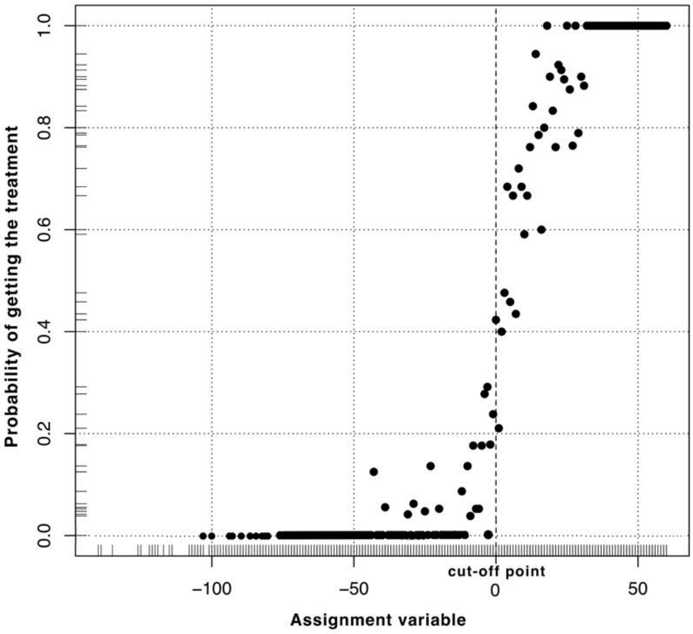

Figure 3: Probability of getting the treatment

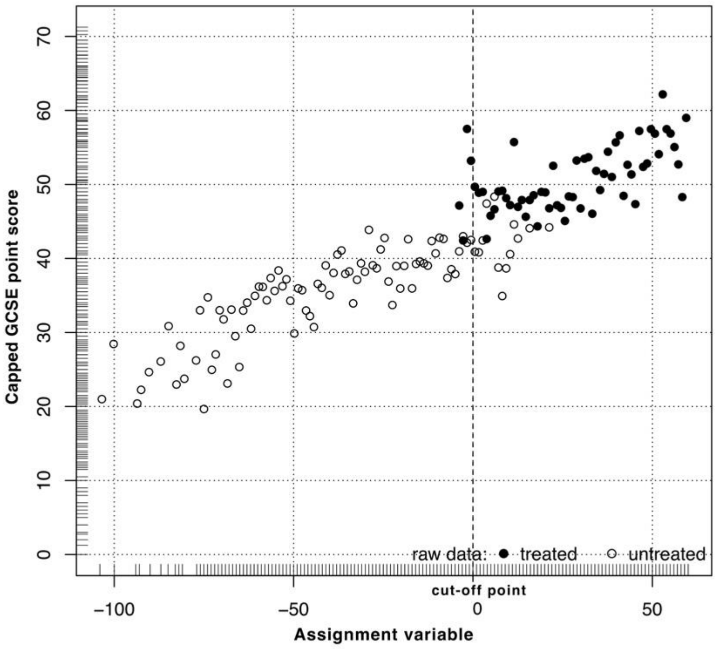

Figure 4: The relationship between the assignment variable and the outcome variable (raw data)

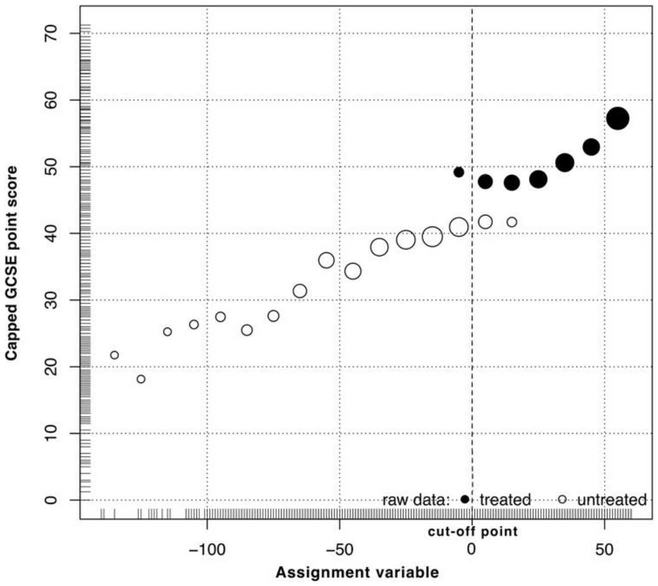

Figure 5: The relationship between the assignment variable and the outcome variable (grouped data)

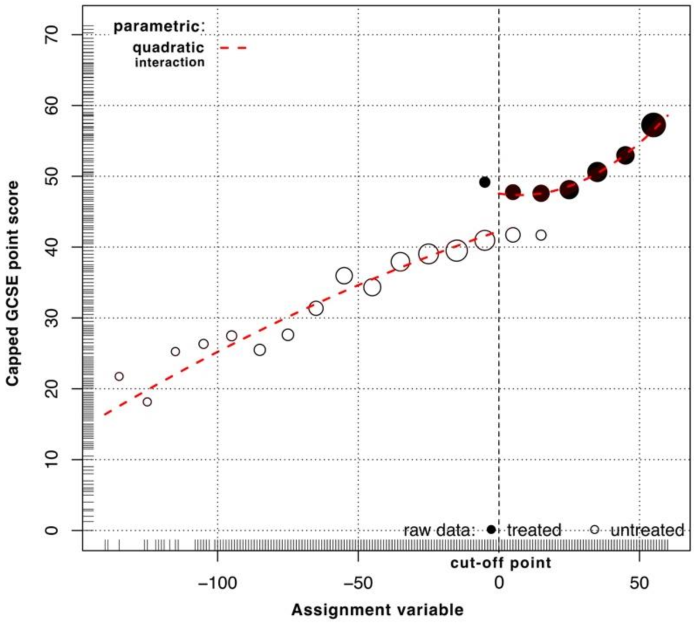

Figure 6: The treatment effect in the parametric approach (quadratic interaction regression lines superimposed)

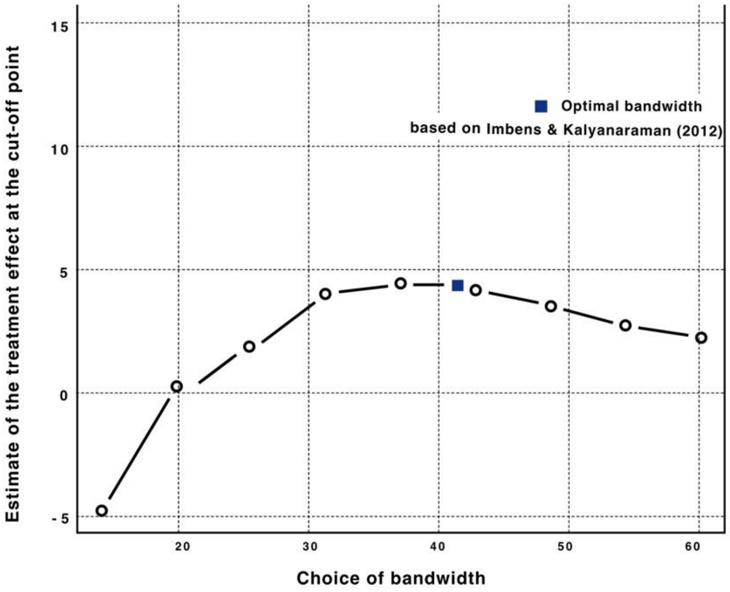

Figure 7: The treatment effect in the non-parametric approach and the choice of bandwidth

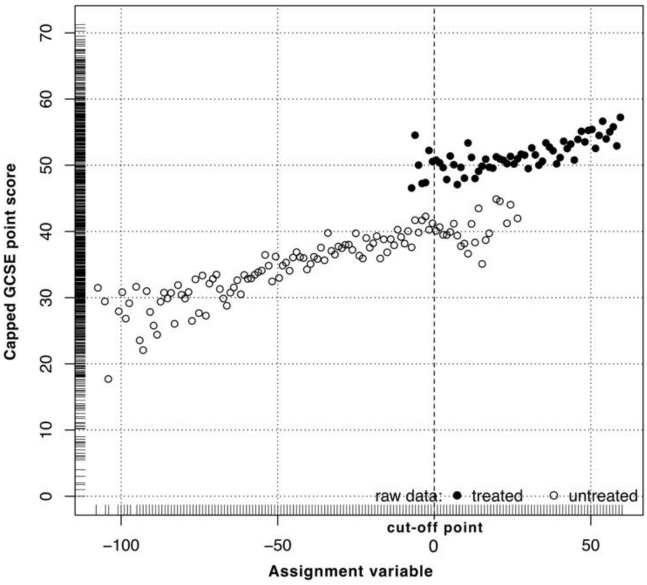

Figure S1: The relationship between the assignment variable and the average capped GCSE of each subgroup (raw data)

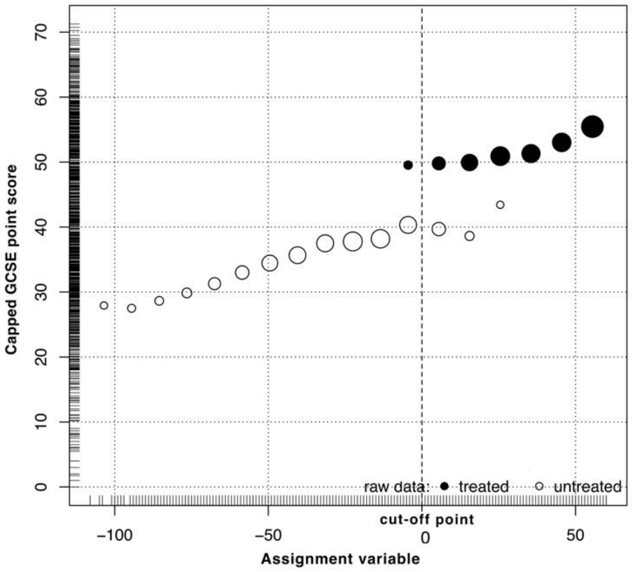

Figure S2: The relationship between the assignment variable and the average capped GCSE of each subgroup (grouped data)

Appendix 1 - Methods and results of previous studies of the effectiveness of grammar schools in England

The table below includes detailed methods and findings of previous literature discussed in the main text. Since some studies have applied extensive sets of explanatory and outcome variables, it is not possible to include all their results in detail, and only general patterns are presented for comparison. Studies are arranged in the same order as appeared in the main text.

The table excludes studies applying the National Child Development Study. In these studies, the sample group attended secondary schools from 1969 to 1974, when the reform of comprehensive schools occurred. Therefore, the results have limited relevance to today's schools and are not discussed.

Schagen & Schagen, 2003	
Area	England
Data	1) 1998 KS3 to 2000 KS4
	2) 1995 KS2 to 1998 KS3
Comparison group	Grammar schools vs. comprehensive and secondary modern schools
Model	Multiple regression and multilevel modelling
Explanatory variables	Pupil level: prior attainment, gender, age
	School level: school type, size of school, proportion of FSM pupils
	LA level: proportion of grammar school places of LA
Outcome variable(s)	1) Total GCSE point score, average GCSE point score 2) KS3 attainment level
Indicator of outcome	Coefficients
Effect of grammar schools	1) 3 grades on total GCSE or 0.4 grade per GCSE subject for average pupils, no effect for high performers
	2) 0.8 level at KS3 for average pupils (about one year's progress), no effect for high performers
Atkinson, Gregg & McConnell, 2006	
Area	19 LAs with more than 10% grammar school places
Data	1997 KS2 to 2002 KS4
Comparison group	Grammar schools vs. comprehensive schools in non-selective LAs
Model	OLS multiple regression
Explanatory variables	Pupil level: prior attainment, gender, age, ethnicity, FSM, SEN, English as a second language
	School level: school type, school size, single sex
Outcome variable(s)	Total GCSE point scores, Capped GCSE point scores
Indicator of outcome	Coefficients
Effect of grammar schools	4 grades on total GCSE or 3 grades on capped GCSE
Levačić & Marsh, 2007	
Area	England
Data	1996 KS2 to 2001 KS4

Comparison group	Grammar schools vs. comprehensive schools in non-selective LAs
Model	Multilevel modelling, logistic regression
Explanatory variables	Pupil level: prior attainment, gender, age
	School level: school type, proportion of FSM pupils, proportion of pupils with SEN statements, proportion of white pupils, pupil-teacher ratio (average for 1997-2001)
Outcome variable(s)	Total GCSE/GNVQ points score, The probability of obtaining 5 or more A*-C GCSE/GNVQ
Indicator of outcome	Coefficients
Effect of grammar schools	6 grades on total GCSE/GNVQ or 25% more likely to achieve 5 A*-C on GCSE/GNVQ
Harris & Rose, 2013	
Area	Buckinghamshire
Data	Borderline pupils in grammar schools and secondary modern schools, who took GCSE between 2007-2009
Comparison group	Grammar schools vs. secondary modern schools
Model	Logistic regression
Explanatory variables	Matched by pupil’s prior attainment, gender, FSM, birth month, Pakistani or not, year of examination
Outcome variable(s)	The probability of obtaining 5 or more A*-C GCSE
Indicator of outcome	Coefficients
Effect of grammar schools	10% more likely to achieve 5 A*-C on GCSE/GNVQ
Gorard & Siddiqui, 2018	
Area	England
Data	2010 KS2 to 2015 KS4 (same results for 2014 and 2016 KS4 cohorts)
Comparison group	1) Grammar schools vs. other state schools 2) Grammar schools vs. other state schools in selective LAs
Model	Multi-stage regression models
Explanatory variables	Pupil level: prior attainment, gender, age, ethnicity, KS4 FSM eligibility, whether a pupil has been eligible for FSM in any of the past six years, the number of years in total a pupil was eligible for FSM up to KS4, IDACI, SEN status, English as an additional or second language, whether the pupil moved to the school in the last two years
	School level: school type, segregation residual for FSM status (the amount by which each school’s intake deviates from the national average)
Outcome variable(s)	Capped GCSE point score
Indicator of outcome	The R value of regression models
Effect of grammar schools	No effect

Coe et al., 2008	
Area	England
Data	2001 KS2 to 2006 KS4
Comparison group	Grammar schools vs. other schools
Model	OLS and multilevel modelling
Explanatory variables	Pupil level: prior attainment, gender, ethnicity, FSM, IDACI,
	School level: school type, average KS2 level, proportion of FSM pupils, average IDACI, single sex
Outcome variable(s)	Total points score on GCSE and equivalents Capped points score on GCSE and equivalents
Indicator of outcome	Coefficient
Effect of grammar schools	0-3/4 grade per subject on GCSE and equivalents

Appendix 2 – Identifying the treatment effect in the ‘fuzzy’ RDD

In an ideal ‘sharp’ RDD, all individuals who have passed the cut-off point would get the treatment and those who have missed it would not. The probability of treatment jumps from 0 to 1 at the cut-off point. However, in a ‘fuzzy’ RDD, as the jump in the probability of treatment at the cut-off point is lower than 1, the discontinuity at the outcome cannot be simply regarded as the treatment effect. The treatment effect in a ‘fuzzy’ RDD needs to be recovered by calculating the ratio of the gap in the outcome variable and the gap in the probability of the treatment at the cut-off point (Jacob et al., 2012; Lee & Lemieux, 2009). Therefore, both the outcome variable and the treatment probability on the two sides of the cut-off point need to be calculated. Accordingly, the estimation can be written as:

$$Y_i = \theta + \pi D_i + f_0(X_i) + \mu_{0_i},$$
$$T_i = \eta + \lambda D_i + g_0(X_i) + v_{0_i},$$

where Y_i is the outcome measure for each individual i; T_i is the treatment dummy; X_i is the assignment variable (X_i=0 is the cut-off point); D_i is the binary indicator of whether individual i reached the cut-off point (D_i=1 if $X_i \geq 0$); μ and ν are the random error for each individual. The coefficient π is the ‘intend to treat’ effect, and the real treatment effect equals the ratio of π/λ . This is the ratio of the discontinuity in the outcome and the discontinuity in the treatment at the cut-off point, as mentioned above. Analytically, the regression equations of the treatment effect can be transformed into equations which have been presented in the main text:

$$Y_i = \alpha + \beta T_i + f(X_i) + u_i, \tag{1}$$
$$T_i = \gamma + \delta D_i + g(X_i) + v_i, \tag{2}$$

where every variable is the same as the above, and the effect of attending grammar school which needs to be estimated now equals β.

As mentioned in the main text, when fitting functional forms in equation (1) and (2), the slopes of the regression lines are allowed to vary on two sides of the cut-off point. This is realised through including interaction terms between the assignment variable (*X*) and the treatment variable (*T*), as well as the interaction between the assignment variable (*X*) and the cut-off point variable (*D*).

Appendix 3 - Alternative sampling strategies

Instead of only keeping cases which can be uniquely matched, Figure S1 (raw data) and Figure S2 (grouped data) below present the difference between grammar schools and non-selective schools using all of the cases in the 11+ file (including those with the same combination of baseline variables). Pupil records in the 11+ file are still matched to the NPD based on characteristics, KS2 attainment, and school type. However, instead of linking to the individual capped GCSE result as the KS4 attainment indicator, the average capped GCSE result of each subgroup with the same combination of background variables is calculated for grammar school pupils and their counterparts. While it is impossible to elucidate the individual relationship between the 11+ result and the capped GCSE result in this way, it presents an overall picture of how each group of pupils with the same characteristics are doing in grammar schools and in non-selective schools. Figure S1 presents the average GCSE result at each 11+ score. Meanwhile, cases are also grouped into bins, and the number of observations within each bin is represented by the size of the dot (Figure S2). In total, 6,732 records in the 11+ file are matched to the NPD data.

Another alternative for dealing with potential misspecification between the indicator of passing the selection (the 11+ file variable) and real attendance in grammar schools (the NPD variable) is to first match all grammar school pupils in the NPD to all pupils who passed the selection in the 11+ file. Then, the non-grammar school pupils in the NPD can be matched to the remaining unmatched cases in the 11+ file. However, after matching, the sample group is different from the original cohort in the 11+ file, with the FSM proportion in the former being twice as high as in the latter. Therefore, an analysis based on this sample group was not conducted.