How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England

Preliminary and incomplete – please do not cite

October 2011

Rebecca Allen, Institute of Education, University of London

Simon Burgess, Centre for Market and Public Organisation, University of Bristol

Abstract

School inspections are an important part of the accountability framework in education in England. In this paper we use a panel of schools to evaluate the effect of a school failing its inspection. We collect a decade's worth of data on how schools are judged across a very large range of sub-criteria, alongside an overall judgement of effectiveness. We use this data to implement a fuzzy regression discontinuity design to model the impact of 'just' failing the inspection, relative to the impact of 'just' passing. This analysis is implemented using a time-series of school performance and background data. Our results suggest that schools only just failing do see an improvement in scores over the following two to three years. The effect size is moderate to large at around 10% of a pupil-level standard deviation in test scores. We also show that this improvement occurs in core compulsory subjects, suggesting that this is not all the result of gaming on the part of schools. There is no positive impact on lower ability pupils, with equally large effects for those in the middle and top end of the ability distribution. In separate analysis we show that failed judgements reduce subsequent demand for places at the school.

Keywords: school inspection; school accountability; school attainment; regression discontinuity

JEL codes: I20; I28

Acknowledgements:

We are very grateful to Ofsted for providing data on Ofsted inspections and to the Department for Education for providing access to the National Pupil Database.

1. Introduction

What is the best policy for dealing with under-performing schools? Most education systems have a mechanism for identifying poorly performing schools[1], typically an inspection system where an external body is responsible for monitoring and reporting on educational standards within schools. But what should then be done with schools which are highlighted as failing their pupils? There are important trade-offs to be considered: rapid intervention may be an over-reaction to a freak year of poor performance, but a more measured approach may leave many cohorts of students to under-achieve. Are penalties or greater support more appropriate to turn a school around? Should the process be left to the market as new pupils shun failing schools, or is a more pro-active policy required if neighbourhoods remain 'loyal' to the local school despite its results? At what point should closure be contemplated?

In this paper, we evaluate the effect of the policy towards poorly performing schools in England. Such schools are identified by a national school inspection body, Ofsted, the Office for Standards in Education, Children's Services and Skills.[2] On the basis of its inspections, Ofsted judges some schools to be 'failing' and this triggers a set of policy actions in those schools. These are detailed below. We implement a regression discontinuity design (RDD) in a panel data context, comparing the time series of performance statistics for schools that are designated as just failing with those just passing. The outcome measure we use is the high-stakes national tests that all students in England take at the end of compulsory schooling. This estimates the causal impact of that policy on school performance. The intuition behind an RDD in this context is that schools around the failure threshold are very similar, except for random measurement of quality by inspectors that causes some to just pass their inspection while others just fail. Our identification strategy is based on defining a running variable to capture continuous variation between schools, on top of which is the discontinuity of a discrete judgement of 'fail' or 'pass'. We then compare the time series differences in performance for such schools.

In principle, the effects of 'failing the Ofsted' could go either way: it could be a catalyst for improvement or a route to decline. It could lead to improved performance if it induces the school to work harder to pass a subsequent inspection: more focussed leadership and more effective teaching could raise test scores. On the other hand, if students and teachers leave the school, the failure may trigger a spiral of decline where falling rolls lead to low morale, financial difficulties and even lower test scores. So it is a meaningful question to ask whether test scores improve or worsen following the treatment.

An RDD is considered to have strong internal validity, but weaker external validity. Schools which fail their inspection by a considerable margin may react very differently to their failure. The RDD would overstate the effect of failing if those schools which are far from the margin have little capacity to improve sufficiently and so cannot respond to the accountability pressure. This is in fact what our data suggests, so we do not extrapolate our results to all failing schools.

This paper contributes to a large international literature on the effect of school accountability systems. In the US a large literature has focussed on the accountability mechanisms built into the No Child Left Behind Act[3]. Our RDD approach and our findings bear similarities to Cooley and Traczynski (2011) and to Ahn and Vigdor (2009), who both find that schools in the North Carolina facing sanctions under No Child Left Behind improve performance. The former confirms Neal and Schanzenbach's (2010) finding that schools facing these forms of accountability tend to focus on students at the threshold of performance measures at the expense of the lowest performing pupils.

[1] See Faubert (2009)
[2] See http://www.ofsted.gov.uk/about-us , accessed 10/10/11.
[3] This includes Reback (2008), Jacob (2005), Neal and Schanzenbach (2010), Rockoff and Turner (2008), Figlio and Rouse (2006), and Chakrabarti (2010), Dee and Jacob (2009), Krieg (2008); Ladd and Lauen (2010], and Ahn and Vigdor (2009).

In the UK, there are fewer studies of the dual accountability systems of publicly published performance tables[4] and Ofsted inspections. Rosenthal (2004) studies the impact of Ofsted visits and has a negative conclusion: there is no gain after the visit and there is a fall in performance in the year of the visit. Using a more sophisticated econometric approach, Hussain (2011) identifies the very short-term impact of failing an Ofsted. He compares the performance of schools failed early in the academic year with those failed later in the academic year, specifically after the exams. This identification strategy compares like with like, and isolates the effect of having some 8 months to respond to the failure. Our approach differs by focussing on the high-stakes exams at the end of secondary school (he uses primary schools), by using data over a run of 9 years, and by adopting another identification strategy that allows us to leverage in more data.

We find that schools failing their Ofsted inspections improve their subsequent performance relative to the score in the pre-visit year. The magnitudes are quantitatively very significant: around 10% of a (student-level) SD. The main impact arises two years after the visit in this data – not unreasonable given that the exam scores we use derive from two-year courses. The typical time pattern in our results is little effect in visit year +1, increasing considerably the following year, remaining flat or slightly increasing in the third post-visit year. We also show that these results do not come from simply gaming the exam system and entering the pupils for easier exams. We show comparable, in fact typically higher, effects in core compulsory subjects of maths and English. This suggests that the improvements in teaching quality are genuine. These results may be underestimates if just-passing schools also react to their result. This may be an internally generated reaction, the schools work hard to avoid future failure, or externally generated if the Ofsted inspectors make it clear that a lot of work is needed. Other results looking at longer-run differences suggest that post-Ofsted failure policies are successful at remedying short-run declines in performance, but are less successful in permanently changing low performance. We argue that this is still a very valuable outcome and, given the nature of the treatment, is not surprising.

In the next section we set out the policies in England regarding school inspections and actions towards poorly performing schools. In section 3 we describe our data and section 4 sets out our identification and estimation strategy. The results are in section 5. In the final section we conclude with some comments on the role of these policies in a suite of responses to schools in different circumstances.

2. Policy Background

Most countries in the world operate a school inspection system where an external body is responsible for monitoring and reporting on educational standards within schools. The nature and purpose of the inspection systems varies considerably. Coleman (1998) describes a continuum with one extreme as the objectives-based approach which has the features of being summative, external and formal, focusing on simply judging whether externally pre-determined objectives for education are being achieved. These types of high-stakes inspections play an important part in an accountability framework by giving parents and officials information about school quality that they can act on, should they wish to. At the other end of a continuum of inspections is a process-based approach which is much more formative, internal and informal and looks to take value from the process of inspection as it advises, supports and helps to improve the education provider. The English schools inspectorate is firmly at the high stakes end of this spectrum.

[4] For example, Burgess et al (2010) show that the publication of league tables does have an impact on school performance, and Allen and Burgess (2011) show that league tables are useful to parents in choosing schools.

2.1 The English school inspection system

The Office for Standards in Education, Children's Services and Skills, known as Ofsted, was created in the 1992 Education Act as part of a new era of parental choice and accountability. It is a national body that provides regular independent inspection of schools with public reporting to both parents and parliament and the provision of advice to ministers. The current intended role of Ofsted is "*to provide an independent external evaluation of a school's effectiveness and a diagnosis of what the school should do to improve, based on a range of evidence including that of first-hand observation*" (Ofsted, 2011a, page 4). The inspectorate has had at its focus the need to make schools accountable for their performance and thus has placed great emphasis on the use of external examination results along with 'snapshot' external inspections to judge schools (Learmonth, 2000). The criteria on which schools are evaluated are both objective, such as exam results, and subjective, such as the inspectors view of teaching quality observed during inspections. While the model is formal and accountability focused, giving both judgments on individual schools and the system as a whole it does have elements of support and improvement through mechanisms such as school action plans.

England's inspection process imposes a relatively high cost on the education system (£207 million or 0.27% of the total schools budget according to Hood et al., (2010)). Added to these costs can be the loss of time by schools as they prepare and the cost of production of materials needed specifically for the inspection. Several early studies showed a small but negative effect on exam results for schools in the year of inspection, suggesting that the preparation time for an Ofsted inspection diverts teacher focus from pupil exam preparation (Rosenthal, 2004; Culliford and Daniels, 1999). However, this negative impact appears to be have shrunk towards zero in recent years, perhaps because the notice period for an inspection is now very short (Allen and Burgess, 2010)).

The legal requirement is for schools to be able to be inspected on a five school year cycle, but since 2005 the frequency of inspections is proportionate to perceived need such that schools judged satisfactory will be on a three year cycle and schools judged as inadequate will be visited far more frequently without notice. The period of notice schools receive before an inspection has shrunk over time from over two months' notice to the current period of between zero and two working days, with no notice possible where there are concerns relating to pupils' welfare, safeguarding, where the school's academic performance has shown rapid decline or where there is a voice of concern raised by parents.

School inspections are currently conducted by Her Majesty's Inspectors (HMI), who are employed by Ofsted, and additional inspectors, who are employed full-time, freelance or otherwise contracted by Ofsted's inspection service providers (CfBT, Tribal and Serco). The intensity of the visits has fallen considerably over time, with full scale week-long visits by a large team of inspectors prior to the 2005 inspection reforms. This regime was criticised by teachers and school heads as greatly disruptive to the operation of the school. Since September 2005 visits are shorter in duration (around two days) and reliant on fewer inspectors in the school. They are more sharply focused on the school's own self-evaluation of their strengths and weaknesses and particularly focus on management, including how well the school is managed, what processes are in place to ensure standards of teaching and learning improve, and how well the management team understand their own strengths and weaknesses.

Before a visit, inspectors draw on attainment data, school performance indicators and reports from previous inspections to decide how to plan their inspection. Parents and staff are always invited to give their views on the school in a short questionnaire. During the visit, inspectors usually talk with pupils, governors, management and staff to gather specific views on provision, observe a large number of lessons, 'track' individual pupils to monitor provision for specific groups and scrutinise school records and documentation (Ofsted, 2011a).

Inspectors make a series of sub-judgements about different aspects of the school. For example, during 2009 judgements were made about pupil outcomes (7 sub-criteria), quality of provision (3 sub-criteria), leadership and management (8 sub-criteria), and early years, sixth form and boarding provision where relevant (Ofsted, 2011a). Under the post-2005 framework a school immediately receives one of four overall judgements: outstanding, good, satisfactory or unsatisfactory/inadequate (Table 1 shows the greater number of judgements in earlier years of the data). This overall judgement first because very prominent in the inspection report in the academic year 2002/3. This judgement is received alongside a publicly available inspection report that is delivered to the school within two weeks.

Those schools judged to be inadequate are deemed to have 'failed' their Ofsted inspection. These schools are currently split into two categories of schools causing concern: notice to improve and special measures. Notice to improve means that "*the school requires significant improvement because either: it is failing to provide an acceptable standard of education, but is demonstrating the capacity to improve; or it is not failing to provide an acceptable standard of education but is performing significantly less well in all the circumstances reasonably be expected to perform"* (Ofsted, 2011a, page 12). These schools are subject to no operating restrictions and will simply receive a monitoring inspection between six and eight months after their last section 5 inspection.[5] They will also be fully inspected around a year after their notice to improve was served. The headteacher, chair of the governing body and a representative from the local authority or proprietor will have been invited to attend a school improvement seminar, but there is no requirement for them to attend. The school does not need to prepare an action plan, but is expected to amend their existing school plans in light of the judgement and submit this to Ofsted within 10 working days (Ofsted, 2011c).

Special measures is a more serious judgement against a school, meaning that "*the school is failing to give its pupils an acceptable standard of education and the persons responsible for leading, managing or governing the school are not demonstrating the capacity to secure the necessary improvement in the school*" (Ofsted, 2011a, page 12). It was introduced as a measure under Section 14 of the Schools Inspection (1996) Act and has more serious consequences than a notice to improve, such as restrictions on the employment of newly qualified teachers. In addition the revision of action plans and invitations to attend seminars outlined above, school under special measures will receive up t give monitoring inspections over two years. It is possible for inspectors to judge that sufficient progress has been made at any of these monitoring inspections, but if special measures have not been removed after two years a second full inspection is held (Ofsted, 2011b). Where a school is still judged to be inadequate a year after the first inspection, the Department for Education requires the local authority to examine carefully the options available to it, including replacement of the governing body by an Interim Executive Board, dismissal of senior managers and teachers and even full closure of the school.

2.2 Model of school and parental behaviour

School inspection has the potential to contribute to improving the efficiency of the schooling system overall, either by making recommendations that improve school effectiveness, or by acting a threat that incentives schools to maintain high standards, or because it leads to the shrinkage or closure of schools with poor inspection judgements. Inspectors might be able to highlight serious problems for schools, assess individual teachers and suggest new ways of working (OECD, 1995).

There are a number of possible mechanisms of the school's response to receiving an unsatisfactory Ofsted rating. Both the stigma of failing and the threat of closure in the event of no improvement are both likely to act as the most important incentives to improve. Schools may improve their overall

[5] Some schools judged to be satisfactory will also receive a monitoring visit.

exam results because teachers work harder, without modifying their education production function. This would be possible if there is slack in the system with some teachers no currently operating at their optimum. Alternatively, individual teachers may direct effort away from broader learning experiences and more directly towards exam preparation. The school itself may decide to develop a more targeted approach of focusing on students who have the greatest capacity to improve or reach some threshold. Or in the longer term it can switch the mix of subjects offered to students towards a set where examination success is more likely.

Of course, the extent to which these responses manifest themselves depends on whether teachers are willing and able to respond. In many countries teachers have a great deal of autonomy and privacy within their classroom (Joyce and Calhoun, 1991) with little discipline or accountability for teachers. Furthermore, there is little praise or reward where teachers do a good job (Heneveld and Craig, 1995). Negative inspection judgements may have little effect because of resistance to change institutionally and personally by teachers (Bush and West-Burnham, 1994).

Information on inspections introduces market forces into education by providing information on quality to consumers, allowing them greater freedom to choose an educational product. A poor Ofsted judgment is likely to result in a fall in applications to a school by parents, which is likely (but not certain) to reduce school income for future years.

<<Formal models to follow>>

3. Data

This paper uses data from two sources. The National Pupil Database (NPD) provides information on pupil test scores from 2002 to 2010. The Ofsted database of school inspection judgements gives information on school visits from the academic year 2002/3 to 2009/10.

3.1 Ofsted Data

We use the publicly available Ofsted Database of school inspections for each year from 2002/3 to 2009/10. This database gives information on every visit made by Ofsted over this time. The schools are judged on a large number of sub-criteria such as value for money and quality of leadership – 19 in 2002/3 and as many as 65 in 2007/8. The sub-criteria are almost always ranked on a scale from 1 (outstanding) through 2 (good), 3 (satisfactory), to 4 (unsatisfactory), although there are some binary indicators of whether standards are met. We also have an overall ranking for the school on a scale of 1 to 4 (unsatisfactory). We exclude the second visit where two take place in consecutive years for a school. Table 1 summarises the Ofsted inspection data for each year.

Not all schools that failed their Ofsted inspection did so equally badly. A regression discontinuity design requires us to identify the set of schools that came close to just passing or 'just' failing their Ofsted inspection. We are able to do this using the large number of sub-criteria that judge how a school performs on a wide variety of metrics. Failing can take place across multiple dimensions, such as unsatisfactory behaviour of learners, but a regression discontinuity approach requires us to create a single dimension running variable of the extent of the fail/pass for each school visit. The idea is that schools that 'just' fail their inspection will only fail a limited number of sub-criteria, compared to schools that 'comprehensively' fail their inspection.

We face a problem because there is no set rule for turning between 19 and 65 sub-scale measures into an overall measure of the judgement for the school. Inspectors are able to exercise an element of discretion in their decision as to whether a school fails their Ofsted inspection. Furthermore, both the sub-criteria and rules for interpreting the sub-criteria changed several times throughout the period of analysis. We use many different approaches to aggregating the sub-criteria information to create an assignment variable, including counting numbers of fails on sub-criteria, weighting scores

for each of the sub-criteria and estimation of the propensity to fail given the sub-criteria. All are highly correlated and our results are robust to using alternative running variables. Our chosen assignment variable is simple to interpret because it counts the number of sub-criteria the schools fails on plus 0.1 * the number of sub-criteria it achieves only a satisfactory rating. For each year we scale our data so that zero represents the approximately discontinuity between passing and failing schools. Our chosen running variable does not perfectly divide schools into those that fail and those that do not because we do not know the assignment rule for failing and in any case we understand that there is discretion on the part of school inspectors. We discuss our approach to dealing with this in the Section 4.

3.2 National Pupil Database (NPD)

We match the Ofsted database to school-level information aggregated from the National Pupil Database (NPD). NPD is an administrative database of pupil demographic and test score information from 2001/2 onwards. It provides basic socio-demographic information on each child in state-maintained schools that is aggregated in this paper to give annual summary statistics on a school's year 7 (age 11/2) and year 11 (age 15/6) cohort. Information on the year 7 cohort is used to describe changes in the intake profile of the school. Information on the year 11 cohort is used to describe changes in school effort and attainment for that year of school leavers.

The size and composition of the year 7 cohort are measured using:

- Number of pupils in year 7;
- Proportion of year 7 pupils eligible for free school meals (FSM), an indicator of level of pupil poverty in the school (see Hobbs and Vignoles, 2010, for drawbacks on this measure);
- Average Key Stage two test result (normalised across pupils as a z-score) for the year 7 cohort of the school. This is a measure of prior attainment of cohort since the tests are sat at the end of primary school at age 10.

The achievement of the year 11 cohorts is measured using a broad outcome measure of the average score across all pupils in their best 8 subjects at GCSE (exit exams at age 16 typically taken in 7-12 subjects). This broad measure – 'capped GCSE' – is standardised across all pupils as a z-score. We also report outcomes for the proportion of pupils achieving five or more 'good' GCSEs at grades A*-C and for average school grades in English and maths measured on a scale of 0 to 8. The threshold measure – '%5AC GCSE' – is reported in school league tables throughout the period of analysis.

In the analysis of the year 11 cohort we use key socio-demographic statistics to control for the changing characteristics of the school cohorts. These are:

- The proportion of pupils eligible for FSM;
- The proportion of white British ethnicity pupils in the cohort;
- The proportion of pupils who are female;
- The proportion speaking English as an additional language;
- The average level of deprivation for the neighbourhoods the pupils live in, measured using the Index of Deprivation Affecting Children Index (IDACI);[6]
- The average pupil prior attainment at Key Stage 2 (end of primary school) in maths, English and science.

Summary statistics for all these variables for each cohort are in Data Appendix Table 1.

[6] The Income Deprivation Affecting Children Index (IDACI) comprises the percentage of children under 16 living in families reliant on various means tested benefits (see http://www.communities.gov.uk/documents/communities/pdf/733520.pdf).

4. Identification and estimation

Our major issue is the endogeneity of failure. We implement a variety of solutions that we describe here. We have data from before and after the inspection so are able to remove some of the bias this way. Next we treat the discontinuity as sharp and run a parametric RDD. Then we take account of the fuzziness in the fail assignment around the threshold and take an IV approach. Finally, we exploit the full panel of data to run a difference-in-differences with a variety of bandwidths. We discuss each of these in turn.

4.1 Before-after analysis

We initially implement a simple before-after analysis of the impact of inspection on exam outcomes, by estimating the following equation:

$$\Delta_\tau Y_s = \beta_0 + \beta_{\text{fail}} fail_s + \beta_{\text{inspyear}} inspyear_s + g\,(\Delta_\tau x_s, x_{s,t-1}) + \varepsilon_s$$

We do this for a variety of difference windows, windows: $\Delta_\tau Y \equiv Y_{t+\tau} - Y_{t-1}$, for $\tau = 1, 2, 3, 4$. Our outcome variable ΔY_s is the change in GCSE results for school *s* over the window. $fail_s$ is a binary indicator for whether the school received an unsatisfactory rating for their inspection, or not. We include dummy indicators for the inspection year, $inspyear_s$, to mop up varying rates of grade inflation in our outcome variables over the time period. We include variables for the pre-treatment level and change over period for a set of school observable composition measures, *x,* that directly affect exam outcome. These are: three variables which measure mean prior attainment of the cohort in English, maths and science; percentage free school meals eligibility; average deprivation level; percentage female; percentage non-English speakers; and the percentage of the school population who are white British.

Clearly this analysis ignores endogeneity of Ofsted failure. Schools that failed their Ofsted inspection are likely to be different from schools that pass in unobservable ways: a simple OLS regression of the level of school attainment would be negative biased, but this before-after regression analysis could equally be positively biased. That said, our treatment has characteristics that do make it amenable to this before-after design since the treatment is clearly defined, takes place quickly, and the effect should be felt quickly before other covariates change (Marcantonio and Cook, 1994).

4.2 Parametric regression discontinuity with before-after data

The estimation approach we describe here uses a regression discontinuity design (RDD) to identify the effect of failing a school inspection on a school's future outcomes, y_s, using schools that passed their inspection as a control group. This is the only available choice of control group since all schools in England are regularly inspected. The dichotomous treatment of failing an Ofsted inspection, *fail*, may have multiple effects on the school since it is a public statement of quality that might impact pupil and teaching sorting but it also brings additional support to the school to help it improve. This treatment is a function of the inspectors' judgments on a wide variety of criteria that are collapsed to a single continuous covariate, *rating*. We do not know the exact rule for the assignment of treatment in each year, so *rating* is constructed from the underlying sub-criteria using the method described in data section 3.1.

We clearly cannot estimate $E[y_1 - y_0]$ as $E[y_1 | fail = 1] - E[y_0 | fail = 0]$ because a set of observed and unobserved characteristics of schools, x_s, such as school and teacher talent and effort and pupil background alter both the probability of receiving the treatment (i.e. failure) *and* future pupil exam performance. The RDD approach assumes subjects near the threshold are likely to be very similar and thus comparable. This 'threshold' randomisation identifies $E[y_1 - y_0 \mid rating \approx 0]$, provided Hahn et al.'s (2001) minimal continuity assumptions hold.[7] It can be interpreted as a weighted average

[7] We haven't applied McCrary's (2007) test of manipulation, which relates to continuity of the running variable density function, because manipulation would have taken place at the sub-criteria level rather than this level.

treatment effect for the entire population, where the weights are the probability that the school draws a *rating* near the threshold (Lee, 2005a). This means we can infer little about the potential effects of failing an Ofsted inspection for those schools who were judged unsatisfactory across a large number of sub-criteria or across none at all.

Applying a regression discontinuity design to this setting is not immediately straightforward because Ofsted failure is determined by a large number of sub-criteria, as described in the data section. This problem has often occurred in the RDD literature (e.g. Bacolod et al, 2009; Ahn and Vigdor, 2009) and bears some similarities to US accountability data used by Cooley and Traczynski (2011) and others because they exploit the set of 10 sub-criteria in No Child Left Behind, producing multi-dimensional scale of closeness to the threshold for failing. Our running variable is a multi-dimension scale based upon between 19 and 65 sub-criteria, depending on the year of inspection visit. Unfortunately there are no strict criteria that are applied by the inspectors to decide how to convert these sub-criteria into the overall judgement, so it is hard for us to accurately measure exactly how far a school should be placed from the discontinuity on a single scale. We deal with this by calculating the running variable as the number of fails on individual sub-criteria plus 0.1 * the number of satisfactory ratings; we normalise this variable around zero. See Section 3.1 for more details. We refer to this running variable as the *rating* variable. For now, we ignore the fuzziness in this measure, but return to it below.

We try to deal with unobservable differences in *x* as we move further from the pass/fail discontinuity by using non-linear approximations to generate simple estimates of the discontinuity gap (up to power 4). These parametric forms are common in the RDD literature (see for example, Angrist and Pischke, 2009), exploit more data than the use of a narrow bandwidth, and can therefore be more efficient. It is also possible that it generates less biased estimates of the true conditional expectation function at the threshold than a simple difference in means on a narrower band, where the true function has a non-zero slope. The critical assumption of this RDD is that the parametric regression function used for extrapolation is correctly specified and this is increasingly important as data further from the discontinuity is drawn upon for efficiency reasons (Lee, 2005a). We do not have good a priori reasoning as to how the *rating* variable should be a function of exam score growth and even pre-treatment data cannot reveal heterogeneities in capacity to benefit from the treatment. We should therefore be cautious in making statistical inferences from parametric regressions. On the one hand, if the polynomials are 'correct', the estimator is efficiently using data that are both close to and far from the discontinuity. On the other hand, if the true functions do not belong to the class of polynomials we select, the discontinuity will in general be biased, and may lead to erroneous inferences of statistical significance (Di Nardo, 2004). We estimate:

$$\Delta_\tau Y_s = \beta_0 + \beta_{\text{fail}} fail_s + \beta_{\text{inspyear}} inspyear_s + h(rating_s, fail_s * rating_s) + g(\Delta_\tau x_s, x_{s,t-1}) + \varepsilon_s$$

In order to capture the idea of the treatment being essentially random close to the threshold, we estimate at different bandwidths. These are data driven, from inspection of the distribution of the rating variable. As always, there is a trade-off between a narrow bandwidth to isolate similar schools, and a broader bandwidth to give greater precision of estimation, given likely effect sizes. Clearly, the wider the band from the pass/fail discontinuity, the higher the likelihood that results are biased by unobservable characteristics underlying the performance of the school.

A kernel estimate of the density of the rating variable is given in Figure 1, along with the fraction of fails for each percentile of the distribution. Note that since the rating variable is the number of fails on sub-criteria a higher value makes an overall fail more likely. Most schools pass their inspections, so most of the weight of the distribution is some way below zero. There is also a very long upper tail to the distribution with some schools failing on most sub-criteria.

Looking at the fail judgements, very few schools fail with a negative rating score; very few pass with ratings between 2 and 4, and none pass with ratings above 10. So while there is clearly a range of fuzziness it is quite narrow both in terms of the measure used and certainly in terms of the mass of the distribution. The range of the running variable for which the realised chance of failure is not zero or one accounts for around 10% of the observations.

We define three bandwidths below the full sample. The "broad" bandwidth comprises all of the 445 fails together with 1454 passes, giving a total sample of 1899 school*visits. The 3,294 passes with the best ratings are excluded. The "narrow" bandwidth is a quarter of the size, a total sample of 521 observations, the highest-rated 252 fails and the lowest-rated 269 passes. Finally, the "very narrow" bandwidth yields just 108 observations, the 55 schools which only just failed and the 53 which only just passed.

We test whether schools just below and just above the threshold for passing are similar in observables, particularly in the pre-test data which allows the partial testing of Hahn et al.'s (2001) minimal continuity assumptions. The results are in Table 2. As we see no jumps in observables, this lends some support to the assumption that there are also no disparities in unobservables that might bias our estimates of the effect of failing Ofsted.

4.3 Fuzzy RDD Analysis

Returning to Figure 1 it is clear that the assignment is fuzzy close to the threshold: the running variable does not perfectly divide schools into those that fail and those that do not. This is because we do not know the rule for the overall judgement relative to the sub-criteria, and in any case we understand that there is some degree of discretion on the part of school inspectors. In this sense our RDD bears similarities to that implemented by Angrist and Lavy's (1999) class size study because we believe the discontinuity is 'sharp' in principle, but the unknown rule and vagaries of implementation by different inspectors makes it slightly fuzzy. To the best of our knowledge we do not believe that the errors in predicting failure are systematic and so biasing our estimates. That said, it is a very good predictor of whether a school fails, as shown in Figure 1, and in Data Appendix 2 which reports the F-statistics from the first stage. Just using the traditional rule of thumb for weak instruments, for all but the very narrow sample, the F-statistics are sufficiently large.

Following standard practice, we use an IV approach. The first stage uses as instruments the polynomial in the running variable, and the threshold indicator of whether the running variable is positive (Angrist and Pischke, 2009)

4.4 Difference-in-differences panel data with regression discontinuity design

Finally we utilise the full panel to get a broader sense of the dynamics of the response to the fail judgement. We are able to include a rich set of student and school characteristics, along with school fixed effects, to deal with any remaining heterogeneity in schools. We estimate this on the narrow bandwidth.

$$\Delta Y_{st} = \beta_0 + \beta_{fail,\,K} fail_s * D(\text{year} = inspyear_s + K) + \beta_{visit}.D(\text{year} = inspyear_s) +$$

$$x_{st} + yeardummies + \mu_s + \varepsilon_{st}$$

For K = 1, 2, 3, 4.

5. Results

We present the main results on school performance, followed by an investigation into how the results are achieved by schools. We then present robustness analysis of the difference-in-difference, RDD and IV assumptions, and finally some results on the possible heterogeneous effects of OFTSED failure. The second set of results relate to the effect of Ofsted failure on demand for places in the school, as measured by the size and composition of the intake cohort in subsequent years.

5.1 Impact of Ofsted failure on school performance

We start with a simple difference-in-difference analysis, treating fail status as if it were exogenous, before going on to utilise the discontinuity.

As we have a panel of up to nine observations per school, we can look at differences over different windows. This is useful because scope for schools to change their practices varies depending on the timescale. It is difficult to achieve immediate (one year) gains as students will be half way through their GCSE courses, teachers will be allocated and so on. Over a two-year window, covering the whole GCSE course, more can be changed. Beyond that, it is a question of the extent to which changes implemented by the school 'stick' or whether the long-run environment of the school – such as its location – re-exert their influence. Also, after five years beyond the event the composition of pupils taking their GCSEs will start to reflect any changes in intake resulting from the failure. We take a pragmatic approach and present results for four different window lengths, and it is lack of data that prevents longer windows.

A first look at the data is presented in Figure 2. This shows the school average normalised capped GCSE score, averaged over all failing schools and all passing schools, before and after the visit date. Two things are very apparent: first, the small minority of schools that fail their Ofsted visits are different to the rest of schools. Much of the paper is devoted to dealing with this issue. Second, the dynamics for these groups are very different. The visit has no discernible effect on the average score of the passing group[8]. For failing schools, however, the average in t+1 is higher than in t-1, and higher still in subsequent years. There is also a decline pre-visit for the schools that fail, reminiscent of 'Ashenfelter's dip' (Ashenfelter, 1978). It may be that this reflects the possibility that in some cases a decline in a school's published performance brought forward an Ofsted visit. However, the post-visit gain in the fail group is such that it exceeds the mean score in t-3 and t-4.

This figure combines all cohorts of schools, so the sample is not the same within in each data point. In Figure 3 we split out three of our cohorts for comparison. Those with visits in 2003 have only one 'before' datapoint, and more 'after' readings; conversely, those visited in 2007 have more 'before' and less 'after' data. The pattern is the same for all these cohorts: no impact on the passing group and a substantial rise in the failing group.

We quantify these effects in the OLS difference-in-difference results in Table 3. The table presents four different difference windows: $\Delta_\tau Y \equiv Y_{t+\tau} - Y_{t-1}$, for $\tau = 1, 2, 3, 4$. Results are given for four different specifications, starting with simply the fail variable and a set of dummies for the year of the visit to control for cohort effects. We add differences of the control variables[9] (the difference of the same order τ as the dependent variable) and then we add the levels of the same set of control variables dated at t-1.

[8] This is not 'ceiling effects' as this group is 90%+ of all schools and none of them have all pupils achieving the maximum.

[9] These are, all averaged over pupils to school level: KS2 English, KS2 Maths, KS2 Science, the fraction of students eligible for Free School Meals, fraction of students female, fraction of students white British, fraction of students with English as an additional language, and neighbourhood poverty score.

There are strong and consistent patterns in the results. In the top three rows, the effect is positive and substantial, and statistically significant. In this table, all of the estimated effects are significant at the 1% level or less. We focus in more depth on the quantitative significance of the effect below, once we have addressed the endogeneity of the fail status, but note here that an impact of around 10% of a pupil SD is a very substantial effect.

The impact increases the further out the window goes, higher for $Y_{t+4} - Y_{t-1}$ than for $Y_{t+1} - Y_{t-1}$. This cannot be interpreted directly as the sample is different in each column – 4384 schools have data to allow us to compute Y_{t+2} but only 2253 have sufficient 'after' periods to yield data for Y_{t+4}. The final row of the table addresses this by imposing a common sample. The pattern is in fact replicated: a minor increase in t+1, a doubling of that effect at t+2, and then further smaller increases in t+3 and t+4.

We now address the determination of the fail status variable and take an RDD approach. To pick up the effect of the unobserved quality of the schools, we include a fourth-order polynomial of the running variable, our school rating variable, and allowing the effect of all these to vary either side of the fail cut-off of zero. The results are presented in Table 4. The top row is equivalent to the "levels" row of Table 3, with these added variables. The effect on the coefficient on fail status is interesting: the coefficients are about 20% lower for $\Delta_\tau Y$ for τ is 1, 2 and 3. This suggests that the rating variable is taking out a good deal of heterogeneity between schools that was previously absorbed into the fail variable. The effects remain strongly statistically significant, apart from for $\Delta_4 Y$, which is considerably smaller. The time pattern also remains the same, doubling between the first and second column and then roughly constant to the third.

The polynomial in the rating variable itself is significant and its exclusion can be strongly rejected. The third and fourth powers can also not be excluded. To give a sense of the direction of the rating variable, Figure 4 presents the fitted values for $\Delta_2 Y$ using just a linear specification. It shows that schools that had only just passed saw a greater increase in their performance than those that with little chance of failing, all else equal. Similarly, schools that had only just failed saw a greater improvement than those with pervasive failings. This all makes a lot of sense and gives some confidence in the role of the running variable. It also makes it clear that we need to narrow the bandwidth around the failure point to focus on alike schools and exclude the always-pass and always-fail units.

Figure 5 displays the difference-in-differences for the four different bandwidths defined above, using the same vertical scale. The narrow bandwidth contains 252 failing schools and 269 passing schools, and the passing schools facing a significant likelihood of failing. The chart shows as intended that the narrower the bandwidth, the more similar are the schools. In all cases other than possibly the narrowest bandwidth, there is evidence of a shift up and subsequent acceleration for failing schools.

This is all made precise in the remaining rows of Table 4. These present the same specification as the top row, but for the broad, narrow and very narrow bandwidths (details of the sample sizes for these are given in Figure 1). Moving to the broad bandwidth in the second row excludes 3,294 observations with very little chance of failing. In this more homogenous population, the impact of the fail status parameter increases substantially from 0.083 to 0.125 for $\Delta_2 Y$. Otherwise the time pattern is the same and statistical significance remains strong. The narrow bandwidth is about a quarter of the size but the effects remain well defined and large for $\Delta_2 Y$ and for $\Delta_3 Y$. The final row of the table displays the results of the very narrow bandwidth and here it is very difficult to estimate effects with any precision: estimating 32 parameters in 103 observations for $\Delta_1 Y$ and 29 parameters in just 40 observations for $\Delta_4 Y$. Nevertheless, the point estimates are stable compared to the broader bandwidths for the first two columns and not significantly different in the third.

The alternative RDD approach described above is to treat this as a fuzzy discontinuity and use the running variable to instrument fail status. The results of this are in Table 6. The top row presents a basic specification, equivalent to that in the top row of Table 3, and the second row adds the differences and levels, equivalent to the third row of Table 3. Focussing on the latter, the coefficients are very similar at around 0.100 for $\Delta_2 Y$ and for $\Delta_3 Y$, and about half that for $\Delta_1 Y$. The rest of the table reports the outcome of the IV procedure for the different bandwidths. At the narrow bandwidth, the effect sizes again follow the same temporal pattern, are about the same magnitude and are precisely estimated for all but the fourth column. None of the coefficients are significant in the very narrow bandwidth, but most of the sizes are comparable to the bigger samples.

The school rating variable is a strong instrument in almost all of the specifications. Some diagnostics are presented in Appendix Table 2. In the very narrow bandwidth, particularly for $\Delta_3 g$ and for $\Delta_4 g$, the F-statistic on the instruments is low.

Finally in this section we utilise the full run of data for each school in a panel regression. This means that we are comparing any given year (say, visit year + 2) to all of the years of data we have for that school, not just to (visit year – 1). We include fixed effects, all the control variables used above, year dummies, and a dummy for the year of the visit. The fixed effects absorb the polynomial in the running variable. We introduce consecutively a dummy for the year after the visit (column 1), two years after the visit (column 2) and so on, and each of these interacted with fail status. This latter is the variable of interest and this is what is reported in each cell of the table. The set of regressions is repeated for each of the four bandwidths. The results are in Table 6.

We see very similar patterns to those reported above for different approaches. The strongest effects are two and three years after the Ofsted visit, and the parameters on these effects are generally stable across the bandwidths. They are smaller than previously, around 0.05 rather than around 0.12; this may be because the (t-1) value is lower than the full 'before' mean. There is no evidence of an effect one year after the visit, and some evidence in the broader samples for an effect at t+4. We consider the implications of the difference between the results in Tables 6 and 4 in the Conclusion.

5.2 Robustness and falsification checks on the performance effects

<<to follow>>

5.3 How do schools achieve their improved performance?

An analysis of more detailed outcome measures yields some useful information on what schools are doing to achieve these gains for their students. In particular, we are interested in whether the improvement is simply due to strategic manipulation of the subjects sat for exams: easier courses are chosen or especially easier courses that offer a high number of GCSE equivalent points. Alternatively, it could be that the improvement is more genuine and more broad-based.

Some light can be shed on this by whether there is any improvement in scores in compulsory core courses. We focus on maths and English[10], and run the same specification as table 5, instrumenting the fail status with the rating variable. The results are in Table 7, in the same format and specification as Table 5, all run on the narrow bandwidth. The table shows substantial and sustained effects of failing the Ofsted on achievement in the core subjects. This suggests that there is some genuine improvement in teaching taking place across the board. It also shows improvement in one year post visit

[10] Comparisons of Science over time are complicated by the many different ways that it can be taken: single science, double science, three separate subjects and so on.

An alternative question is whether the improvement is narrowly focussed on simply getting more students over the 5A*-C threshold. In fact, this appears not to be true, as the second row of the Table shows. The metrics are not the same so the magnitudes of the coefficients cannot be compared, but it is clear that the effect is statistically much weaker. Again, this suggests a broader – based improvement in teaching.

There are obviously other factors that might contribute to schools' improvement, but which cannot be addressed here. These include changes in the pattern of expenditure by schools, changes in teacher turnover leading to changes in average teacher effectiveness, and changes in school leadership (headteachers and/or governors). Some of these we hope to address in future work.

5.4 Differential impact of Ofsted failure on marginal students

One issue discussed in the NCLB literature is the extent to which there are differential effects on different groups of students. In particular, marginal students for whom a gain in performance is particularly valuable to the school might be expected to be targeted by the school.

Here we define three groups of students, based on their chances of achieving at least 5 C grades or better. We run a student-level probit on this indicator as a function of our prior ability measures, KS2 scores, and all the individual background variables discussed above. We then split the fitted probability index into thirds which we label 'lower ability', 'marginal' and 'higher ability'. The idea is that the performance of the marginal students might improve to pass the 5A* to C grade criterion with some improvement in the schools' effectiveness, and so are of particular interest.

We re-run the procedures in table 7 separately for the school means over these groups of students. The results are in Table 8. The top row reports on the mean capped GCSE score, we see a positive and significant effect for marginal students. The impact on higher ability students is larger still and more precisely determined. But the effect is lower and less significant for lower ability students. The results for English suggest that the biggest effects are on marginal students, with lower (though not significantly lower) effects on higher ability students, and much less well estimated effects on lower ability students. There are similar patterns in maths, though in general these coefficients are less precisely estimated.

These patterns are suggestive of a strategic response by the failed schools to how to allocate their increased effort. They may be responding to what is measured in the school performance tables and so are focussing their activity on students with good chances of getting at least 5 good passes.

5.5 Impact of Ofsted failure on changes in the school intake

Finally, we address a question about the long run viability of a school. One of the potential outcomes of an Ofsted failure is that local families turn against the school, parents choose not to send their children, and falling roles and consequent financial pressures force the school's closure. Within that, it could be that more affluent families look elsewhere for school places, so that while numbers hold up, the characteristics of the peer group change adversely.

To address this we look at the intake cohorts into secondary school, rather than the finishing cohorts taking their GCSEs. We analyse three metrics: the number of students entering the school; the mean prior ability score (KS2), and the fraction eligible for free school meals. The results in Table 9 use the narrow sample and again instrument the fail variable with the running variable. They show that enrolment does decline on average in the first three years by around 10 students (relative to a school-cohort mean of around 180, see Data Appendix 1), but that this decline stays steady and if anything starts to fade out. However, this decline appears to be evenly spread and there is no change in the average ability or disadvantage of the intake.

6. Conclusion

Every education system needs a method for identifying and dealing with underperforming schools. In England, the accountability system is composed of published performance tables and an independent inspectorate. Schools judged to be failing by the inspectors are then monitored and given a Notice to Improve. In this paper, we evaluate the impact on subsequent school performance of this judgement. We use a panel of 9 years of data and inspections of all secondary schools in England. Using data on the outcomes of all the sub-criteria used by the inspectors, we implement an RDD in the panel.

Our results suggest a quantitative and statistically significant effect. Relative to the year before the visit, school performance improves by around 10% of (student-level) SDs of GSCE exam performance. We show that this result is repeated across a number of different empirical approaches. The impact is significantly higher in the second year post visit than the first, and in most specifications remains level into the third year after the inspection. We also investigate the impact on individual subjects and show similar effects in the core compulsory subjects of maths and English. Finally, we analyse the impact on different groups of students separately. We show that the lowest ability group see little or no improvement in their scores after the inspection, compared to marginal and high ability students.

It may be that this represents a focus by the school on the marginal students, though the evidence of impact on maths and English suggests that the effect is not all gaming. It may be that failed schools are too closely monitored to get away with improving simply through manipulating course entries. Or it may be that low students have less capacity to benefit from the school changing ethos and putting more focus on attainment.

Comparing the longer-run of effects in the panel suggests a lower but still positive and significant effect of the fail judgement. This arises in part because we have shown that school scores display the Ashenfelter dip, declining two years before the inspection. Thus the fail judgement causes a substantial increase relative to (t-1) and a smaller but still positive one relative to further back in the school's history. An improvement over three to five years is very valuable as that affects the attainment records of a number of cohorts of students. Finally we note that because of the RDD approach, our results relate primarily to schools that were only just deemed to fail, and may have nothing to say about the prospects for severely failing schools. It may well be that a different policy response is needed for those schools.

References

Ahn, T. and Vigdor, J. (2009*) Does No Child Left Behind have teeth? Examining the impact of federal accountability sanctions in North Carolina*. Working Paper, October 2009.

Allen, R. and Burgess, S. (2011) Can school league tables help parents choose schools? *Fiscal Studies*, 32(2)245-261.

Allen, R. and Burgess, S. (2010) *The role and performance of Ofsted,* Memorandum submitted to the House of Commons Select Committee for Education, October 2010.

Angrist, J.D. and Lavy, V. (1999) Using Maimonides' Rule To Estimate The Effect Of Class Size On Scholastic Achievement, *The Quarterly Journal of Economics,* 114(2) 533-575.

Angrist, J.D. and Pischke, J-S. (2009) *Basically harmless econometrics*, Princeton, New Jersey: Princeton University Press.

Ashenfelter, O. (1978) Estimating the Effects of Training Programmes on Earnings. *The Review of Economics and Statistics* 60 (1) 47-57.

Bacolod, M., DiNardo, J. and Jacobson, M. (2009) *Beyond incentives: Do schools use accountability rewards productively?* NBER Working Paper No. 14775.

Burgess, S., Wilson, D. and Worth, J. (2010) *A natural experiment in school accountability: the impact of school performance information on pupil progress and sorting*, CMPO Working Paper 10/246.

Bush, T. and West-Burnham, J. (1994) *The Principles of Educational Management,* Harlow: Longman.

Chakrabarti, R. (2010) *Vouchers, public school response, and the role of incentives: Evidence from Florida*, Federal Reserve Bank of New York Report 306.

Coleman, J. (1998) The Place of External Inspection, in Middlewood, D. *Strategic Management in Schools and Colleges,* London: Paul Chapman.

Cooley, J. and Traczynski, J. (2011) *Spare the Rod? The Effect of No Child Left Behind on Failing Schools*, University of Wisconsin: mimeo.

Cullingford, C. and Daniels, S. (1999) *An inspector calls: Ofsted and its effect on school standards*, Kogan Page, London.

Dee, T. and Brian, J. (2009) *The impact of No Child Left Behind on student achievement*, NBER Working Paper No. 15531.

Faubert, V. (2009), School Evaluation: Current Practices in OECD Countries and a Literature Review, *OECD Education Working Papers*, No. 42, OECD Publishing.

Figlio, D.N. and Rouse, C.E. (2006) Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, 90(1-2) 239-255.

Heyneman, S. and Loxley, W. (1984) The effect of primary school on academic achievement across twenty nine high and low income countries, *The American Journal of Sociology*, 88(6) 1162-1194.

Hood, C., Dixon, R. and Wilson, D. (2010) Keeping Up the Standards? The Use of Targets and Rankings to Improve Performance, *School Leadership Today*, March 2010.

Hussain, I. (2011) *Subjective performance evaluation in the public sector: Evidence from school inspections*, Mimeo, University of Oxford.

Jacob,B. (2005) Accountability, incentives and behavior: Evidence from school reform in Chicago, *Journal of Public Economics*, 89(5-6) 761-796.

Joyce, B and Calhoun, E (1991) Review: The New Meaning of Educational Change, *School Effectiveness and School Improvement*, 2(4) 336-343.

Krieg, J. M. (2008) Are students left behind? The distributional effects of the No Child Left Behind Act, *Education Finance and Policy*, 3(2) 250-281.

Ladd, H.F. and Lauen, D.L. (2010) Status vs. growth: The distributional effects of school accountability policies, *Journal of Policy Analysis and Management*, 29(3) 424-450.

Learmonth, J. (2000) *Inspection: what's in it for schools?* London, Routledge Falmer

Marcantanio, R.J. and Cook, T.D. (1994) Convincing quasi-experiments: The interrupted time series and regression-discontinuity designs, in J.S. Wholey, H.P. Hatry and K.E. Newcomer (eds) *Handbook of Practical Program Evaluation,* San Francisco: Jossey-Bass.

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

Ofsted (2011a) *The framework for school inspection in England under section 5 of the Education Act 2005, from September 2009*, Report reference 090019 (September, 2011).

Ofsted (2011b) *Monitoring inspections of schools that are subject to special measures*, Report reference 090272 (September, 2011).

Ofsted (2011c) *Monitoring inspections of schools with a notice to improve,* Report reference 090277 (September, 2011).

Organisation for Economic Cooperation and Development (OECD) (1995) *Schools under Scrutiny*, Paris, OECD

Reback R. (2008) Teaching to the rating: School accountability and the distribution of student achievement, *Journal of Public Economics*, 92(5-6) 139-415.

Rockoff J.E. and Turner, L.J. (2008) *Short run impacts of accountability on school quality*, NBER Working Paper No. 14564.

Rosenthal, L. (2004) Do school inspections improve school quality? Ofsted inspections and school examination results in the UK, *Economics of Education Review*, 23 (2) 143-151.

Tables

Table 1: Ofsted inspections data

Year	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09
Number of school visits	475	558	452	929	1152	993	652
Number of sub-criteria used	19	33	33	55	41	58	65
Rating = Excellent	18	10	13	n/a	n/a	n/a	n/a
Outstanding/v good	117	97	109	98	165	179	151
Good	202	264	190	358	439	420	284
Satisfactory	113	130	107	350	451	321	178
Unsatisfactory	18	43	24	123	97	73	39
Poor	5	14	9	n/a	n/a	n/a	n/a
Very poor	2	0	0	n/a	n/a	n/a	n/a
Proportion failing (%)	5.26	10.22	7.30	13.24	8.42	7.35	5.98

Note: figures based on data used in year 11 (age 16) analysis

Table 2: Discontinuity tests for observable background variables

	***x* at t-1**			***x* at t+2 minus *x* at t-1**			***x* at t+4 minus *x* at t-1**		
	***fail*=1**	***fail*=0**	**p-value on diff**	***fail*=1**	***fail*=0**	**p-value on diff**	***fail*=1**	***fail*=0**	**p-value on diff**
FSM:									
bandwidth=all	17.81%	13.12%	0.00	-0.77%	-0.35%	0.14	-0.36%	-0.75%	0.37
bandwidth=broad	17.82%	15.73%	0.00	-0.77%	-0.32%	0.20	-0.31%	-0.31%	1.00
bandwidth=narrow	17.06%	20.23%	0.01	-0.35%	-0.54%	0.74	-0.05%	-0.46%	0.65
bandwidth=very narrow	19.23%	22.21%	0.29	-2.48%	-1.02%	0.23	-0.99%	-1.64%	0.77
KS2 English:									
bandwidth=all	-0.21	0.04	0.00	0.00	-0.01	0.42	-0.02	-0.01	0.62
bandwidth=broad	-0.21	-0.12	0.00	0.00	-0.02	0.16	-0.02	-0.01	0.52
bandwidth=narrow	-0.19	-0.25	0.01	-0.02	-0.01	0.62	-0.02	0.01	0.35
bandwidth=very narrow	-0.17	-0.24	0.23	-0.06	-0.03	0.41	-0.04	-0.03	0.90
% White British ethnicity:									
bandwidth=all	81.51%	80.61%	0.46	-0.04%	-1.03%	0.11	-1.30%	-1.31%	1.00
bandwidth=broad	81.40%	81.99%	0.64	-0.02%	-1.75%	0.02	-1.24%	-1.93%	0.62
bandwidth=narrow	81.90%	80.35%	0.45	0.55%	-2.53%	0.02	1.80%	-3.42%	0.03
bandwidth=very narrow	79.23%	78.51%	0.88	0.11%	-1.07%	0.78	3.12%	1.09%	0.82

Table 3: Before-After regressions for all schools

Dependent variable is the difference in school mean GCSE score (capped z-score)
Unit of observation is a school*visit; Metric is (pupil-level) SDs of GCSE score
Just the coefficient on "Ofsted failed" and its standard error reported

Difference in GCSE:	(t+1) – (t-1)	(t+2) – (t-1)	(t+3) – (t-1)	(t+4) – (t-1)
Basic[(b)]	0.069***	0.124***	0.126***	0.138***
	(0.009)	(0.011)	(0.014)	(0.019)
Adj-Rsqd	0.014	0.032	0.026	0.031
N	5086	4384	3353	2255
Differences[(c)]	0.066***	0.116***	0.124***	0.138***
	(0.008)	(0.010)	(0.013)	(0.017)
Adj-Rsqd	0.149	0.163	0.143	0.160
N	5083	4382	3351	2253
Levels[(d)]	0.053***	0.097***	0.096***	0.107***
	(0.008)	(0.010)	(0.013)	(0.017)
Adj-Rsqd	0.176	0.206	0.211	0.227
N	5083	4382	3351	2253
Common sample[(e)]	0.035***	0.071***	0.091***	0.107***
	(0.012)	(0.013)	(0.015)	(0.017)
Adj-Rsqd	0.206	0.222	0.214	0.227
N	2250	2253	2253	2253

Notes:
(a) Each cell of this table reports the results of a separate regression.
(b) Other variables included in 'Basic': Dummies for the year of the Ofsted visit.
(c) Other variables included in 'Differences': Dummies for the year of the Ofsted visit; differences of the same order as the column of the school mean: KS2 English, KS2 Maths, KS2 Science, fraction students eligible for Free School Meals, fraction students female, fraction students white British, fraction students with English as an additional language; IDACI score;
(d) Other variables included in 'Levels': Dummies for the year of the Ofsted visit; differences of the same order as the column of the school mean: KS2 English, KS2 Maths, KS2 Science, fraction students eligible for Free School Meals, fraction students female, fraction students white British, fraction students with English as an additional language; IDACI score; plus the level of all these same variables at (t-1).
(e) Common sample just uses the schools with data available for up to the fourth difference.
(f) Levels of significance indicated as * 0.10, ** 0.05, *** 0.01.

Table 4: Parametric Sharp RDD Analysis

Dependent variable is the difference in school mean GCSE score (capped z-score)
Unit of observation is a school*visit; Metric is (pupil-level) SDs of GCSE score
Just the coefficient on "Ofsted failed" and its standard error reported

Difference in GCSE:	(t+1) – (t-1)	(t+2) – (t-1)	(t+3) – (t-1)	(t+4) – (t-1)
All	0.039**	0.083***	0.078***	0.022
	(0.018)	(0.022)	(0.027)	(0.038)
Adj-Rsqd	0.180	0.210	0.220	0.235
N	5083	4382	3351	2253
Broad bandwidth	0.061***	0.125***	0.124***	0.055
	(0.022)	(0.027)	(0.034)	(0.046)
Adj-Rsqd	0.170	0.218	0.222	0.247
N	1851	1577	1098	670
Narrow bandwidth	0.019	0.135***	0.130**	0.023
	(0.036)	(0.041)	(0.055)	(0.079)
Adj-Rsqd	0.119	0.222	0.156	0.267
N	501	437	314	193
Very narrow bandwidth	0.060	0.156*	-0.083	-0.468**
	(0.092)	(0.091)	(0.134)	(0.177)
Adj-Rsqd	0.014	0.279	0.283	0.543
N	103	91	72	40

Notes:
(a) Each cell of this table reports the results of a separate regression.
(b) Other variables included in all regressions: Fourth order polynomial in the rating variable, and fourth order polynomial in the rating variable interacted with Fail status; dummies for the year of the Ofsted visit; differences of the same order as the column of the school mean: KS2 English, KS2 Maths, KS2 Science, fraction students eligible for Free School Meals, fraction students female, fraction students white British, fraction students with English as an additional language; IDACI score; plus the level of all these same variables at (t-1).
(c) Bandwidths defined on the rating variable: All: all school*visits;
(d) Broad: The broad sample is defined by rating variable greater than or equal to -5, and this gives a maximum of 1899 school*visits, of which 445 (23.4%) were fails and 1454 were passes.
(e) Narrow: The narrow sample is defined by rating variable greater than or equal to -2 and less than or equal to 7, and this gives a maximum of 521 school*visits, of which 252 (48.4%) were fails and 269 were passes.
(f) Very narrow: The very narrow sample is defined by rating variable greater than or equal to -0.5 and less than or equal to 1.5, and this gives a maximum of 108 school*visits, of which 55 (50.9%) were fails and 53 were passes.
(g) Levels of significance indicated as * 0.10, ** 0.05, *** 0.01.

Table 5: Fuzzy RDD IV regression analysis

Dependent variable is the difference in school mean GCSE score (capped z-score)
Unit of observation is a school*visit; Metric is (pupil-level) SDs of GCSE score
Just the coefficient on "Ofsted failed" and its standard error reported
The fail status variable is instrumented by the rating variable.

Difference in GCSE:	(t+1) – (t-1)	(t+2) – (t-1)	(t+3) – (t-1)	(t+4) – (t-1)
Basic[(b)], full sample	0.079***	0.151***	0.176***	0.189***
	(0.011)	(0.014)	(0.018)	(0.023)
Adj-Rsqd	0.013	0.030	0.023	0.028
N	5086	4384	3353	2255
Levels, full sample	0.046***	0.102***	0.116***	0.126***
	(0.010)	(0.013)	(0.017)	(0.021)
Adj-Rsqd	0.176	0.206	0.211	0.226
N	5083	4382	3351	2253
Broad bandwidth	0.026**	0.069***	0.066***	0.098***
	(0.013)	(0.016)	(0.022)	(0.030)
Adj-Rsqd	0.166	0.219	0.220	0.242
N	1851	1577	1098	670
Narrow bandwidth	0.040*	0.100***	0.085**	0.066
	(0.024)	(0.027)	(0.041)	(0.055)
Adj-Rsqd	0.124	0.231	0.171	0.285
N	501	437	314	193
Very narrow bandwidth	0.067	0.085	0.008	0.284
	(0.082)	(0.080)	(0.117)	(0.283)
Adj-Rsqd	0.066	0.320	0.296	0.262
N	103	91	72	40

Notes:
(a) Each cell of this table reports the results of a separate regression.
(b) Other variables included in 'Basic': Dummies for the year of the Ofsted visit.
(c) Other variables included in all other regressions: dummies for the year of the Ofsted visit; differences of the same order as the column of the school mean: KS2 English, KS2 Maths, KS2 Science, fraction students eligible for Free School Meals, fraction students female, fraction students white British, fraction students with English as an additional language; IDACI score; plus the level of all these same variables at (t-1).
(d) Bandwidths defined on the rating variable: All: all school*visits;
(e) Broad: The broad sample is defined by rating variable greater than or equal to -5, and this gives a maximum of 1899 school*visits, of which 445 (23.4%) were fails and 1454 were passes.
(f) Narrow: The narrow sample is defined by rating variable greater than or equal to -2 and less than or equal to 7, and this gives a maximum of 521 school*visits, of which 252 (48.4%) were fails and 269 were passes.
(g) Very narrow: The very narrow sample is defined by rating variable greater than or equal to -0.5 and less than or equal to 1.5, and this gives a maximum of 108 school*visits, of which 55 (50.9%) were fails and 53 were passes.
(h) Levels of significance indicated as * 0.10, ** 0.05, *** 0.01.

Table 6: Difference-in-difference panel analysis

Dependent variable is the difference in school mean GCSE score (capped z-score)
Unit of observation is a school*visit; Metric is (pupil-level) SDs of GCSE score
Just the coefficient on "Ofsted failed" times D(t+#) and its standard error reported

Fail*Dummy for year:	Year = visit + 1	Year = visit + 2	Year = visit + 3	Year = visit + 4
Full sample	-0.004	0.053***	0.061***	0.075***
	(0.007)	(0.008)	(0.009)	(0.011)
Adj-Rsqd	0.046	0.047	0.047	0.047
N*T	45319	45319	45319	45319
N	5195	5195	5195	5195
Broad bandwidth	-0.001	0.047***	0.042***	0.064***
	(0.008)	(0.009)	(0.010)	(0.013)
Adj-Rsqd	0.059	0.061	0.060	0.061
N*T	16332	16332	16319	16319
N	1899	1899	1899	1899
Narrow bandwidth	-0.009	0.058***	0.052***	0.011
	(0.013)	(0.014)	(0.017)	(0.022)
Adj-Rsqd	0.091	0.095	0.093	0.091
N*T	4409	4409	4409	4409
N	515	515	515	515
Very narrow bandwidth	0.007	0.108***	0.057	-0.061
	(0.03)	(0.032)	(0.036)	(0.048)
Adj-Rsqd	0.120	0.133	0.123	0.123
N*T	905	905	905	905
N	108	108	108	108

Notes:
(a) Each cell of this table reports the results of a separate regression.
(b) Other variables included in all other regressions: year dummies; dummy for the year of the visit; contemporaneous values of the school mean of: KS2 English, KS2 Maths, KS2 Science, fraction students eligible for Free School Meals, fraction students female, fraction students white British, fraction students with English as an additional language; IDACI score.
(c) Bandwidths defined on the rating variable: All: all school*visits;
(d) Broad: The broad sample is defined by rating variable greater than or equal to -5, and this gives a maximum of 1899 school*visits, of which 445 (23.4%) were fails and 1454 were passes.
(e) Narrow: The narrow sample is defined by rating variable greater than or equal to -2 and less than or equal to 7, and this gives a maximum of 521 school*visits, of which 252 (48.4%) were fails and 269 were passes.
(f) Very narrow: The very narrow sample is defined by rating variable greater than or equal to -0.5 and less than or equal to 1.5, and this gives a maximum of 108 school*visits, of which 55 (50.9%) were fails and 53 were passes.
(g) Levels of significance indicated as * 0.10, ** 0.05, *** 0.01.

Table 7: Fuzzy RDD IV regression analysis, different outcome variables

Dependent variable is the difference in school mean of various outcome measures
Unit of observation is a school*visit; Metric is (pupil-level) SDs for row 1, fraction for row 2, and mean point score (from 1 to 8) for rows 3 and 4.
Just the coefficient on "Ofsted failed" and its standard error reported.
The fail status variable is instrumented by the rating variable.

Difference in GCSE:	(t+1) – (t-1)	(t+2) – (t-1)	(t+3) – (t-1)	(t+4) – (t-1)
Capped mean GCSE score	0.040*	0.100***	0.085**	0.066
	(0.024)	(0.027)	(0.041)	(0.055)
Adj-Rsqd	0.124	0.231	0.171	0.285
N	501	437	314	193
Fraction of pupils achieving at least 5 A*-C grades	0.018	0.049***	0.018	0.019
	(0.014)	(0.017)	(0.024)	(0.031)
Adj-Rsqd	0.182	0.255	0.215	0.329
N	501	437	314	193
Mean English GCSE score	0.117**	0.194***	0.179**	0.142
	(0.050)	(0.054)	(0.077)	(0.091)
Adj-Rsqd	0.236	0.256	0.225	0.356
N	501	437	314	193
Mean Maths GCSE score	0.096**	0.161***	0.099	0.051
	(0.046)	(0.054)	(0.069)	(0.091)
Adj-Rsqd	0.275	0.242	0.249	0.226
N	501	437	314	193

Notes:
(a) Each cell of this table reports the results of a separate regression.
(b) Other variables included in all other regressions: dummies for the year of the Ofsted visit; differences of the same order as the column of the school mean: KS2 English, KS2 Maths, KS2 Science, fraction students eligible for Free School Meals, fraction students female, fraction students white British, fraction students with English as an additional language; IDACI score; plus the level of all these same variables at (t-1).
(c) Bandwidths defined on the rating variable: This is the narrow sample, defined by rating variable greater than or equal to -2 and less than or equal to 7, and this gives a maximum of 521 school*visits, of which 252 (48.4%) were fails and 269 were passes.
(d) Levels of significance indicated as * 0.10, ** 0.05, *** 0.01.

Table 8: Analysis of marginal pupils versus others

Dependent variable is the difference in school mean of various outcome measures
Unit of observation is a school*visit; Metric is (pupil-level) SDs for row 1, fraction for row 2, and mean point score (from 1 to 8) for rows 3 and 4.
Just the coefficient on "Ofsted failed" and its standard error reported.
The fail status variable is instrumented by the rating variable.

Difference in GCSE:	(t+1) – (t-1)	(t+2) – (t-1)	(t+3) – (t-1)	(t+4) – (t-1)
Capped mean GCSE score				
Lower ability students	0.022	0.067**	0.075	0.037
	(0.028)	(0.033)	(0.049)	(0.067)
Marginal students	0.048	0.097***	0.123**	0.134*
	(0.031)	(0.035)	(0.052)	(0.070)
Higher ability students	0.101***	0.153***	0.159***	0.154**
	(0.037)	(0.043)	(0.055)	(0.070)
5+ A*-C GCSE score				
Lower ability students	-0.009	0.026	-0.001	-0.005
	(0.018)	(0.022)	(0.033)	(0.044)
Marginal students	0.056***	0.066***	0.062*	0.081*
	(0.021)	(0.024)	(0.033)	(0.042)
Higher ability students	0.047***	0.068***	0.057***	0.044*
	(0.017)	(0.018)	(0.021)	(0.025)
English GCSE				
Lower ability students	0.071	0.140**	0.128	0.077
	(0.062)	(0.067)	(0.101)	(0.126)
Marginal students	0.222***	0.239***	0.279***	0.300***
	(0.056)	(0.061)	(0.090)	(0.102)
Higher ability students	0.184***	0.267***	0.172*	0.124
	(0.064)	(0.074)	(0.096)	(0.121)
Maths GCSE				
Lower ability students	0.064	0.114*	0.097	0.041
	(0.052)	(0.063)	(0.089)	(0.115)
Marginal students	0.097	0.176**	0.126	0.144
	(0.060)	(0.070)	(0.084)	(0.107)
Higher ability students	0.097	0.179**	0.162*	0.154
	(0.061)	(0.076)	(0.094)	(0.123)
N	495	431	311	193

Notes:
(a) Each cell of this table reports the results of a separate regression.
(b) Other variables included in all other regressions: dummies for the year of the Ofsted visit; differences of the same order as the column of the school mean: KS2 English, KS2 Maths, KS2 Science, fraction students eligible for Free School Meals, fraction students female, fraction students white British, fraction students with English as an additional language; IDACI score; plus the level of all these same variables at (t-1).
(c) Bandwidths defined on the rating variable: This is the narrow sample, defined by rating variable greater than or equal to -2 and less than or equal to 7, and this gives a maximum of 521 school*visits, of which 252 (48.4%) were fails and 269 were passes.
(d) Levels of significance indicated as * 0.10, ** 0.05, *** 0.01.

Table 9: Analysis of change in pupil intake

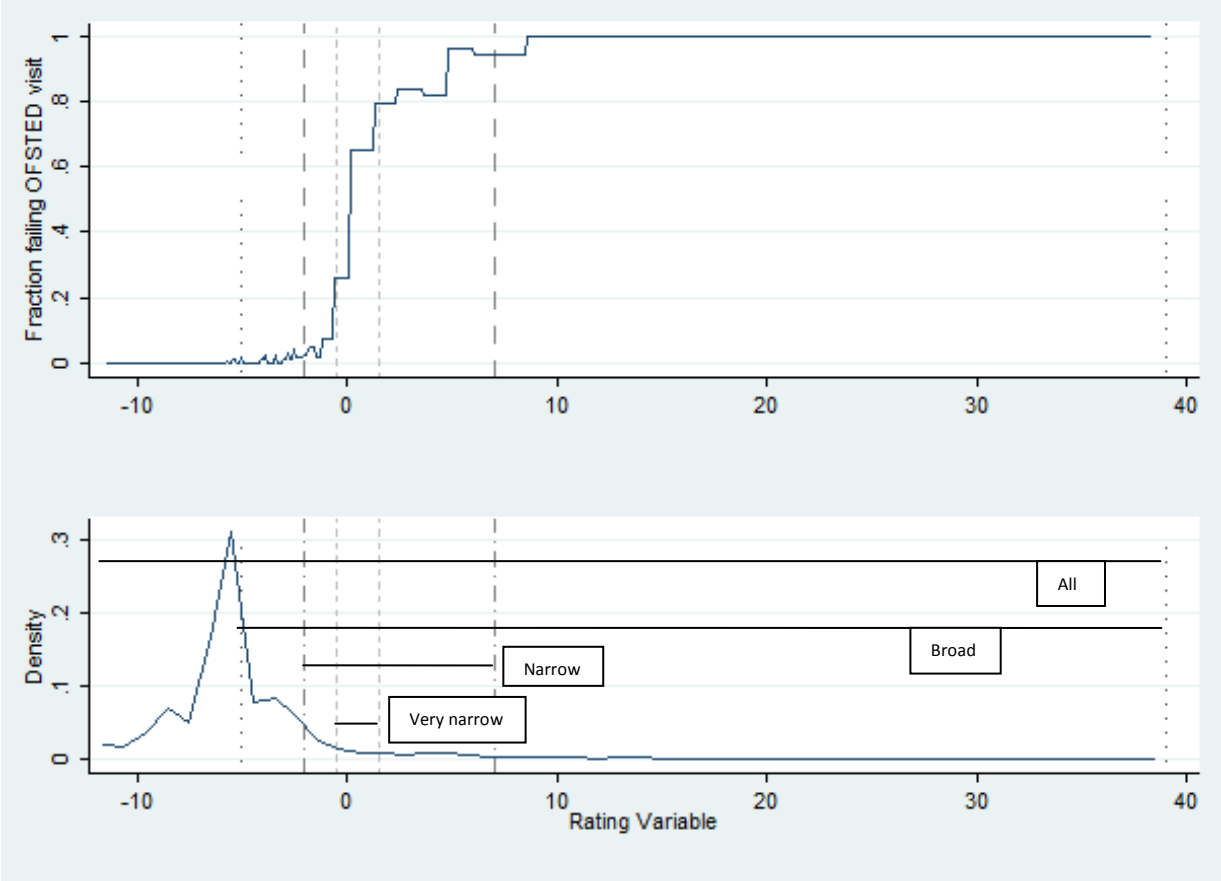
Dependent variable is the difference in school mean of various outcome measures
Unit of observation is a school*visit; Metric is number of pupils in row 1, the mean points score in row 2, and the proportion in row 3.
Just the coefficient on "Ofsted failed" and its standard error reported.
The fail status variable is instrumented by the rating variable.

Difference in year 7 metric:	(t+1) – (t-1)	(t+2) – (t-1)	(t+3) – (t-1)	(t+4) – (t-1)
Number of pupils in cohort	-11.509***	-10.412***	-10.622*	-7.029
	(3.837)	(4.475)	(5.704)	(7.716)
Adj-Rsqd	0.025	0.014	0.005	0.015
N	494	461	394	285
Mean KS2 score	0.017	0.004	0.029	0.032
	(0.019)	(0.022)	(0.026)	(0.039)
Adj-Rsqd	0.017	0.004	0.000	0.015
N	494	461	394	285
FSM proportion	-0.001	-0.000	-0.009	-0.003
	(0.011)	(0.013)	(0.014)	(0.019)
Adj-Rsqd	0.003	0.003	0.015	0.011
N	494	461	394	285

Notes:
(a) Each cell of this table reports the results of a separate regression.
(b) Other variables included in all other regressions: dummies for the year of the Ofsted visit.
(c) Bandwidths defined on the rating variable: This is the narrow sample, defined by rating variable greater than or equal to -2 and less than or equal to 7, and this gives a maximum of 521 school*visits, of which 252 (48.4%) were fails and 269 were passes.
(d) Levels of significance indicated as * 0.10, ** 0.05, *** 0.01.

Figures

Figure 1: RDD assignment variable



Bandwidth:	N	Passes (% of N)	Fails (% of N)
All	5196	4748 (91.4)	448 (8.6)
Broad	1899	1454 (76.6)	445 (23.4)
Narrow	521	269 (51.6)	252 (48.4)
Very Narrow	108	53 (49.1)	55 (50.9)

Figure 2: Mean scores before and after Ofsted inspection, by pass/fail

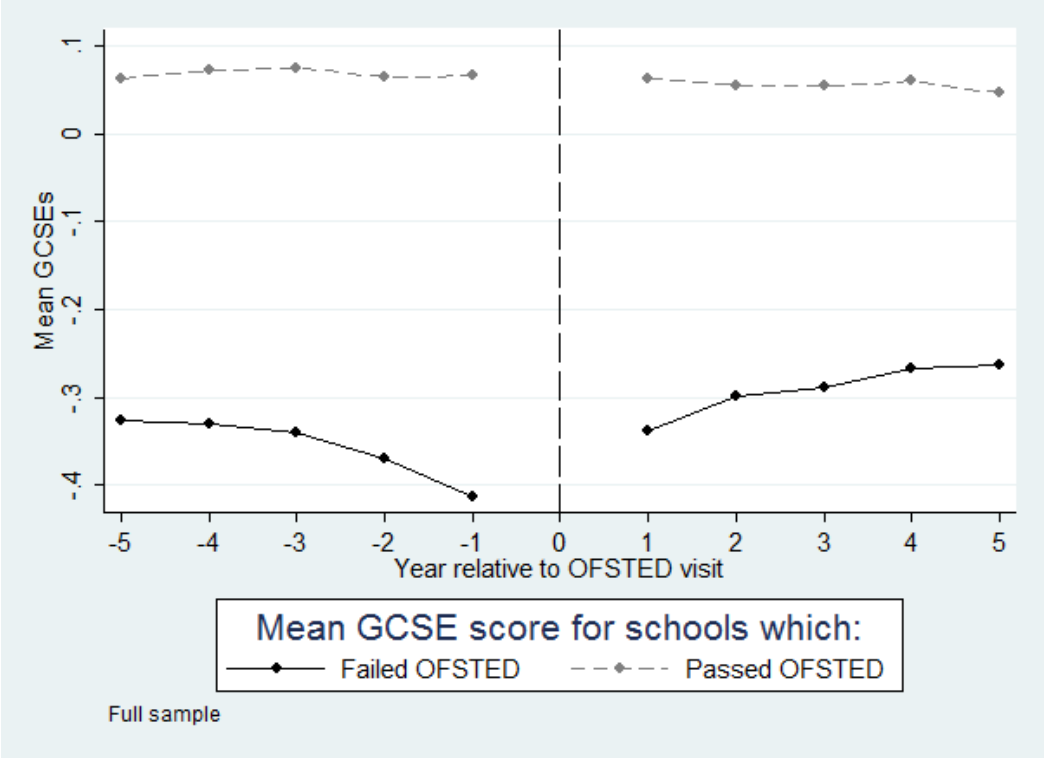


Figure 3: Mean scores before and after Ofsted inspection, by pass/fail and year of inspection

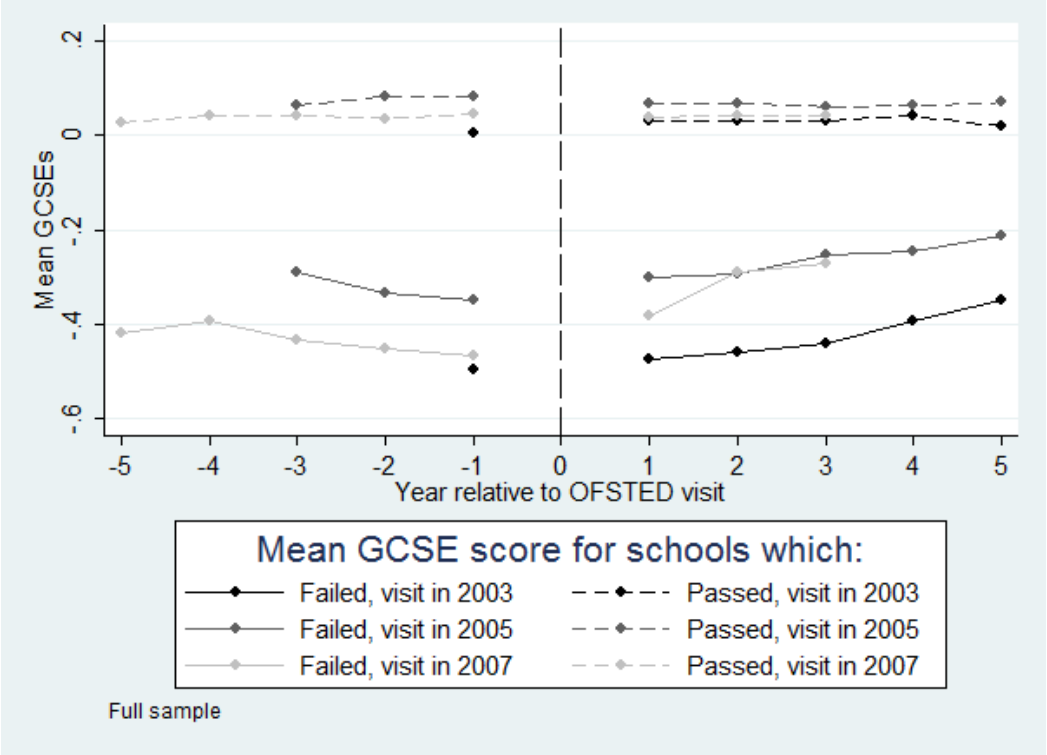


Figure 4: Running variable – Linear Fit and Distribution

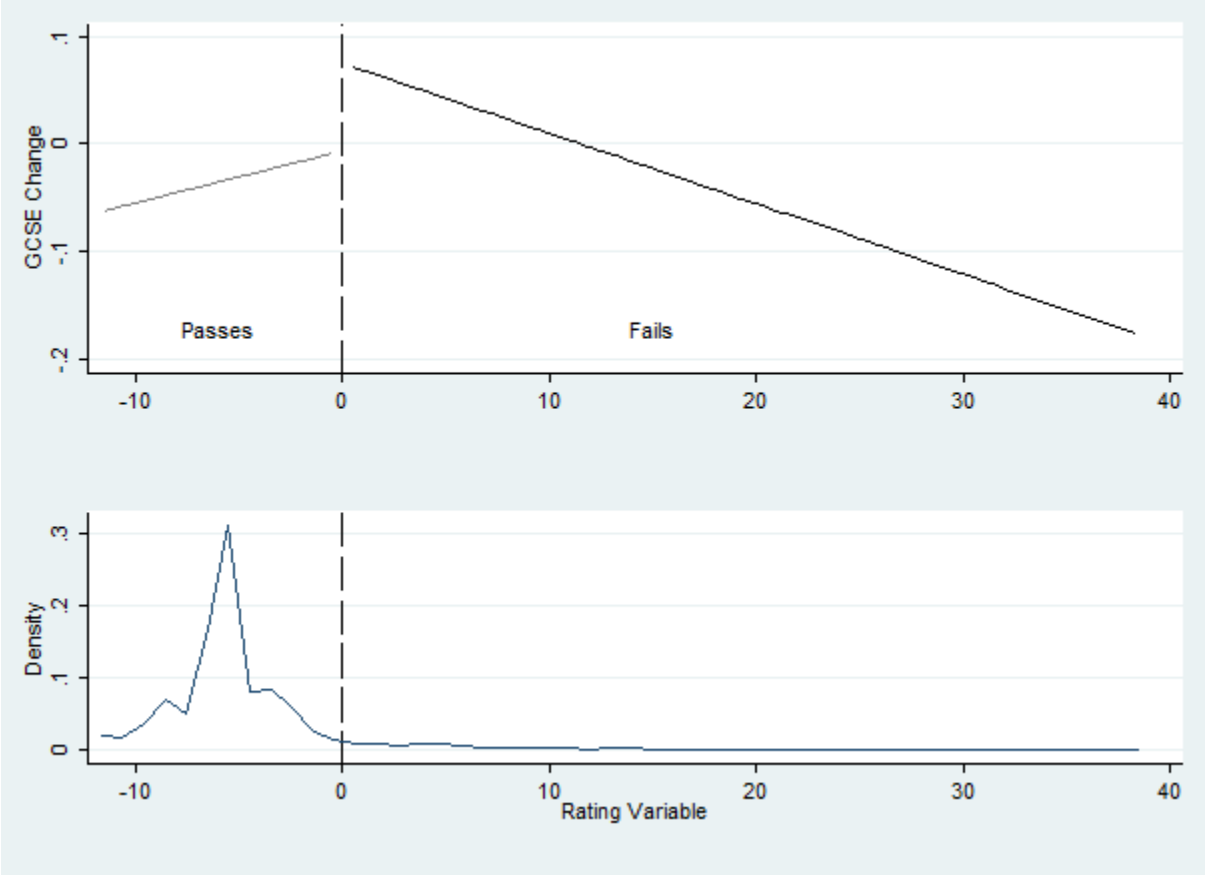
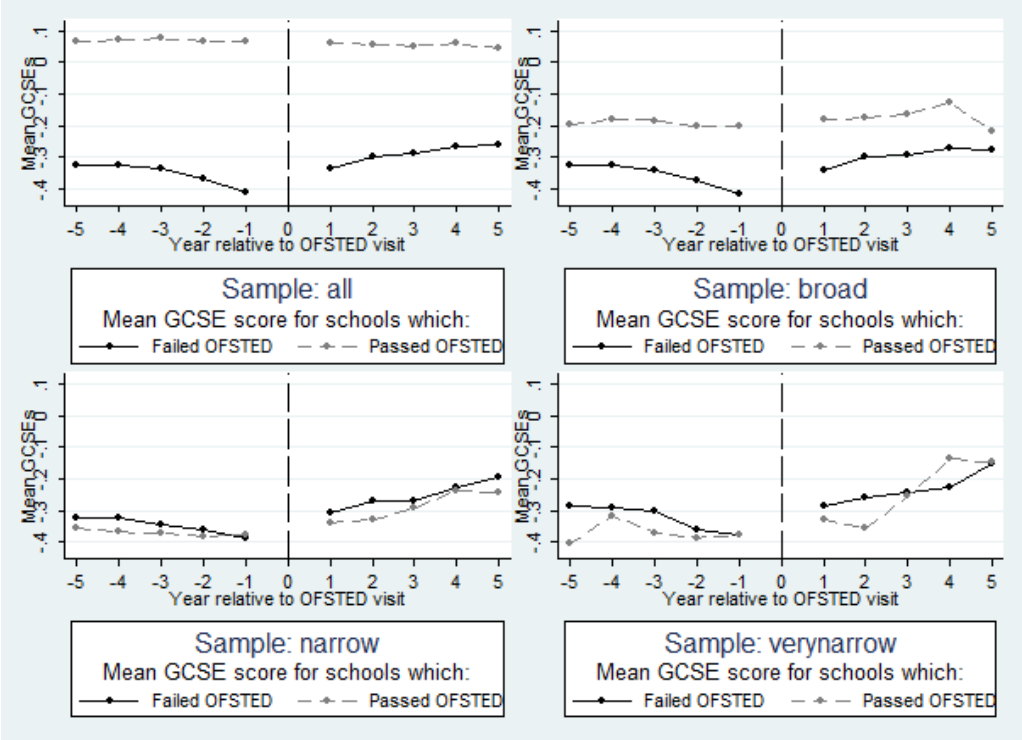


Figure 5: Mean scores before and after Ofsted inspection, by pass/fail and bandwidth



Data Appendix

Data Appendix Table 1: Summary statistics for pupil background and prior attainment and outcomes

	2002	2003	2004	2005	2006	2007	2008	2009	2010
Year 11 cohorts:									
GCSE capped (best 8) z-score	0.02	0.04	0.03	0.03	0.03	0.03	0.02	0.03	0.04
GCSE 5+ A*-C	51.48%	52.12%	53.05%	55.65%	58.02%	60.84%	65.18%	70.23%	77.45%
GCSE English score	4.74	4.69	4.46	4.54	4.58	4.72	4.77	4.75	4.97
GCSE maths score	4.24	4.10	4.17	4.29	4.35	4.52	4.60	4.61	4.82
Number of pupils	174.40	182.09	186.56	185.65	188.77	191.04	190.70	186.76	186.28
KS2 English score (age 11)	0.02	0.03	0.03	0.02	0.02	0.03	0.00	0.03	0.04
KS2 maths score (age 11)	-0.01	0.03	0.03	0.03	0.02	0.03	0.00	0.03	0.04
KS2 science score (age 11)	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.03	0.04
Proportion free school meals	14.22%	13.85%	13.96%	13.90%	13.30%	12.86%	12.57%	12.83%	13.08%
Mean deprivation IDACI score	0.21	0.21	0.22	0.22	0.22	0.22	0.23	0.23	0.22
Proportion female	49.45%	49.46%	49.69%	49.62%	49.37%	49.53%	49.45%	49.47%	49.57%
Proportion English Additional Language	9.00%	9.06%	8.93%	9.22%	9.60%	9.81%	10.21%	11.10%	11.38%
Proportion ethnicity white British	82.25%	79.25%	80.59%	80.66%	80.87%	80.80%	80.31%	78.73%	78.41%
Year 7 cohorts:									
Number of pupils	188.39	188.40	186.17	179.69	179.25	175.42	170.06	174.72	175.16
Mean KS2 score (age 11)	0.03	0.04	0.03	0.03	0.02	0.00	0.01	0.01	0.03
Proportion in top 25% at KS2	25.98%	26.00%	25.79%	25.54%	25.36%	25.07%	25.13%	25.50%	25.75%
Proportion at bottom 25% at KS2	24.06%	24.01%	24.38%	24.54%	24.70%	25.10%	24.99%	24.70%	24.24%
Proportion free school meals	17.27%	16.92%	16.94%	16.91%	17.18%	16.88%	16.80%	17.09%	17.78%
Proportion ethnicity white British	82.43%	80.93%	81.84%	80.50%	79.38%	78.50%	78.33%	77.51%	76.53%

Data Appendix Table 2: Statistics for first stage regressions for Table 5

Model	Difference order	N	F-stat	p-value	t-stat on instrument
Basics	1	5086	1286.4	0.0	94.3
	2	4384	1275.1	0.0	86.9
	3	3353	1157.2	0.0	75.6
	4	2255	1063.3	0.0	64.9
Levels, all	1	5083	396.7	0.0	91.7
	2	4382	353.4	0.0	84.4
	3	3351	278.9	0.0	73.5
	4	2253	216.0	0.0	62.9
Broad	1	1851	123.6	0.0	48.9
	2	1577	111.0	0.0	44.8
	3	1098	79.4	0.0	36.6
	4	670	54.1	0.0	28.2
Narrow	1	501	24.1	0.0	21.1
	2	437	23.4	0.0	20.2
	3	314	14.0	0.0	15.0
	4	193	10.6	0.0	12.1
V. Narrow	1	103	2.5	0.0	5.1
	2	91	2.0	0.0	5.1
	3	72	1.3	0.2	3.9
	4	40	0.8	0.7	2.0

See notes to table 5 for details.