# Subjective Performance Evaluation in the Public Sector: Evidence From School Inspections

Iftikhar Hussain*
University of Sussex and Centre for Economic Performance, LSE

Draft, September 2011
Comments welcome

**Abstract**

Performance measurement in the public sector is largely based on 'hard' metrics, which have the benefit of being transparent, but may be subject to gaming behavior. Subjective performance evaluation offers the potential advantage of 'measuring what matters', but is open to manipulation by the bureaucrats charged with oversight. This paper investigates a novel school inspection system where independent inspectors visit schools at very short notice, write and disclose school quality reports and sanction those schools rated 'Fail'.

After demonstrating that inspection ratings are valid in the sense of being conditionally correlated with independent underlying school quality measures, the study evaluates the causal effect of a fail inspection on subsequent student performance. The evidence shows that a fail inspection leads to test score gains. The largest gains accrue to students with lower prior ability; this result cannot be accounted for by 'ceiling effects' for high ability students. The evidence also shows that at least some of these gains persist in the medium term. Furthermore, and in contrast with much evidence from test-based accountability regimes, the study finds no evidence to suggest that fail schools are able to inflate test score performance by gaming the system, suggesting that oversight by the inspectors may limit such strategic behavior.

*For helpful discussions and comments I would like to thank Orazio Attanasio, Oriana Bandiera, Steve Bond, Martin Browning, Ian Crawford, Avinash Dixit, Sergio Firpo, Meg Meyer, Caroline Hoxby, Andrea Ichino, Ian Jewitt, Kevin Lang, Valentino Larcinese, Clare Leaver, Steve Machin and Imran Rasul as well as seminar participants at Oxford University, Sussex University and the Second Workshop On the Economics of Education, Barcelona. All remaining errors are my own. Email: i.hussain@lse.ac.uk

# 1 Introduction

In an effort to make public organizations more efficient, governments around the world make use of 'hard' performance targets to evaluate the quality of service delivery. Examples include student test scores for the schooling sectors in the US, England and Chile (see the survey by Figlio and Loeb, 2011) and hospital waiting times in the English public health care system (Besley, Bevan and Buchardi, 2008; Propper et al, 2008). Accountability based on hard or objective performance measures has the benefit of being transparent but a potential drawback is that such schemes may lead to gaming behavior in a setting where incentives focus on just one dimension of a mutifaceted outcome.[1]

Subjective performance evaluation, on the other hand, holds out the promise of 'measuring what matters' if the evaluator can combine both hard and soft information to measure the outcome. However, a system where the evaluator is allowed to exercise his or her own judgement, rather than following a formal decision rule, raises a new set of concerns. For example, results from the theoretical literature emphasize 'influence activities' and 'favoritism' (Milgrom and Roberts 1988; Prendergast and Topel 1996) which make the subjective measure 'corruptible' (Dixit, 2002). Empirical evidence on the effectiveness of subjective evaluation remains thin.[2]

This paper seeks to fill this gap by empirically evaluating a subjective performance evaluation regime for schools. The setting is the English public (state) schooling system, where independent inspectors visit, monitor and report on the quality of schools. Furthermore, schools rated 'Fail' may be subject to sanctions, such as more frequent and intensive inspections. As explained in detail below, inspectors combine hard metrics, such as test scores, with softer ones, such as observations of classroom teaching, in order to arrive at a judgement of school quality. There is almost no empirical evidence on whether such a system for the schooling sector works in practice.

I provide evidence on the effectiveness of this regime along the following two dimensions. First, do inspector ratings provide any extra information on school quality, over and above that already available in the public sphere? This 'validity test' is implemented as follows: I ask whether inspection ratings are correlated with underlying school quality measures - constructed from teenage student survey reports of teacher practices as well as parental satisfaction - conditional on standard observable school characteristics such as test score rankings and the proportion of students

[1] See Holmstrom and Milgrom (1991) for a formal statement of the multitasking model. Dixit (2002) discusses incentive issues in the public sector. For empirical examples of dysfunctional behaviour in the education sector see the references below.

[2] As noted by Prendergast (1999, p.33), the economics literature has largely focused on "workers with easily observed output [who are] a small fraction of the population." See also the survey by Lazear and Oyer (forthcoming). In many settings good objective measures may not be immediately available. For example, in their analysis of active labor market programs, Heckman et al (2011, p.10) note that: "..the short-term measures that continue to be used in...performance standards systems are only weakly related to the true long-run impacts of the program... Researchers and policymakers have yet to identify performance measures that will promote key, long-term program objectives while simultaneously generating information for ongoing program management." Whether complementing objective performance evaluation with subjective assessment is an effective strategy in such settings remains an open question.

eligible for a free lunch.[3]

Second, I examine whether a fail inspection rating leads to subsequent gains in student test scores. Identifying the effect of a fail rating on test scores is plagued by the kind of mean reversion problems encountered in evaluations of active labour market programs (see Ashenfelter, 1978; Heckman, Lalonde and Smith, 1999). This is because assignment to treatment, fail, is at least partly based on past realizations of the outcome variable, test scores. The threat to identification is that poor performance prior to inspection is simply due to bad luck and that test scores at fail schools would have risen even in the absence of a fail inspection. Any credible strategy must overcome such concerns. Figure 1 illustrates the problem in the current setting. Between 2000 and 2005, test score performance on the age-11 Key Stage 2 mathematics test declined at schools failed in 2006 relative to schools rated satisfactory or better in the same inspection year. There is a dramatic pickup in performance at failed schools both in the year of inspection and subsequently.[4] The key question is: to what extent can this rise in performance can be attributed to the fail inspection?

< Fig. 1 here >

This study exploits a design feature of the English testing system to assess the causal effect of a fail inspection. As explained in detail below, tests for Year 6 (age 11) students in England are administered in the second week of May in each year. These tests are marked externally, and results are released to schools and parents in mid-July. The short window between May and July allows me to address the issue of mean reversion: schools failed in June are failed *after* the test in May but *before* the inspectors know the outcome of the tests.[5] By comparing schools failed early in the academic year - September, say - with schools failed in June I can isolate mean reversion from the effect of the fail inspection.[6]

An additional issue examined in this study is whether any estimated positive effect of a fail inspection on test scores can be explained by strategic or dysfunctional responses by teachers. A growing literature has by now established the empirical importance of such phenomena in

[3] One way to motivate this test is to ask whether parents engaged in searching for a school for their child should place any weight on inspection ratings. If the inspection ratings pass the validity test, then prospective parents may have some confidence that the ratings help forecast the views of the school's current stock of students and parents, even after conditioning on publicly available school indicators.

[4] Note that '2000' here refers to the academic year 1999/2000, which runs from September 1999 to July 2000; and so on for the other years. Inspections are undertaken throughout the academic year. The Key Stage 2 test is administered in May of each year. Thus, test score information is typically not available for the year in which the inspection takes place.

[5] So that the May test outcome for these schools is not affected by the subsequent fail, but neither do inspectors select them for failure on the basis of this outcome.

[6] The descriptive analysis demonstrates that there is little difference in observable characteristics between schools failed in June (the control group) and schools failed in the early part of the academic year (the treatment group). This, combined with the fact that timing is determined by a mechanical rule, suggests that there are unlikely to be unobservable differences between control and treatment schools. The claim then is that when comparing early and late fail schools within a year, treatment (early inspection) is as good as randomly assigned.

the context of schools. The overall message from this body of evidence is that when a school's incentives are closely tied to test scores teachers will often adopt strategies which artificially boost the school's measured test score performance.[7] I test to what extent such behavior can be detected in the current context.

The findings of this study are as follows. On whether inspection ratings are correlated with student (and parent) reports of school practices, the evidence shows that inspection ratings are strongly associated with these survey measures of school quality. For example, the association between inspection ratings and student survey reports of teacher practices is economically meaningful and statistically significant, even after conditioning on the school's test rank, proportion of students eligible for a free lunch and other school and student characteristics. This result implies that students enrolled in schools with better inspection ratings experience an environment where, according to student self-reports, teachers practices are superior. Similar findings hold for other measures of school quality constructed from the student and parent survey responses, including parent satisfaction. These results suggest that inspector ratings are informative about 'insider' views of the school, conditional on traditional measures of school attributes.

Turning to the effect of a fail rating on test scores, results using student-level data from a panel of all the schools failed in England between 2006 and 2009 show that students at early failed schools (the 'treatment' group) gain 0.12 of a standard deviation on national standardized mathematics test scores relative to students enrolled in late fail schools (the 'control' group). The treatment effect for English is a gain of 0.08 of a standard deviation. These results are robust to different methods of estimation: simple comparisons of post-treatment outcomes for the control and treatment groups as well as difference-in-differences models yield very similar results.

I do not find evidence of gaming behavior. Teachers do not exclude low ability students from tests nor do they appear to target students on the margin of attaining the official test target (attainment 'Level 4' on the Key Stage 2 age-11 test) to the detrimant of students far from this standard. In addition, test gains appear to be long lasting, at least for some students, suggesting that gains are not driven purely by 'teaching to the test'.

Given the prior evidence on strategic behavior these results are revealing. In the English setting the stakes - certainly for the principal - are potentially very high.[8] The fact that I find no evidence of the sort of strategic behavior encountered in many other similar contexts suggests that inspectors may have a mitigating effect on such dysfunctional responses.

The overall effect masks substantial heterogeneity in the treatment effect. The largest gains are for students scoring low on the prior (age seven) Key Stage 1 test. Within this subgroup, quantile regression analysis reveals that higher achieving students gain the most. For example,

[7] Strategies include excluding low ability students from the test taking pool and targeting students on the margin of passing proficiency thresholds, see, for example, Burgess et al 2005, Cullen and Reback 2006, Figlio 2006, Figlio and Getzler 2006, Jacob 2005, Neal and Schanzenbach 2010 and Reback 2008.

[8] As demonstrated in Hussain (2009), when schools fail an inspection there is a significant rise in the probability that the school principal exits the teaching labour market. Thus the incentives to generate improvements in school performance are strong. (Note that the school's governing board has the power to fire the principal if the school fails an inspection.)

for mathematics, students in the bottom quartile of the prior ability distribution gain between 0.1 and 0.3 of a standard deviation, with the estimated effect rising steadily from 0.1 for the lowest quantiles to 0.3 for the highest quantiles. These results are consistent with the view that children of poorer parents gain the most from a fail inspection. One reason for this may be that poorer parents have limited capacity to assess quality of instruction provided by the school and hence their children may receive less attention from teachers. Inspectors may then play an important role in mitigating the effects of such information asymmetries.

The main contribution of this study is to offer an empirical evaluation of the effectiveness of school inspections. A number of countries are experimenting with school inspections and there are some indications that interest in adopting such schemes is growing.[9]

In addition to the literature cited earlier on subjective performance evaluation, this study is also related to a small empirical literature investigating bureaucratic behavior. For example, Heckman, Smith and Taber (1996) show that in the context of a job training programme, case workers, who are allowed to use their own judgement in allocating training, tend to indulge their own preferences by enrolling the least advantaged and least employable applicants into the programme. The question the present study seeks to address is whether the bureaucrats charged with inspecting schools - who may be also be subject to biases and prejudices, as in the Heckman et al (1996) study - are effective along the dimensions highlighted above.

Finally, this paper also links to studies examining the role of monitoring in raising public service quality, especially for the poor. A number of examples exist from developing country settings, including Olken (2007) who finds positive effects of top-down monitoring and Bjorkman and Svensson (2009) who report beneficial effects of community based monitoring.

The remainder of this paper is laid out as follows. Section 2 describes the context for this study and the relevant theoretical background. Section 3 reports findings on the validity of inspection ratings. Section 4 lays out the empirical strategy used in order to evaluate the effect of a fail inspection on student test scores. This section also describes the empirical methods employed to test for strategic behavior by teachers in response to the fail rating. Section 5 reports the results and section 6 concludes.

# 2 Institutional Context

The English public (state) schooling system combines elements of local control and autonomy with centralized testing and a national curriculum. There is restricted school choice and school budgets are linked to enrolment levels. Each school has its own governing board, consisting of parent

[9]For example, in the context of pre-school programmes, Haskins and Barnett (2010) note that the US administration has proposed a review of all Head Start centers, a central component of which will be in-class evaluation of the quality of teacher support and instruction by external assessors. In recent reviews of test-based school accountability, Lang (2010) and Neal (2010) highlight some of the pitfalls of an exclusive focus on test scores; both authors speculate that inspections may provide a more nuanced approach. Grubb (2000) notes small pockets of experimentation with such systems in the US.

governors and representatives from the local authority as well as the local community. Schools manage their own budgets and the school's governing board is responsible for hiring the school principal.[10] Testing of students takes place at ages 7, 11, 14 and 16; these are known as the Key Stage 1 to Key Stage 4 tests, respectively[11]. Successive governments have used these Key Stage tests, especially Key Stages 2 and 4, as key measures of school performance in holding schools to account. For further details see, for example, Machin and Vignoles (2005).

A second plank in England's school accountability regime is the school inspection system. Since the early 1990s all English public schools have been inspected by the Office for Standards in Education, or Ofsted, an independent government agency. As noted by Johnson (2004), Ofsted has three primary functions: (i) offer feedback and advice to the school principal and teachers; (ii) provide information to parents to aid their decision-making process; and (iii) identify schools which suffer from 'serious weakness'.[12] Although Ofsted employs its own in-house team of inspectors, the body contracts out the majority of inspections to a handful of private sector and not-for-profit organizations via a competitive tendering process.[13] Setting overall strategic goals and objectives, putting in place an inspection framework which guides the process of inspection, as well as maintaining responsibility for the quality of inspections, remain with Ofsted.

Over the period relevant to this study, schools were generally inspected once during an inspection cycle, which typically lasted between three and six years.[14] An inspection involves an assessment of a school's performance on academic and other measured outcomes, followed by an on-site visit to the school, typically lasting between one and two days for primary schools.[15] The Ofsted inspection exercise may be viewed as a two-stage process. In the first stage, inspectors form a prior about the school on the basis of 'hard' data', namely test scores, combined with background characteristics used to match the school with its peer group. The second stage - visiting the school - can be interpreted as an attempt to uncover the unobserved heterogeneity, which may help explain why the school under- or over-performs relative to its peer group of schools. Inspectors arrive at the school at very short notice (maximum of three days), which in theory should enable inspectors to see the school as it 'really is' and limit disruptive 'window dressing' in

[10] In ascending order of autonomy, the main types of public schools in England are as follows: Community (these make up the majority of schools); Voluntary Controlled and Voluntary Aided (usually religious and linked to a charitable foundation); and Foundation schools. See Clark (2009) for an analysis of the effects of school autonomy in England.

[11] Note that the age-14 (Key Stage 3) tests were abolished in 2008.

[12] In its own words, the inspectorate reports the following as the primary purpose of a school inspection: "The inspection of a school provides an independent external evaluation of its effectiveness and a diagnosis of what it should do to improve, based upon a range of evidence including that from first-hand observation. Ofsted's school inspection reports present a written commentary on the outcomes achieved and the quality of a school's provision (especially the quality of teaching and its impact on learning), the effectiveness of leadership and management and the school's capacity to improve." (Ofsted, 2011, p.4).

[13] As of 2011, Ofsted tendered school inspections to three organizations, two are private sector firms, the third is not-for-profit.

[14] From September 2009 schools judged to be good or better are subject to fewer inspections than those judges to be mediocre or worse.

[15] English primary schools - the focus of this study - cater for students between the age of 5 and 11.

preparation for the inspections.[16] Importantly for the empirical strategy employed in the current study, inspections take place throughout the academic year, September to July.

During the inspection visit inspectors collect qualitative evidence on performance and practices at the school. A key element of this is classroom observation. As noted in Ofsted (2011b): "The most important source of evidence is the classroom observation of teaching and the impact it is having on learning. Observations provide direct evidence for [inspector] judgements..." (p. 18, paragraph 64). In addition, inspectors hold in-depth interviews with the school leadership, examine students' work and have discussions with pupils and parents. The evidence gathered by the inspectors during their visit as well as the test performance data form the evidence base for each school's report, which is released soon after the inspection. The school is given an explicit headline grade, ranging between 1 ('Outstanding') and 4 ('Unsatisfactory'; also known as a fail rating) and the inspection report is made available to parents and posted on the Internet.[17]

Simple regression analysis (not reported here for brevity) suggests that although inspectors place substantial weight on test outcomes, schools ranked relatively low on this measure can nevertheless receive the best inspection ratings. Conditional on test scores, a *rise* in the proportion of students eligible for free lunch is associated with an *improvement* in the inspection outcome. This provides support for the view that inspectors take the circumstances of the school into account when determining the inspection rating.

There are no direct incentives for schools receiving the highest ratings. Schools rated 'fail' are subject to frequent re-inspection. Typically, around five per cent of schools fail in any given year. At a theoretical level, a fail rating may lead to 'exit' and 'voice' responses from parents. Hussain (2009) provides some initial evidence of the former: a fail rating leads to declines in enrolment levels in the years after inspection. (There is little response in the year of inspection itself, perhaps because of the high costs associated with switching schools mid-year.) Repeated inspections and greater oversight, as well as the threat of sanctions for the school leaders, may lead to higher effort on the part of teachers.[18] A priori, it seems plausible that there is a great deal of heterogeneity in parents' ability to hold individual teachers to account, even in the same school. In this case, the potential benefits of a fail inspection may fall unevenly across students from different socioeconomic backgrounds. This idea partly motivates the investigation of heterogeneous treatment effects examined in section 5 below.

[16] This short notice period has been in place since September 2005. Prior to this, schools had many weeks, sometimes many months, of notice of the exact date of the inspection. Anecdotal evidence suggests that these long notice periods resulted in disruptive preparations for the inspections. There is some evidence to suggest that inspections may have had a small adverse effect on test scores in the year of inspection during this long-notice inspection regime (see Rosenthal 2004). See also Matthews and Sammons (2004).

[17] These can be obtained from http://www.ofsted.gov.uk/.

[18] On the possibility of increased monitoring crowding out trust and intrinsic motivation see, for example, Frey (1993).

# 3 Evidence on the Validity of Inspection Ratings

This section investigates whether inspection ratings convey any information on school quality beyond that which is already captured by, for example, test score rankings. The critical question is whether inspectors visiting the school are able to gather and summarize information about underlying school quality which is not already available in the public sphere.[19]

In the analysis below a measure of underlying school quality is constructed from student (age 14) survey responses to questions about teacher behavior and practices. These data come from the Longitudinal Survey of Young People in England (LSYPE), a major survey supported by the Department for Education. (The online appendix 1 provides details of the survey and further results for survey questions relating to school discipline as well as parental satisfaction.) The survey asks the following six questions on how likely teachers are to: take action when a student breaks rules; make students work to their full capacity; keep order in class; set homework; check that any homework that is set is done; and mark students' work.

A composite student-level score is computed by taking the mean of the responses to these six questions (see the online appendix 1 for further details). These student-level means are then converted into z-scores by normalizing them to mean zero and standard deviation one.[20] The validity test is implemented by regressing the composite z-scores, $q$, on inspection ratings as well as other school and respondent family background characteristics:

$$q_{ijk} = aX_{ij} + b.\text{Rating}_j + \lambda_k + u_{ijk}. \tag{1}$$

$i$ indicates individual survey respondent (the unit of observation) at school $j$ in local authority ('district') $k$. $X_{ij}$ captures school- and student-level variables and $\lambda_k$ represents local authority fixed effects. School-level variables include the school's national percentile test rank and the proportion of students eligible for a free lunch. '$\text{Rating}_j$' is the school's inspection rating.[21]

The key issue here is whether inspection ratings are correlated with the underlying measure of school quality not observed by inspectors, conditional on observed school characteristics such as test rankings, the proportion of students receiving a free lunch, whether the school is secular or religious, as well as survey respondent background characteristics. The inspection ratings are then

[19] Prior evidence suggests that inspectors' findings are *reliable*: Matthews et al (1998) show that two inspectors independently observing the same lesson come to very similar judgements regarding the quality of classroom teaching. The question addressed here is whether inspection ratings are also *valid* in the sense of being correlated with underlying measures of school quality not observed by the inspectors.

[20] A higher z-score corresponds to higher quality as reported by the student.

[21] On the timing of the survey and the inspections, note that students were asked questions relating to teacher practices in the academic year 2003/04. I extract from the data those LSYPE respondents who attend a school inspected soon after the survey, i.e. between 2004/05 and 2006/07. '$\text{Rating}_s$' corresponds to the rating from one of these three years. (Recall that over this periodschools were inspected once every five or six years.) Most schools will also have been inspected prior to 2003/04. To guard against the possibility that the previous inspection rating may influence student survey responses, the regression results reported below also control for past inspection ratings. Students in schools inspected in 2003/04, the year of the survey, are excluded, because it is difficult to pinpoint whether the survey took place before or after the inspection. This yields a sample of just over 10,000 students enrolled in 435 secondary schools.

said to be *valid* if the coefficient on the inspection rating variable, $b$, remains statistically significant and economically meaningful in the 'long' regression (1). Note that this parameter simply captures the association between the inspection rating and the measure of 'quality' (teacher practices), $q$; it does not estimate a causal effect.

Another way to view this validity test is as follows. 'Insider' views of the school from students (and their parents) potentially provide useful information to other decision makers.[22] Such feedback information from consumers is typically not observed in the public sector. Inspection ratings may be useful in forecasting consumers' views of quality: for example, if inspection ratings can be used to predict student perceptions of the quality of teaching, then parents currently engaged in choosing schools may put some weight on these ratings when making their decisions.

*Results*

Although the main focus of the analysis here is the relationship between the survey school quality measures and inspection ratings, it is useful to first investigate the relationship between the survey z-scores and the school test rank. This will then provide the analyst (or parent) a benchmark by which to assess the relationship between the survey z-scores and the inspection ratings. Column 1 of Table 1 demonstrates that there is a strong and statistically significant relationship between teacher practices as reported by students and the school test ranking: a rise of 50 national percentile ranks is related to 0.32 (50 x 0.0064) of a standard deviation improvement in the teacher practices composite score.[23]

Turning now to the main issue of interest, column 3 of Panel A shows the unconditional relationship between the teacher practice z-score and the inspection rating after the survey was administered. The result suggests that each unit decline in performance on the inspection rating is associated with 0.22 of a standard deviation decline in the teacher practices z-score. Thus, the gap in the teacher practices z-scores between an Outstanding (Grade 1) and a Fail (Grade 4) school is around 0.7 of a standard deviation. If we take the results for test rankings (column 1) as a benchmark then these are clearly large effects.[24]

Controlling for test rankings and the proportion of students receiving a free school meal in column 4 leads to a 40% reduction in the association between the inspection rating and the teacher practices z-score, but in relative terms, the estimate remains large, and is highly significant. (Also included as controls in column 4 are the size of the school and the type of school as well as local

[22] As Heckman (2000) has noted of public schools in the US: "One valuable source of information - parental and student perception of the qualities of teachers and schools - is rarely used to punish poor teaching" (Heckman, 2000, p. 24).

[23] Interestingly, the results in column 2 suggest that on one measure of poverty at least, the link with school quality is relatively weak: a rise of 0.2 (one standard deviation) in the fraction of students eligible for free lunch is associated with a fall in the teacher practices z-score of 0.04 of a standard deviation. As seen below, this relationship becomes weaker still as additional controls are added.

[24] Note that at 0.03, the adjusted R-squared value for the regressions in columns (1) and (3) is of a similar size. The fact that the this is small should not come as a surprise, given the noisy nature of the left hand side variable, school quality. Recall that this is derived from a survey of around 20 students from each school, representing just 2 or 3 per cent of the total student body.

education authority fixed effects.) There are two potential criticisms to this exercise. First, there may be a concern that students from different socioeconomic backgrounds respond to the survey questions in systematically different ways, even if underlying teacher practices are the same. For example, students from poorer backgrounds or those scoring low marks on prior tests may have more negative or positive opinions about teachers than richer or better performing students. There is then the possibility that the relationship between inspection ratings and the survey z-scores is an artefact of this sort of bias in response to the survey questions. Column 5 includes detailed controls on students' family background and prior test scores.[25] These lead to a minor fall in the absolute size of the coefficient on inspection ratings, which remains statistically significant at the 1 per cent level.

A second potential criticism is that students' survey responses may be influenced by past inspection ratings. If a school's inspection ratings are correlated over time then the effects of inspection ratings after the survey interview shown in Table 1 may simply be capturing the effect of past inspections on respondents' views. In order to investigate the possibility that this mechanism is driving the results, column 6 includes additional controls for the inspection rating prior to the year of the student interview, 2003/04. The results show that the effect of including dummies for the most recent inspection rating before the interview has only a small effect on the estimated effect.

The results in column 6 with the full set of controls suggest that worse inspection ratings are associated with sharply declining school quality as measured by student reports of teacher practices. The strength of this gradient may be gauged by comparing the decline in quality associated with declines in test rankings: the results show that a 50 percentile point decline in a school's test ranking is associated with a decline of 0.15 (0.0029 x 50) of one standard deviation in the teacher practices z-score. Compare this with a one unit deterioration in the inspection rating: this is associated with a decline of 0.10 of one standard deviation in the teacher practices z-score.

Finally, by including inspection dummies, column 7 investigates whether the linearity assumption implicit in the previous models is justified. The results in column 7 suggest a concave relationship between teacher practices and inspection ratings: the gap is largest when we move from a Grade 1 school (the omitted category) to Grade 2; it is smallest between Grade 3 and Grade 4 (Fail) schools. This is noteworthy in that it suggests that on this measure at least, there are a large number of schools (Grade 3) which are not dramatically different from outright Fail schools.

The online appendix 1 repeats the above analysis for student-level mean z-score for the three school discipline questions (relating to class disruptions and behavior and overall discipline in the school) and five parent satisfaction questions (relating to interest teachers show in the child, school discipline, feedback from the school and overall satisfaction in child's school progress). The results for these outcomes are very similar to those reported for the teacher practices outcome in Table 1: the association between inspection ratings and student- and parent-reported school

[25] These include controls for age 11 test score, administered in the student's primary school prior to enrolling at the current secondary school; parents' education, income, employment history and whether in receipt of various government benefits; whether a single parent household; and number of children in the household.

quality (discipline) outcomes is strong.

In summary, this analysis reveals that inspection ratings can help detect good and poor teacher practices (or high and low parental satisfaction as reported in the online appendix) among schools with the same test rankings and socioeconomic composition of students. The results paint a highly consistent picture across all the student and parent measures: the inspection ratings do indeed convey information about school quality over and above that already contained in publicly available information such as test scores, school type, the proportion of students eligible for free lunch, etc. Moreover, separate regression results (not reproduced here) for each of the items which make up the composite scores also point to the same conclusion. For example, each of the six items which make up the teacher practices composite score show that the relationship with inspection ratings is negative and statistically significant. I.e., a better inspection rating is associated with better teacher practices on each of the six underlying measures. This implies that conditional on observable school and student characteristics, students at higher rated schools experience an environment where teachers are more likely to: take action when a student breaks rules; make students work to their full capacity; keep order in class; set homework; check that any homework that is set is done; and mark students' work.[26]

# 4 The Effect of a Fail Inspection on Test Scores: Empirical Strategy

The primary question addressed here is: What is the effect of a fail inspection on students' subsequent test scores? An analysis using before- and after-fail test score data for a panel of schools quite possibly confounds any effect of a fail rating with mean reverting behavior of test scores. For example, if inspectors are not fully able to account for idiosyncratic negative shocks unrelated to actual school quality, then poor test score outcomes a year or two prior to the inspection may lead to a fail. The concern is that any rise in test scores following inspection would in fact have occurred even in the absence of a fail rating.[27]

This study exploits a design feature of the English testing system to address such concerns. The age-11 'Key Stage 2' tests – administered at the national level and a central plank in student and school assessment – take place over five days in the second week of May each year. The results of these tests are then released in mid-July. The short window between May and July allows me

[26] Previous theoretical and empirical work suggests that the measures of school practices used in this study 'matter'. For example, theoretical work by Lazear (2001) and the empirical evidence in Hoxby (2000) and Figlio (2007) suggests that disruptions have a negative effect on student outcomes.

[27] Examples of such idiosyncratic shocks might include an outbreak of a cold virus during exam time or other disruptions such as building works. Mean reversion problems plague evaluations of labor market programs. The fundamental issue in these cases and our context is that assignment to treatment (enrolment in training; a fail inspection rating) is based on past values of the outcome variable (earnings; test scores). If a bad draw in the pre-treatment year is reversed in the post-treatment year, then the re-bound in earnings (test scores) cannot be wholly attributed to the treatment. See, for example, Ashenfelter (1978) and Heckman, Lalonde and Smith (1999) for job training programs and Kane and Staiger (2002) and Chay, McEwan and Urquioal (2005) for discussions relevant to evaluation of school interventions.

to address the issue of mean reversion: schools failed in June are failed *after* the test in May but *before* the inspectors know the outcome of the tests. Thus the May test outcome for these schools is not affected by the subsequent fail, but neither do inspectors select them for failure on the basis of this outcome. (See Figure 2 for an example time line for the year 2005/06.)

< Figure 2 here >

This insight enables me to identify credible causal estimates of the short-term effects of a fail inspection. In particular, and taking the year 2005/06 as an example again, the question addressed is: for schools failed in September 2005, what is the effect of the fail inspection on May 2006 test scores?

The evaluation is undertaken by comparing outcomes for schools inspected early in the academic year, September – the treatment group – with schools inspected in June, the control group.[28] Schools failed in September have had almost a whole academic year to respond to the fail treatment. The identification problem, that the counterfactual outcome for schools failed in September is not observed, is solved via comparisons with June failed schools. The details of these comparisons are described below.

*Descriptive Statistics*

A key question is *why* some schools are inspected earlier in the year than others. The descriptive analysis in Table 2 helps shed light on this question. This table shows mean characteristics for *all* primary schools inspected and failed in England in the four years 2005/06 to 2008/09.[29] For each year the first two columns show means for schools failed early in the academic year (September to November[30]) and those failed late in the year (from mid-May, after the Key Stage 2 test, to mid-July, before the release of test score results). The former category of schools are the 'treatment' group and the latter the 'control' group. The first row simply shows the mean of the month of inspection. Given the selection rules for the analysis, these are simply June (between 6.1 and 6.2) and October (between 10.1 and 10.2) for the control and treatment groups.

The second row, which shows the year of the previous inspection, offers an explanation why some schools are inspected early in the year and others later on. The 2005/06 fail columns, which are typical of all four fail years, show that the mean year of inspection for late inspected schools is 2000.6; for early inspected schools it is 2000.1.[31] This suggests that schools inspected slightly

[28] For an evaluation of the effects on test scores it is important to note that the latest available test score information at the time of inspection is the same for both control and treated schools: i.e. from the year before inspection.

[29] I focus on these four years because 2005/06 is the first year when the inspection system moved from one where schools were given many weeks notice to one where inspectors arrived in schools with a maximum of two days notice. In order to analyze the effect of a first fail (a 'fresh fail') , schools which may have failed in 2004/05 or earlier are dropped from the analysis. This results in a loss of 10 per cent of schools.

[30] The early inspection category is exapnded to three months in order to increase the sample of 'treated' schools.

[31] Note that an inspection in the academic year 1999/00 is recorded as '2000'; an inspection in 2000/01 is recorded as '2001', and so on.

earlier in the previous inspection round are also inspected slightly earlier in 2005/06. The online appendix Table 2 shows that in general, over this period, inspectors followed a mechanical rule with regards to the timing of inspections - schools which were inspected early in the first inspection round in the mid-1990s were inspected early in subsequent inspection rounds. Table 2 shows that for fail schools, within a given year, the month of the inspection appears to be determined by the timing of the previous inspection.[32]

The third and fourth rows report the proportion of students receiving a free school meal (lunch) and the proportion of students who are white British at treatment and control schools, respectively. Across each of the four inspection years the differences in means between the two groups appear to be small and are statistically insignificant. Similarly, there are no statistically significant differences between the early and late inspected schools in the previous inspection rating, except for the year 2008/09.

Finally, national standardized test scores for the cohort of 11-year olds in the year prior to the inspection are reported in rows six and seven. Once again, these show no evidence of statistically significant differences between the two groups. It is noteworthy that this set of fail schools perform between 0.4 and 0.5 of one standard deviation below the national mean.

In sum, the evidence in Table 2 that there is little difference between control and treatment schools on observable characteristics combined with the fact that timing is determined by a mechanical rule suggests that there are unlikely to unobservable differences between control and treatment schools. Thus it would appear that when comparing early and late failed schools within a year, treatment is as good as randomly assigned.

*OLS and Difference-in-Differences Models*

For ease of exposition, I will consider the case of the schools failed in 2005/06 in the months of September and June. The analysis extends to schools failed in the early part of the year (September to November) versus those failed late in the year (mid-May to mid-July) in each of the four inspection years analyzed. First, define a treatment dummy, $D_s = 1$ if school $s$ is failed in September 2005 and $D_s = 0$ if the school is failed in June 2006. For student $i$ at each of these two groups of schools the two potential outcomes for the May 2006 standardized test score are given as follows:

$$y^0_{is,06} = \alpha + u_{is},$$
$$y^1_{is,06} = \alpha + \delta_i + u_{is}$$

where $y^0_{is,06}$ is the outcome if the school *is not* failed in September 2005 and $y^1_{is,06}$ is the outcome if the school *is* failed in September 2005. For students attending schools failed in September 2005 the counterfactual outcome, $y^0_{is,06}$, is not observed. $\delta_i$ is the student-specific gain from treatment. The

[32]It should be noted that uncertainty about the exact timing of inspections remains. For example, the standard deviation for the year of previous inspection is 0.9 years for schools in 2005/06 fail columns, Table 2.

role of conditioning variables in the analysis is discussed in section 4.1 below. Realized outcomes can then be stated as follows:

$$\begin{aligned} y_{is,06} &= (1 - D_s)y^0_{is,06} + D_s y^1_{is,06} \\ &= \alpha + \delta_i D_s + u_{is}. \end{aligned} \tag{2}$$

Given the evidence on assignment to September versus June inspections presented in the previous sub-section, we can credibly argue that treatment status $D_s$ is uncorrelated with both the residual $u_{is}$ and the student-specific gain $\delta_i$. Thus, a comparison of means for the treatment and control outcomes yields the parameter of interest, the average effect of treatment on the treated (ATT), $E(y^1_{is,06} - y^0_{is,06} \mid D_s = 1) = E(\delta_i \mid D_s = 1)$. This the effect of a fail inspection rating for schools inspectors judge to be failing.[33]

Below, results are also presented using difference-in-differences (DID) models. The evidence in Table 2 suggests that although there are no statistically significant differences in the levels of prior test score outcomes and other school characteristics across the control and treatment groups, small differences remain. These small differences may lead to biased estimates from simple comparisons of post-treatment outcomes, especially if the gains from treatment are also small. The DID approach may then be viewed as a robustness check. It is implemented as follows. Continuing with the example of schools failed in 2005/06, data are taken from two periods for the DID analysis, form 2004/05 (the 'pre' year) and 2005/06 (the 'post' year). In the DID model realized test scores are determined as follows:

$$y_{ist} = \alpha + \gamma post_{06} + \delta_i D_{st} + \lambda_s + u_{ist}, \tag{3}$$

where $t = 2005$ or 2006; $post_{06}$ is a dummy indicator, switched on when $t = 2006$; $D_{st}$ is now a time-varying treatment dummy, switched on in 2006 for schools inspected in September (i.e. the interaction between $post_{06}$ and the dummy indicating early inspection, $D_s$); and $\lambda_s$ is a school fixed effect. $\delta_i$ is the student-specific gain from treatment.

The DID assumption, which embodies the assumption of common trends across treatment and control groups, is that conditional on the school fixed effect ($\lambda_s$) and year ($post_{06}$) the treatment dummy $D_{st}$ is uncorrelated with the residual, i.e. $E(u_{ist} \mid \lambda_s, post_{06}, D_{st}) = 0$. The regression version of the DID model estimates $E(\delta_i \mid D_s = 1)$, which is the ATT.[34]

Essentially, the effect of a fail inspection is uncovered by comparing the change in scores between the May 2005 and May 2006 tests for schools inspected early in the academic year (September 2005) versus those inspected late in the year (June 2006).[35] The key assumption is that any rebound in test scores which would have occurred in the absence of a fail inspection for Sep-

[33] Another parameter of policy interest might be the Marginal Tratement Effect, i.e. the test score gain for students in schools on the margin of being failed.

[34] For details, see, for example, Blundell and Costa-Dias.

[35] Note that age-11 test scores are observed for two different cohorts across the two years.

tember failed schools is captured by changes observed for the June failed schools (represented by the coefficient on the $post_{06}$ dummy in regression model (3)). The difference in performance between the two groups of schools yields the effect of the treatment. One prior would be that some of the decline in test scores observed in 2004/05 is temporary, so that we would expect some mean reversion in test scores. In this case a simple school fixed effect strategy would overestimate the effect of a fail inspection. In the DID setup we would expect a positive coefficient on the $post_{06}$ dummy and hence a lower estimated effect of treatment when compared with the effect implied by the simple fixed effect approach.

## 4.1 Testing for Strategic Behavior

A growing body of evidence has demonstrated that when schools face strong incentives to perform on test scores they game the system. Such gaming behavior may have distributional consequences and may also lead to misallocation of resourses.[36] Evidence of the following types of dysfunctional response has been documented. First, studies show that under test-based accountability systems teachers may remove low ability students from the testing pool, for example by suspending them over testing periods or reclassifying them as special needs (Jacob 2005, Figlio 2006, Figlio and Getzler 2006, Cullen and Reback 2006). Second, teachers may 'teach to the test,' so that the performance measure (the high stakes test) rises whilst other aspects of learning may be neglected (see e.g. Koretz, 2002). Third, when schools are judged on the number of students attaining a given proficiency level it has been shown that teachers target students close to the proficiency threshold (see, for example, Burgess et al 2005, Reback 2008 and Neal and Schanzenbach 2010). Fourth, there may be outright cheating by teachers (Jacob and Levitt 2003).

In the analysis below, I test for the presence of the first three of these types of strategic responses. First, I examine to what extent gains in test scores following the fail rating are accounted for by selectively removing low ability students.[37] This involves checking whether the estimated effect of treatment in the OLS and DID regressions ($\delta$ in equations (2) and (3) above) changes with the inclusion of student characteristics such as prior test scores, special education needs status, free lunch status and ethnic background. For example, suppose that in order to raise test performance fail schools respond by removing low ability students from the test pool. This would potentially yield large *raw* improvements in test scores for treated schools relative to control schools. However, conditioning on prior test scores would then reveal that these gains are much smaller or non-existent. This test enables me to directly gauge the effect of gaming behavior on

[36] Courty and Marschke (2011, p. 205) provide the following definition: "A dysfunctional response is an action that increases the performance measure but is unsupported by the designer because it does not efficiently further the true objective of the organization." The authors go on to propose a formal classification of dysfunctional responses based on the multitasking model.

[37] It should be noted that potentially distortionary incentives may well exist prior to the fail rating. However, these incentives become even more powerful once a school is failed. Thus the tests for gaming behaviour outlined here shed light on the effects of any *extra* incentives to game the system following a fail inspection. As noted previously, the evidence shows that the incentives to improve test score performance following a fail inspection are indeed very strong.

test scores.[38]

Second, I test for whether any gains in test scores in the year of the fail inspection are sustained in the medium term. This provides an indirect test of the extent of teaching to the test. More precisely, the 11-year-old students are in their last year of primary school at the time of the fail. The issue is whether any gains in test scores observed in that year can still be detected when the students are tested again at age 14, three years after the students left the fail primary school. Note that this is a fairly stringent test of gaming behavior since fade-out of test score gains is typically observed in settings even when there are no strong incentives to artificially boost test scores (see, for example, Currie and Thomas, 1995).

Third, I analyze the distributional consequences of a fail inspection. In particular, I investigate whether there is any evidence that teachers target students on the margin of achieving the key government target for Year 6 (age 11) students.[39] It was noted above that the percentage of students attaining the 'Level 4' proficiency on the age-11 Key Stage 2 test is a key measure of performance used by the government. It is also the headline school performance measure and hence is commonly used to rank schools. We might then expect teachers and schools to target resources towards students on the margin of attaining this threshold, to the detriment of students far below and far above this critical level.

A number of strategies are adopted to explore this issue. In the first approach I examine whether gains in student test scores vary by prior ability. Prior ability predicts the likelihood of a student attaining the performance threshold. Previous evidence has shown that teachers may neglect students at the bottom of the prior ability distribution in response to the introduction of performance thresholds (see Neal and Schanzenbach, 2010).

In the current setting, the official expectation is for students to attain 'Level 4' on the Year-6 Key Stage 2 test. Table 3 shows the distribution of Year 6 students achieving this target for mathematics and English at fail schools, in the year before the fail, by quartile of prior ability. Prior ability is measured by age seven test scores. As should be expected, Table 3 shows that ability at age seven is a strong predictor of whether a student attains the official target: the proportion doing so rises from between a quarter and a third for the bottom quartile to almost 100 per cent at the top quartile of prior ability. As the final rows of Table 3 show, in the year before the fail inspection the average number of students achieving the Level 4 threshold is 67 and 72 per cent for mathematics and English, respectively. One implication of the evidence presented in Table 3 is that students in the lowest ability quartile are the least likely to attain the official threshold, and so teachers may substitute effort away from them towards students in the second quartile. The analysis below tests this prediction.

A second approach to analyzing whether teachers selectively target effort towards students on

[38]Note that the evidence presented in Table 2 suggests that the school's treatment status is uncorrelated with observable student characteristics. Hence, although the inclusion covariates may reduce the estimated standard error of the treatment effect, we would not expect the inclusion of student background characteristics to change the estimated coefficient. That is *unless* the schools are engaging in the type of gaming behaviour highlighted here.

[39]In primary schools, national tests are administered to students in England at ages seven and 11.

the margin of attaining the mandated threshold is to investigate the distributional effects of a fail rating *within* prior ability quartiles. For example, if teachers set or track students within or among classrooms by ability, then they may target the marginal students within these ability groups.

Figure 3 illustrates this idea. Suppose that test scores in the year before a fail inspection are distributed as in this stylized example. The figure shows the distribution of test scores for each of the four prior ability quartiles, as well as the proportion of students who pass the official proficiency threshold, 'T0'. For illustrative purposes, suppose that 20 per cent of students from the bottom quartile attain proficiency; 50, 75 and 90 per cent do so in the second, third and top quartiles, respectively. Following a fail inspection the incentives to maximize students passing over the threshold may be more intense than prior to the fail rating. If schools are able to game the system (for example, if inspectors are unable to detect such strategic behavior) then they may target the students on the margin of attaining the proficiency level. Suppose that the distribution of potential test scores is similar in the year of inspection as it is in the year before, so that Figure 3 also depicts the potential test scores for students in the year of inspection. Then if teachers are able to detect the marginal students, they may allocate greater effort towards the students who lie on the boundary of the shaded area in each of the four charts in Figure 3.

The analysis below tests for such teacher behavior by examining the effect of treatment at specific quantiles of the test score distribution. Thus, quantile treatment effects are estimated to establish whether or not the largest gains are around the performance threshold boundary, as predicted by this simple theory.

# 5 Results

## 5.1 Basic Results

Table 4 shows results for the effects of a fail inspection on mathematics and English test scores for schools failed in one of the four academic years, 2006 to 2009.[40] The top panel reports results from the OLS model and the bottom panel reports results from the difference-in-differences model. For ease of presentation, the four inspection years are pooled together. The estimated OLS model is as follows:

$$y_{is} = \alpha + \delta D_s + \beta X_{is} + u_{is}, \tag{4}$$

where $y_{is}$ is the age 11 (Key Stage 2) national standardized test score for student $i$ attending school $s$. $D_s$ is turned on for schools inspected early in the academic year and $X_{is}$corresponds to student characteristics such as prior test scores, special education needs status, free lunch status and ethnic background.[41] In this setup, the comparison is between students attending schools failed early in the academic year in one of the years 2006, 2007, 2008 and 2009 - the treatment

[40] Note that '2006' refers to the academic year 2005/06 and so on for the other years.

[41] Homogeneous treatment effect notation has been adopted for model (4). The discussion in section 4 made it clear that if the treatment effect varies across units then (4) identifies the average effect of treatment on the treated.

group - with those attending schools failed late in the academic year in one of these four years - the control group.

Pooling over the four years is justified because, first, the timing of inspections is arbitrarily determined[42] and, second, over these four years schools were inspected and rated in a consistent manner.[43] The evidence presented in Table 2 shows that schools are indeed comparable across the different years. As a robustness check, results from regression analysis conducted for each year separately are also reported (in Appendix Tables A1 and A2). As will be seen, these show that results for the pooled sample and for individual years produce a consistent picture of the effects of a fail inspection.

Turning first to mathematics test scores, the row 'early Fail' in Panel A of Table 4 corresponds to the estimate of the treatment effect $\delta$ in equation (4). Column (1) reports the 'raw' effect of a fail inspection, i.e. without any controls. The result in column (1) suggests that the effect of a fail rating is to raise standardized test scores by 0.11 of a standard deviation. This effect is statistically significant at conventional levels (standard errors are clustered at the school level).

As explained in section 4.1 above, the estimated effect in column (1) may in part reflect distortionary behavior by teachers. If schools respond to a fail inspection strategically, for example, by excluding low ability students from tests via suspensions, then we should see the relatively large gains in column (1) diminish once prior ability controls are introduced in the regression analysis. In order to address such concerns, columns (2) and (3) introduce student-level controls. Regression results reported in column (2) include the following student characteristics: gender; eligibility for free lunch; special education needs; month of birth; whether first language is English; ethnic background; and census information on the home neighborhood deprivation index. The model in column (3) also includes age seven (Key Stage 1) test scores. Dummies for missing covariates are also included.

The rise in the R-squared statistics as we move from columns (1) to (2) and then (3) clearly indicates that student background characteristics and early test scores are powerful predictors of students' test outcomes. However, the addition of these controls appears to have little effect on the estimated effects of the fail rating. Overall, the evidence in Panel A for mathematics suggests that (i) the effect of a fail inspection is to raise test scores and (ii) this rise does not appear to be driven by schools selectively excluding (by ability, for example) students from the tests.

Turning to the difference-in-differences estimates for mathematics reported in Panel B, a nice feature of this approach is that it provides direct evidence on the importance of mean reversion. For the DID analysis the 'pre' year corresponds to test scores prior to the year of inspection (i.e. test scores from the 2004/05 exam for schools failed in 2005/06; 2005/06 tests for schools failed in 2006/07, etc.) whilst the 'post' year corresponds to test scores from the year of inspection. The

[42] As discussed earlier and demonstrated in Appendix Table 1, whether a school is inspected early or late in an inspection cycle depends on whether the school was inspected early or late in the previous cycle.

[43] As described earlier, changes to the inspection process were introduced in September 2005. Arguably, the biggest change was a move to very short (two days) notice for inspections, down from up to two months notice. This regime has remained in place since September 2005.

estimate of mean reversion is provided by the difference between test scores in the pre-inspection year and test scores in year of inspection for schools failed *late* in the academics year (i.e. the control group). This estimate is indicated in the row labeled 'post'.

Meanwhile, the DID estimate of the effect of a fail inspection is identified from any extra gain in test scores between these two periods for schools failed *early* in the academic year (the treatment group). This estimate is provided in the first row of Panel B, labeled 'post x early Fail' which corresponds to the treatment dummy $D_{st}$ in equation (3).

The DID results are exactly in line with the OLS results: column (3) of Panel B shows that students at early failed schools gain by 0.12 of a standard deviation relative to students enrolled in late fail schools. In addition, comparing results with and without student-level controls - column (1) versus columns (2) and (3) - shows that there is little change in the estimated effect. These results support the contention that a fail inspection raises student test scores and, further, that these gains are unlikely to be accounted for by the kind of strategic behavior outlined above.

As for evidence on mean reversion, the results in the second row of Panel B show that there is only mild mean reversion for mathematics. With the full set of controls, the coefficients on the 'post' dummy is 0.03 of a standard deviation and is not statistically significant at conventional levels. This suggests that in the absence of a fail rating from the inspectors, we should expect very small or even zero gains in test scores from the low levels in the base year reported in the descriptive statistics in Table 2.

Columns (4) to (6) report results for English test scores. The OLS results in column (6), Panel A show that the effect of a fail inspection is to raise standardized test scores by 0.08 of a standard deviation. The DID estimates in Panel B point to gains of around 0.07 of a standard deviation. These estimates are statistically significant.

In line with the results for mathematics, the results for English provide no evidence of gaming behavior: although the predictive power of the controls is large, as indicated by the rise in the R-squared statistics, there is little change in the estimates when we move from the column (4), no controls, to column (6), full set of controls.

Finally, the evidence on mean reversion of English test scores presented in the second row of Panel B is noteworthy. This time there is stronger evidence of a re-bound in test scores from the low level in the base year. The coefficients on the 'post' dummy is now 0.08 of a standard deviation, indicating a substantial re-bound in test scores even in the absence of a fail inspection. As seen below, this re-bound in fact corresponds to a 'pre-program' dip observed in the year before inspection.[44]

*Falsification Test and the 'Pre-Program Dip'*

[44]Note that the above results from data pooled over the four inspection years are in line with results from regression analysis conducted for each year separately, reported in Appendix Tables A2 and A3. For example, the results in each of the columns labeled (3) in Table A2 show that the effect of a fail inspection on mathematics test scores ranges between 0.06 and 0.15 of a standard deviation across all four years and both OLS and DID estimation strategies.

Table 5 presents analysis from a falsification exercise. This makes use of the fact that data are available in both the year before and two years before treatment in order to conduct a placebo study. The question addressed is: when we compare the treatment and control groups in the year before treatment, can we detect a treatment effect when there was none?

Table 5 pools the data over the four inspection years. The OLS estimates in Panel A compare test score outcomes in the year before inspection for students at early and late failed schools. Focusing on columns (3) and (6) with the full set of controls, these show that the estimated effect of the placebo treatment is small, statistically insignificant and close to zero for mathematics and English. The DID estimates in Panel B, which compare the change in test scores one and two years before inspection for early and late failed schools, also show no evidence of placebo effects, supporting the common trends assumption underlying the DID strategy.

The online appendix tables 3 and 4 present the results for individual inspection years. The results in these two tables confirm the finding that the placebo treatment produces no discernible effect. For example, the OLS results in the columns labeled (3) in Panel A of the online appendix table 3 show that the estimated effect of the placebo treatment is small, statistically insignificant and close to zero on average across the four years for mathematics and English.

There is one remaining feature of the results in Table 5 which is worthy of mention. This is the evidence on the preprogran dip in test scores, presented in the row labeled 'post' in Panel B. The results in column (3) for English show that there is a large, statistically significant decline in test scores in the year prior to the fail rating which cannot be explained by student characteristics or their prior test scores. This effect, -0.08 of a standard deviation, is the same as the re-bound reported in the corresponding cell of Table 4.

## 5.2 Heterogeneous Treatment Effects

In this section I explore the distributional consequences of a fail inspection. The analysis below first assesses whether the treatment effect varies by prior ability. The discussion then turns to quantile treatment effects, followed by some further subgroup analysis. The final section summarizes the results from the heterogeneous effects analysis.

*Effects by Prior Ability*

As discussed in section 4.1 above, variation in treatment effect by prior ability may provide evidence of distortionary teacher behavior. But in order to assess whether teachers strategically allocate effort among students so that the number of students passing the performance threshold is maximized, it is important to first consider who might be the 'marginal' students. Recall that the official expectation is for students to attain the performance threshold of 'Level 4' on the Key Stage 2 test (typically taken at age 11). As noted earlier teachers may substitute effort away from the lowest ability students if there is little chance of these students crossing this threshold.[45]

[45]Table 2, discussed in section 4.1, highlighted the relationship between prior ability and the probability of

In order to test the prediction that low ability students are adversely affected when incentives to attain the performance threshold are strengthened (following a fail inspection), I test whether the effect of treatment varies with prior ability. The following model incorporating the interaction between the treatment dummy and prior ability is estimated:

$$y_{is} = \alpha + \delta D_s + \gamma . Rank_{is} * D_s + X_{is}\beta_1 + W_s\beta_2 + u_{is}, \tag{5}$$

where the treatment dummy $D_s$ is turned on for schools inspected early in the academic year and $Rank_{is}$ is the percentile rank on prior ability for student $i$ as measured by the student's performance on the Key Stage 1 (age seven) test. $\gamma$ then estimates how the effect of treatment varies by prior ability. The effects of treatment may in fact vary non-linearly by prior ability. This will be the case if, for example, teachers target students in the middle of the prior test score distribution and neglect students at the top and bottom. In order to allow for such non-linear interactions the following regression is also estimated:

$$y_{is} = \alpha + \delta D_s + \sum_{k=2}^{4} \gamma_k Q_{isk} D_s + X_{is}\beta_1 + W_s\beta_2 + u_{is}, \tag{6}$$

where the dummy variable $Q_{isk}$ is switched on for student $i$ if her rank on the prior test score lies in quartile $k$. Thus, $\gamma_k$ estimates the effect of treatment for students lying in quartile $k$ in the prior ability distribution, relative to the omitted category, the bottom quartile.

Table 6, columns (1) and (3), presents estimates of the main ($\delta$) and interaction ($\gamma$) effects for mathematics and English, respectively, for the linear interactions model (5). In each column, the row 'early Fail' corresponds to the estimate of $\delta$ and 'early Fail x prior ability percentile rank' corresponds to the estimate of $\gamma$. The results for both mathematics and English in columns (1) and (3) show that there is a strong *inverse* relationship between prior ability and the gains from treatment. Students from the lowest end of the prior ability distribution gain 0.19 and 0.14 of a standard deviation for mathematics and English, respectively. The interaction term in the second row of columns (1) and (3) suggests that for students at the very top end of the ability distribution gains are close to zero.

The estimates for the nonlinear interactions model, equation (6), are reported in columns (2) and (4).[46] Allowing for non-linearities leaves the above conclusion unchanged: the biggest gains are posted for students from the bottom quartile (the omitted category); students in the middle of the prior ability distribution also experience substantial gains, though not as large as the ones for low ability students. At 0.05 and 0.025 of a standard deviation for mathematics and English, respectively, gains for students in the top quartile appear to be positive, though substantially smaller than for those at lower ability levels.

attaining the target level on the Key Stage 2 test. This showed that only around a quarter or one third of students in the lowest prior ability quartile attain the stipulated Level 4.

[46] Note that running four separate regressions by prior ability quartile subgroup leads to results virtually identical to those reported in columns (2) and (4) of Table 5.

One explanation that may account for the relatively small gains observed for high ability students is that their test scores are at or close to the ceiling of 100 per cent attainment. However, it should be noted that even for students in the highest ability quartile, the mean test scores in the year before treatment are some way below the 100 per cent mark (76 per cent and 68 per cent for mathematics and English, respectively). This hypothesis is explored further (and rejected) in the quantile treatment effect analysis below.

In summary, the results presented in Table 6 show that low ability students reap relatively large test score gains from a fail inspection. This is in contrast to findings from some strands of the test-based accountability literature which show that low ability students may suffer under such regimes.[47] One explanation for the findings reported here may lie in the role played by inspectors. I discuss this at greater length below.

*Quantile Treatment Effects*

Low ability students are at a relative disadvantage in an accountability regime based on performance thresholds if the only information teachers have regarding the probability of a student clearing this hurdle is prior ability, as in Table 3. In this case the number of students attaining the required standard may well be maximized by substituting teacher effort away from those least likely to attain the mandated standard - the average student in the bottom ability quartile - towards those students most likely to attain the threshold as a result of greater teacher focus (students in the second quartile, say).

This conclusion - which is in stark contrast to the findings reported in Table 6 - is based on the assumption that teachers must target the *average* student within each prior ability quartile, say, and that they cannot identify the *marginal* student in a given quartile. The payoff, in terms of passing the mandated threshold, from investing greater effort on the average student in the low prior ability category may indeed be low. However, if teachers can successfully identify the marginal students in, for example, the bottom quartile of the prior ability distribution, then the returns to extra teacher effort may be substantial. The intuition for this line of reasoning was discussed in section 4.1 and illustrated by Figure 3.

One way to detect such teacher behavior is by examining the effect of treatment at specific quantiles of the test score distribution. In particular, the above argument suggests that we should examine both the full sample quantile treatment effects as well as quantile treatment effects within prior ability subgroups. Looking for heterogeneous effects within prior ability subgroups accords with the notion that teachers may set (track) students within (among) classes by ability. They may then target effort towards marginal students within these subgroups.

In the analysis below, I examine how the conditional distribution of test scores is affected by

[47] For example, Neal and Schanzenbach (2002) find that test scores improve for students in the middle of the prior ability distribution whilst low ability students experience zero or even negative effects on test scores.

treatment at each quantile $\tau \in [0,1]$ by estimating models of the following form:

$$Q_\tau(y_{is} \mid .) = \alpha_\tau + \delta_\tau D_s + X_{is}\beta_{1\tau} + W_s\beta_{2\tau}, \tag{7}$$

where $Q_\tau(. \mid .)$ is the $\tau^{\text{th}}$ conditional quantile function and $\delta_\tau$ is the quantile treatment effect (QTE) at quantile $\tau$. Figure 4 plots $\delta_\tau$ as well as the associated 95 per cent confidence interval, for the full sample of fail schools. Figures 5 and 6 plot the QTE within each prior ability quartile, for mathematics and English, respectively.

Panel A of Figure 4 shows that the effect of a fail inspection is to raise national standardized test scores by between 0.08 and 0.13 of a standard deviation for all quantiles.

Is there any evidence to suggest that teachers are acting strategically to raise performance of 'marginal' students? Recall from Table 3 that 67 per cent of students attain this target at fail schools in the year prior to the inspection. Thus, if teachers are strategically targeting the marginal student we would expect the treatment effect to peak at around quantile 0.33. This is not the case; in fact the treatment effects are relatively stable across most of the test score distribution. There is some evidence in Figure 4, Panel A that students at the highest ability levels gain less. Nevertheless, the gains even here are substantial and decline only modestly below 0.1 of a standard deviation. Thus, on this evidence there is little to suggest that teachers are acting strategically to raise performance of students on the margin of attaining the official government target.

An additional point to note here is that the pattern of treatment effects across quantiles reported in Figure 4, Panel A strongly rejects the notion that ceiling effects bite. If this was the case then high scoring students would not post gains from treatment. In fact the figure shows that even at high quantiles, treatment effects remain large.

Panel B of Figure 4 shows results for English test scores. As with the case of mathematics, there is no strong evidence to suggest that teachers are targeting the 'marginal' students. However, for English there is stronger evidence of lower gains for students at higher quantiles: at around 0.05 of a standard deviation for the 0.9 quantile, this effect is half that for quantiles below 0.7.

I turn now to the analysis of QTE within each prior ability quartile, reported in Figures 5 and 6. These are revealing. First, the OLS results reported in each panel of the two figures confirm the results in Table 6: the largest effects are for students in the bottom quartile of the prior ability distribution, and the smallest effects are for those in the top quartile. Second, and more importantly, within quartiles there is evidence of a great deal of heterogeneity, especially for students in quartile one. For students in the bottom prior ability quartile, the treatment effect for mathematics rises steadily from around 0.1 of a standard deviation for the lowest quantiles to just below 0.3 for the highest quantiles (Figure 5, Panel A). For English, Panel A of Figure 6 shows that the treatment effect is around 0.1 of a standard deviation for students below the median of the test score distribution and close to 0.2 for students at or above the median.

One explanation for the pattern of results reported in Panel A of Figures 5 and 6 is that teachers

target the students on the margin of attaining the Level 4 performance threshold.[48] However, the evidence from the remaining three panels (prior ability quartiles 2, 3 and 4) in each of Figures 5 and 6 does not support this view. For example, for the second quartile prior ability subgroup the evidence in Table 3 indicates test gains should peak around quantile 0.4 for mathematics and English. Panel B of Figure 5 shows some support for this but the English results in Panel B, Figure 6 show no evidence of such behavior. Similarly, for students in the third prior ability quartile the descriptive statistics in Table 3 indicate that if teachers are behaving strategically then test performance gains should peak around quantile 0.1 or 0.2 for mathematics and English and decline thereafter. The evidence in Panel C in each of Figures 5 and 6 shows no such pattern.

On balance, the results from the full sample quantile treatment effects as well as quantile treatment effects within prior ability subgroups tend to reject the view that teachers target students on the margin of attaining the official 'Level 4' threshold. What might then explain the strong rise in gains from treatment across quantiles for students in the lowest prior ability quartile (panel A in each of Figure 5 and Figure 6)? Discussion of this question is postponed until after the following subgroup analysis of treatment heterogeneity.

*Further Subgroup Analysis*

Table 7 reports results from separate regressions for subgroups determined by free lunch status and whether English is the first language spoken at home. The results by free lunch status suggest modestly higher gains in mathematics for free lunch students but smaller gains for this group relative to no-free lunch students in English. However, there are large differences in gains for students according to whether or not their first language is English. For mathematics, students whose first language is not English record gains of 0.19 of a standard deviation, compared to 0.12 of standard deviation for those whose first language is English. Similarly, gains on the English test are 0.12 of a s.d. (though only marginally significant) for the first group of students and 0.08 of a s.d. for the latter group.

*Discussion of Heterogenous Treatment Effects Results*

The analysis above points to strong gains on the age 11 (Key Stage 2) test for students classed as low ability on the prior (age seven) test. On the basis of the evidence presented above, two potential explanations for this finding can be rejected. First, these gains for low ability students do not appear to be a result of teachers strategically allocating effort among students: there is only weak support for the hypothesis that teachers target students on the margin of attaining the official performance threshold. Second, it also seems unlikely that ceiling effects for high ability students account for this result. So what then explains the gains for low ability students reported

[48] As indicated in Table 2, these students are likely to be in the higher quantiles of the test score distribution: Table 2 shows that in the year before the fail inspection 23 per cent and 33 per cent of students reach the mathematics and English threshold, respectively. Thus, if teachers successfully target the marginal students, we would expect to see the largest gains at quantiles 0.77 (mathematics) and 0.67 (English).

in Table 6 and the shape of the quantile treatment effects in Panel A, Figure 5 and Panel A, Figure 6?

One explanation that fits the facts is the argument that there may be a great deal of heterogeneity within the same school and even the same classroom in the degree to which parents are able to hold teachers to account. Parents of children scoring low on the age seven test are likely poorer than average and less able to assess their child's progress and the quality of instruction provided by the school. Teachers may therefore exert lower levels of effort for students whose parents are less vocal about quality of instruction. Following a fail inspection and the subsequent increased oversight of schools, teachers raise effort. This rise in effort may be greatest where previously there was the greatest slack. Thus lower ability students, whose parents face the highest costs in terms of assessing teaching quality, may gain the most from a fail inspection. This would then help explain the strong rise for low ability students, as reported in Table 6.

Furthermore, if students in the low prior ability group do indeed receive greater attention from teachers following a fail inspection, the expectation may be that within this group, students with higher innate ability benefit the most. This would accord with the usual assumption that investment and student ability are complementary in the test score production function. This is exactly in line with the results of Panel A, Figure 5 and Panel A, Figure 6, which show rising treatment effects across quantiles for students in the lowest prior ability quartile.

The above interpretation of the results is also supported by the subgroup analysis of Table 7, which shows that children from poorer, minority groups tend to gain relatively more from the fail inspection. Children from families where English is not the first language at home most likely have parents who are less able to interrogate teachers and hold them accountable. The results in Table 7 boost the conclusion that it is children from these sorts of families who are helped most by the fail inspection.

## 5.3 Evidence on Medium-Term Effects

The results reported above show that a fail inspection leads to test score gains for Year 6 students, who are in the last year of primary school. One question is whether these gains are sustained following the move to secondary school. This analysis provides an indirect assessment of whether the initial test score gains at the primary school are due to 'teaching to the test' rather than a result of greater mastery or deeper understanding of the material being examined. In the former case, any gains would be expected to dissipate quickly. Note that such fadeout of initial gains is in fact common in settings even where educators are not under pressure to artificially distort measured student performance (see for example Currie and Thomas, 1995). Thus, the fading of test score gains does not necessarily indicate distortionary response on the part of teachers. On the other hand, if some of the initial test score gains persist in to the medium term then this would suggest that the initial gains from the fail treatment are 'real'.

Table 8 reports results for the Key Stage 3 test score outcome for Year 9 students (age 14),

i.e. three years after leaving the fail primary school. This exercise is limited by the fact that these tests are teacher assessments (and not externally marked, as is the case for Key Stage 2 tests used in the primary analysis above) and are currently only available for students at primary schools failed in 2006 (Year 9 test taken in 2009) and 2007 (Year 9 test taken in 2010). This leads to a reduced sample size, relative to the earlier analysis of Key Stage 2 test outcomes. In order to reduce noise, mathematics and English test scores are combined into a single measure by taking the mean for the two tests for each student.

The results in column 1 of Table 8 suggest that the mean effect of treatment three years after leaving the fail primary school is a gain in test score of 0.05 of a standard deviation (statistically significant at the 10 per cent level). Analysis of heterogeneity in treatment impact suggests that the medium-term gains are largest for lower ability students (columns 2 and 3), in line with earlier results showing large gains for these groups in the year of inspection. Finally, subgroup analysis reported in columns 4 through 7 suggests that fadeout is stronger for poorer students and students whose first language is not English.

Overall, the analysis of test scores three years after the treatment show that the positive effects are not as large as the immediate impacts, suggesting fadeout is an important factor. Nevertheless, the evidence shows that some of the gains do persist in to the medium term.

# 6 Conclusion

How best to design incentives for public organizations such as schools is a fundamental public policy issue. One solution, performance evaluation on the basis of test scores, is prevalent in many countries. This paper evaluates an alternative approach, school inspections, which may better capture the multifaceted nature of education production.

The first set of results in this study demonstrate that inspector ratings are correlated with underlying measures of school quality - constructed using survey measures from the school's current stock of students and parents - even after conditioning on standard observed school characteristics such as test rankings and the proportion of students receiving a free lunch. The evidence suggests that inspectors are able to discriminate between more and less effective schools, and, significantly, report on their findings in a high stakes setting.[49] Thus, this aspect of school inspections - simply disseminating inspection ratings and reports - may help relax information constraints facing consumers and other decision makers. Such constraints are thought to be pervasive in the public sector.[50]

The main body of this study is concerned with evaluating the causal effect of a fail inspection on test scores. Employing a novel strategy to address the classic mean reversion problem, the

[49] In an evaluation of *school principals'* subjective assessment of teacher effectiveness, Jacob and Lefgren (2008) find that principals are able to identify teachers who produce the largest and smallest standardized achievement gains. There are no stakes associated with the principal's assessment.

[50] For example, on the effects of relaxing information constraints on families' school choices, see Hastings and Weinstein (2008).

basic finding is that a fail inspection leads to test score improvements of around 0.1 of a standard deviation. These results are robust to different methods of estimation: simple comparisons of post-treatment outcomes for the control and treatment groups as well as difference-in-differences models yield very similar results.

Furthermore, there is little evidence to suggest that schools are able to inflate test performance by gaming the system. First, teachers do not appear to exclude low ability students from the test-taking pool. Second, the evidence does not support the notion that teachers target students on the margin of attaining the official proficiency level at the expense of students far above or below this threshold. Third, although test gains fade somewhat over time, there is evidence to suggest that for some students gains last into the medium term, even after they have left the fail school. This suggests that teachers inculcate real learning and not just test-taking skills in response to the fail rating. These negative findings on strategic behavior are in stark contrast to a significant body of evidence demonstrating dysfunctional responses to test-based performance evaluation in other settings. My interpretation of these results is that by subjecting schools to close scrutiny, inspectors may play an important role in limiting such distortionary activities.

Finally, examining treatment heterogeneity reveals that the largest gains are for students scoring low on the prior (age seven) test. Within this group, students in the middle and upper quantiles gain the most, around 0.3 of a standard deviation for mathematics and 0.2 of a standard deviation for English. These effects are large when compared to other possible policy interventions, such as the effects of higher quality teachers (Rivkin et al, 2005); attending a school with higher attainment levels (Hastings et al, 2009); or enrolling in a charter school (Abdulkadiroglu et al, 2011). These results are consistent with the view that children of low income parents - arguably, the least vocal in holding teachers to account - benefit the most from inspections. Consequently, the findings of this study may be especially relevant in the current policy environment where, first, there is heightened concern about raising standards for this group of children and, second, they are hard to reach using other policy levers.[51]

These findings are noteworthy given the prior empirical evidence suggesting that subjective assessments may give rise to various kinds of biases. For example, Prendergast (1999) notes that subjective evaluations of workers may lead to 'leniency' and 'centrality' bias in private sector firms. Evidence from the public sector points to bureaucrats indulging their preferences when allowed to exercise discretion rather than following formal rules (Heckman et al, 1996). Although such biases in inspector behavior cannot be ruled out, this study demonstrates that the inspection system appears to be effective along the following two dimensions: first, inspectors produce ratings which are 'valid' and, second, they are able to identify poorly performing schools, leading to test score gains. One important difference between the bureaucrats in charge of school inspections and

[51] For example, a 'hard to reach' group may be students whose parents choose *not* to participate in a voucher program or apply for a spot in a charter school. On the other hand, the most credible evidence in the current literature tends to focus on students whose parents *are* active in such programs, often making use of lottery assignment to, for example, determine the effects of attending a higher performing school. Examples include Hastings et al (2009) and Abdulkadiroglu et al (2011).

those charged with allocating training in the Heckman et al (1996) study is that the key inspector output - an inspection rating and report - is available on the Internet for public consumption. Consequently, inspector decisions themselves may be subject to scrutiny and oversight. One hypothesis for future research is that this is a key element in driving the positive results found in this study.

**References**


Abdulkadiroglu, A., J. Angrist, S. Dynarski, T. Kane, and P. Pathak (2011) "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots", *The Quarterly Journal of Economics*, 126 (2), pp. 699-748.

Ashenfelter, Orley (1978), "Estimating the Effect of Training Programs on Earnings", *The Review of Economics and Statistics*, 60(1), pp. 47-57.

Besley, Timothy, Gwyn Bevan and Konrad Burchardi (2008), "Accountability and incentives: The impacts of different regimes on hospital waiting times in England and Wales", mimeo, London School of Economics.

Björkman, Martina and Jakob Svensson (2009). "Power to the People: Evidence from a Randomized Field Experiment of Community-Based Monitoring in Uganda", *Quarterly Journal of Economics*, 124(2).

Blundell, Richard and Monica Costa Dias (2010), "Alternative approaches to evaluation in empirical microeconomics", *Journal of Human Resources*, Vol. 44 (3), pp. 565-640.

Burgess, Simon, Carol Propper, Helen Slater and Deborah Wilson (2005), "Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools," CMPO working paper (July 2005).

Chay, Kenneth Y, Patrick J. McEwan and Miguel Urquiola (2005), "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools", *The American Economic Review*, Vol. 95, No. 4 (Sep., 2005), pp. 1237-1258.

Chiang, Hanley (2009), "How accountability pressure on failing schools affects student achievement", *Journal of Public Economics*, 93(9-10), pp.1045-1057.

Clark, Damon (2009), "The Performance and Competitive Effects of School Autonomy", *Journal of Political Economy*, 117 (4), pp. 745-783.

Courty, Pascal and Gerald Marschke (2011), "Measuring Government Performance: An Overview of Dysfunctional Responses", in Heckamn et al (eds.) (2011), pp. 203-229.

Cullen, J.B. & Reback, R. (2006), "Tinkering toward accolades: School gaming under a performance accountability system" In T. Gronberg & D. Jansen (Eds), *Advances in Applied Microeconomics*, 14.

Currie, Janet and Duncan Thomas (1995), "Does Head Start Make a Difference?", *The American Economic Review*, 85 (3), pp. 341-364.

Dixit, Avinash (2002), "Incentives and Organizations in the Public Sector: An Interpretative Review." J. Human Resources 37:696–727.

Figlio, David (2007) "Boys Named Sue: Disruptive Children and Their Peers", *Education Finance and Policy*, 2 (4), pp. 376-394.

Figlio, David, "Testing, crime and punishment", *Journal of Public Economics*, Volume 90, Issues 4-5, May 2006, Pages 837-851.

Figlio, D. and L Getzler (2006), "Accountability, ability, and disability: Gaming the system?" In T. Gronberg & D. Jansen (Eds), *Advances in Applied Microeconomics*, 14.

Figlio, David and Susanna Loeb (2011), "School Accountability" in Handbook of the Economics of Education, edited by Hanushek, Eric & Machin, Stephen & Woessmann, Ludger, Volume 3, 2011, Pages 383-421.

Frey, Bruno S. 1993, "Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty", *Economic Inquiry*, 31(4), pp. 663-70.

Grubb, N (2000), "Opening Classrooms and Improving Teaching: Lessons from School Inspections in England", *Teachers College Record*, 102(4), pp. 696–723.

Haskins, Ron and W. Steven Barnett (2010), "Finally, the Obama Administration Is Putting Head Start to the Test", *The Washington Post*, October 11:
http://www.brookings.edu/opinions/2010/1011_head_start_haskins.aspx

Hastings, Justine S. and Jeffrey M. Weinstein (2008), "Information, School Choice, and Academic Achievement: Evidence from Two Experiments", The Quarterly Journal of Economics, 123(4), pp 1373-1414.

Heckman, James (2000), "Policies to foster human capital", *Research in Economics*, vol. 54(1), pp. 3–56.

Heckman, J, C. Heinrich, P. Courty, G. Marschke and J Smith (eds.) (2011), *The Performance of Performance Standards*, Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Heckman, James J.; Lalonde, Robert J. and Smith, Jeffrey A. (1999), "The Economics and Econometrics of Active Labor Market Programs", in Orley Ashenfelter and David Card, eds.,

*Handbook of Labor Economics*, Vol. 3A. Amsterdam: Elsevier Science, North-Holland, pp. 1865-2097.

Heckman, James J.; Smith, Jeffrey A. and Taber, Christopher (1996), "What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance Into the JTPA Program", in G. Libecap, ed., Advances in the study of entrepreneurship, innovation and growth. Vol. 7. Greenwich, CT: JAI Press, pp. 191-217.

Holmstrom, Bengt, and Paul Milgrom (1991), "Multitask Principal-Agent Analysis: Incentive Contracts Asset Ownership, and Job Design", Journal of Law, Economics, and Organization 7(Special Issue): 24-52.

Hoxby, C (2000) "Peer Effects in the Classroom: Learning from Gender and Race Variation", National Bureau of Economic Research working paper 7866.

Hussain, Iftikhar (2009), "Essays in Household Economics and Economics of Education", PhD Thesis, University College London.

Jacob, B (2005), "Accountability, Incentives and Behavior: Evidence from School Reform in Chicago," Journal of Public Economics. 89(5-6): 761-796.

Jacob, Brian A. and Lars Lefgren (2008), "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education", *Journal of Labor Economics*, 26(1), pp. 101-136.

Jacob, Brian A. and Steven D. Levitt (2003), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating", *The Quarterly Journal of Economics*, 118(3), pp. 843-877.

Johnson, Paul (2004), "Education Policy in England", *Oxford Review of Economic Policy* , 20 (2), pp. 173-197.

Kane, Thomas J. and Staiger, Douglas O. (2002), "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives*, 2002a, 16(4), pp. 91-114.

Koretz, Daniel (2002), "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity", *The Journal of Human Resources*, 37 (4), pp. 752-777.

Lang, Kevin (2010), "Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member", *Journal of Economic Perspectives*, V.24(3), pp.167-82.

Lazear, E (2001) "Educational Production", *The Quarterly Journal of Economics*, 116 (3), pp. 777-803.

Lazear, Edward P. and Paul Oyer (forthcoming), "Personnel Economics", Robert Gibbons and John Roberts (eds.), *Handbook of Organizational Economics*, Princeton University Press.

Machin, Stephen and Anna Vignoles (2005), *What's the Good of Education? The Economics of Education in the UK*, Princeton University Press.

Matthews, P, J Holmes, P Vickers and B Corporaal (1998) "Aspects of the Reliability and Validity of School Inspection Judgements of Teaching Quality", *Educational Research and Evaluation*, 4(2), pp. 167-188.

Matthews, P and P. Sammons (2004), *Improvement Through Inspection: An Evaluation of the Impact of Ofsted's Work*, London: Ofsted/Institute of Education.

Milgrom, Paul and John Roberts (1988), "An Economic Approach to Influence Activities in Organizations", *The American Journal of Sociology*, V.94, pp. S154-S179.

Neal, Derek (2010), "Aiming for Efficiency Rather Than Proficiency", *Journal of Economic Perspectives*, V.24(3), pp. 119-32.

Neal, Derek and Diane Whitmore Schanzenbach (2010) "Left Behind by Design: Proficiency Counts and Test-Based Accountability", Review of Economics and Statistics May 2010, Vol. 92, No. 2: 263–283.

Ofsted (2011a), *The Framework For School Inspection*, September, document reference number 090019.

Ofsted (2011b), "*Conducting School Inspections: Guidance For Inspecting Schools*", September, document reference number 090097.

Olken, Benjamin (2007), "Monitoring Corruption: Evidence from a Field Experiment in Indonesia", *Journal of Political Economy*, 115 (2), pp. 200-249.

Prendergast, Canice (1999), "The Provision of Incentives in Firms", *Journal of Economic Literature*, 37(1), pp. 7-63.

Prendergast, Canice, Robert H. Topel (1996), "Favoritism in Organizations", *The Journal of Political Economy*, V.104(5), pp. 958-978.

Propper, Carol, Sutton, Matt, Whitnall, Carolyn and Windmeijer, Frank (2008) "Did 'Targets and Terror' Reduce Waiting Times in England for Hospital Care?," The B.E. Journal of Economic Analysis & Policy: Vol. 8: Iss. 2 (Contributions), Article 5.

Reback, Randall (2008), "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," Journal of Public Economics, 92:5–6 (2008), 1394–1415.

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

Fig.1:**Test Score Performance, Fail and Non-Fail Schools, 2006 Inspections**

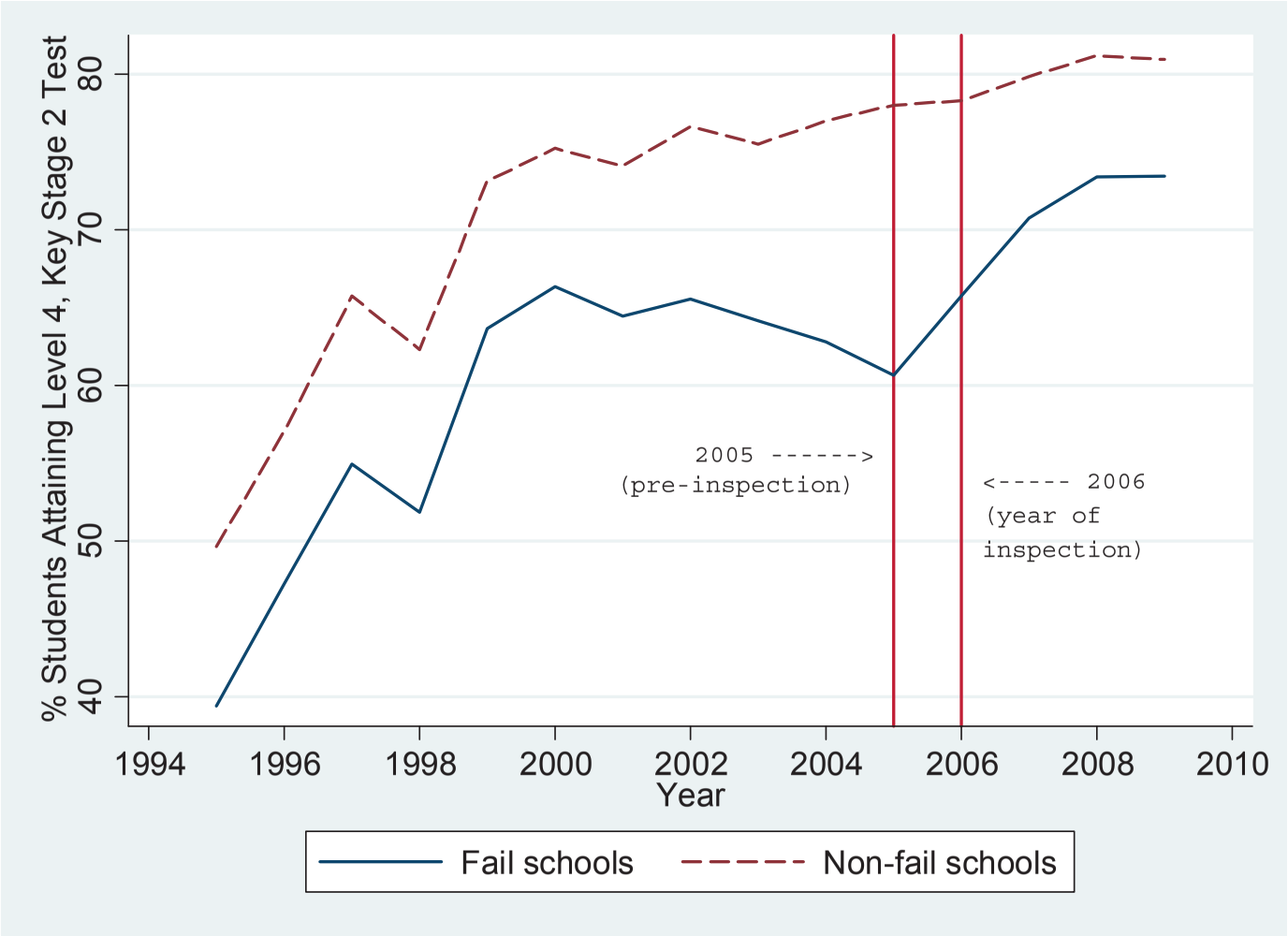
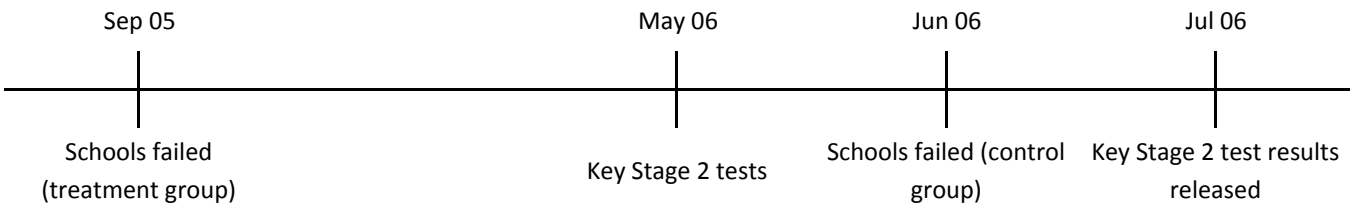


**Figure 2: Time line showing treatment and control schools for academic year 2005/06**




Note: This time line depicts two groups of schools: those failed in September and those failed in June. Also shown is the month (May) in which the national age-11 Key Satge 2 tests are taken and the month when the results are leased (July).

**Figure 3: Stylized Example of Distribution of Students Passing Proficiency Thresholds, by Quartile of Prior Ability**

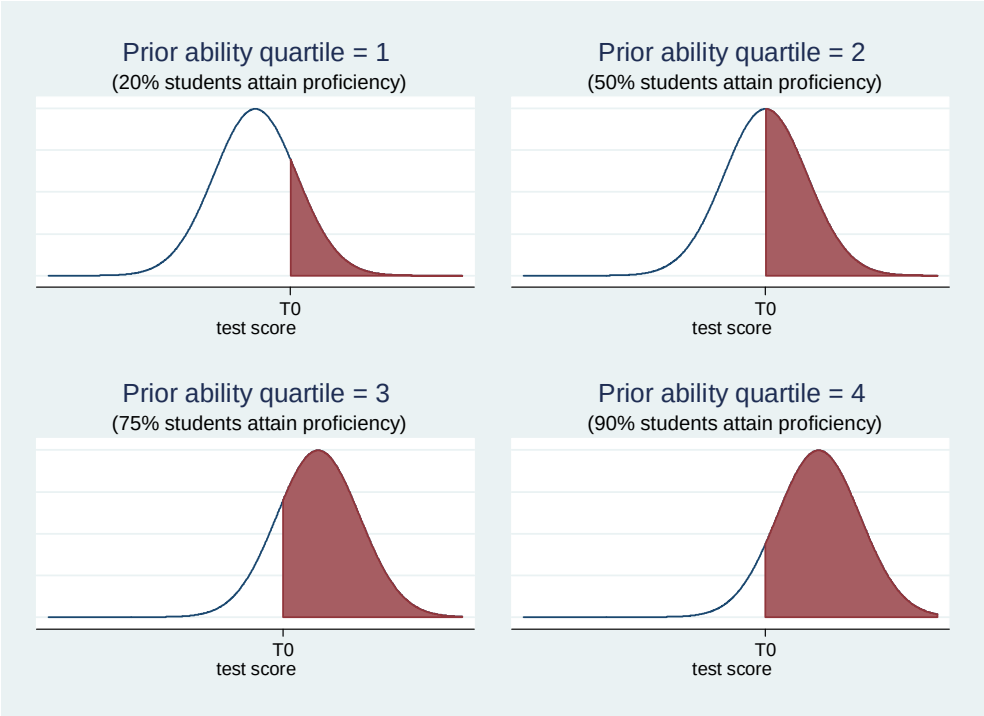
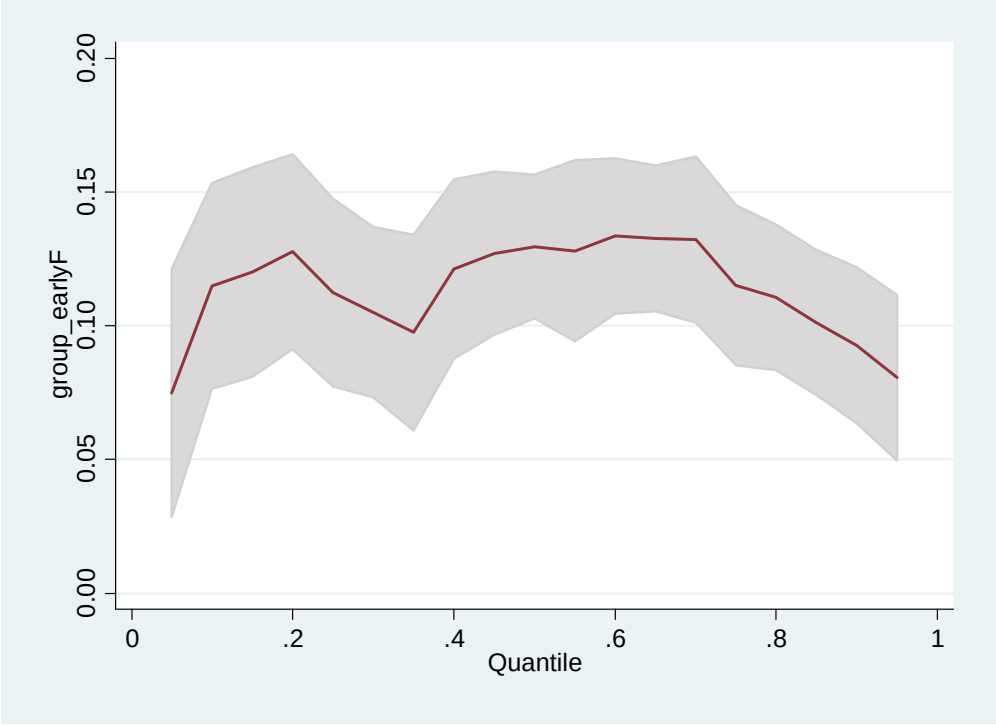


**Figure 4: Quantile Regression Estimates of the Effect of a Fail Inspection**

Outcome variable: age 11 (Key Stage 2) national standardised test score

**Panel A: Mathematics**




**Panel B: English**

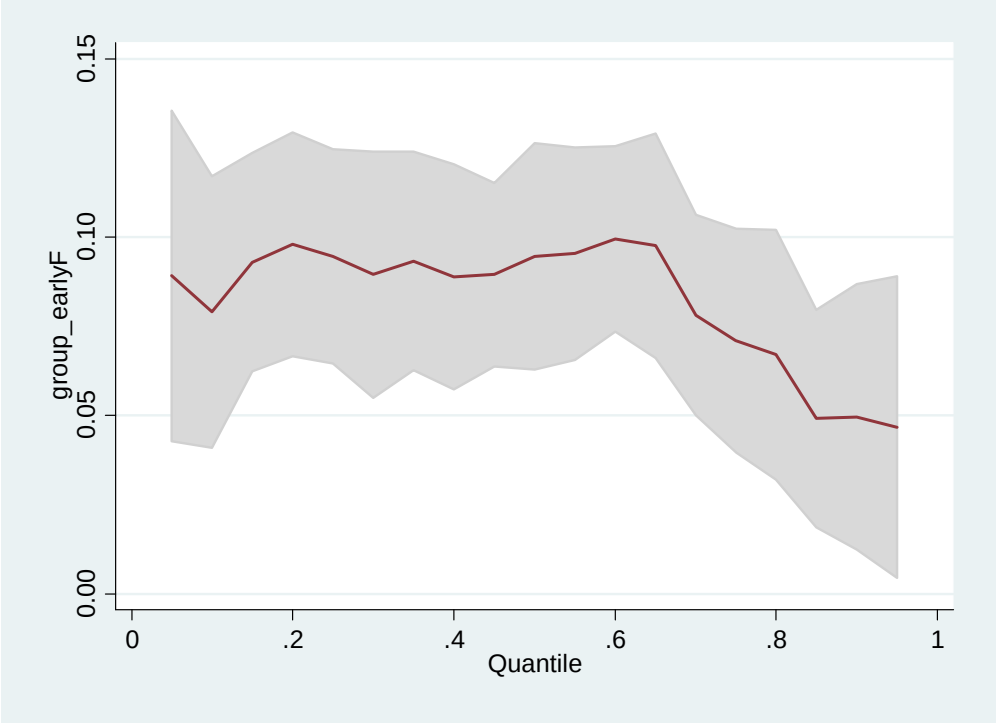
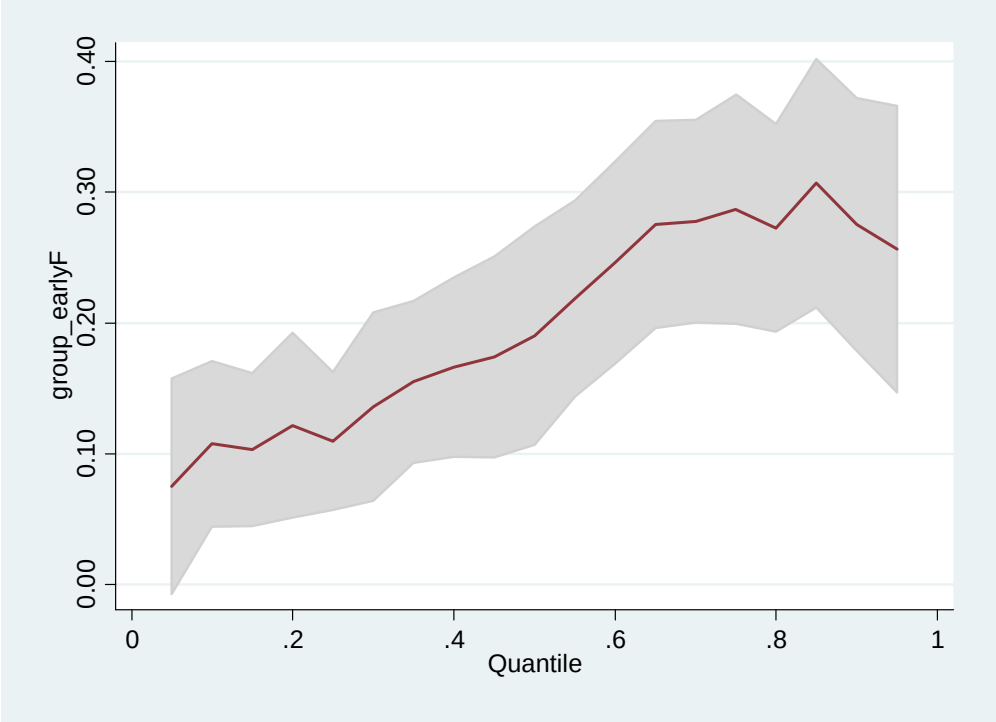


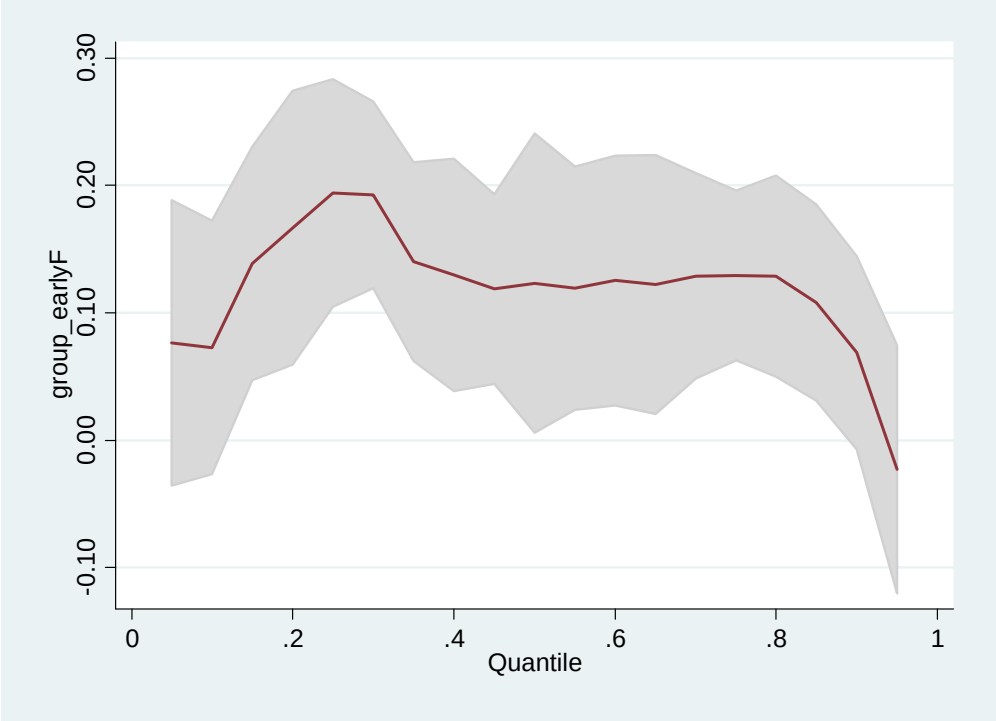
**Figure 5: Quantile Regression Estimates: by Prior Ability Quartile (Mathematics)**

Outcome variable: age 11 (Key Stage 2) national standardised test score

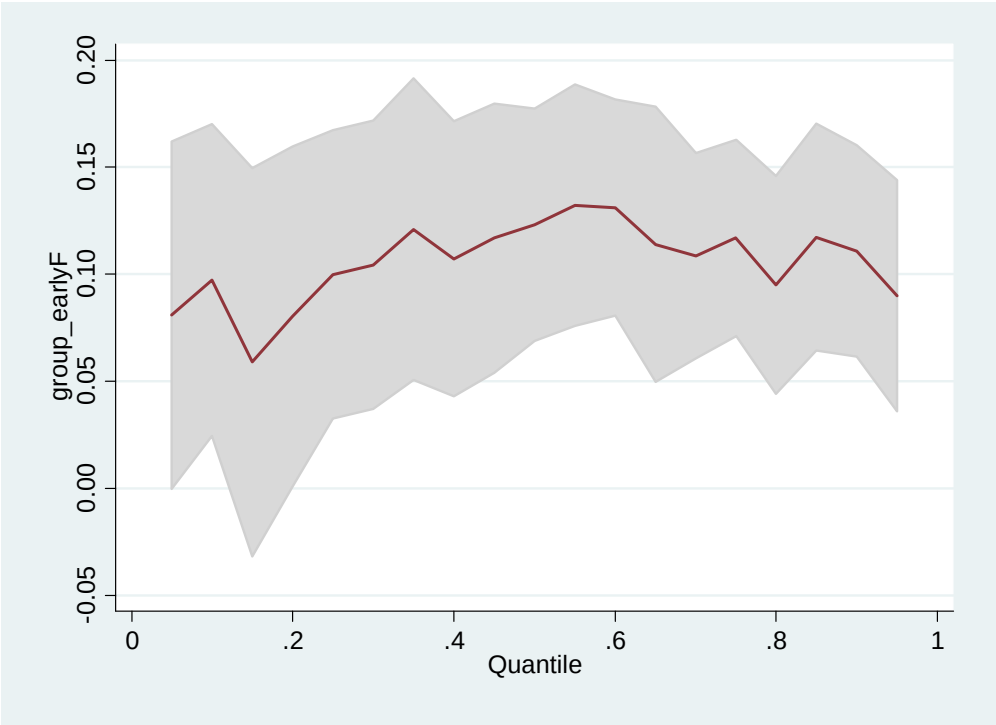
**Panel A: Prior ability (age 7 test) quartile = 1**




**Panel B: Prior ability (age 7 test) quartile = 2**



**Panel C: Prior ability (age 7 test) quartile = 3**




**Panel D: Prior ability (age 7 test) quartile = 4**

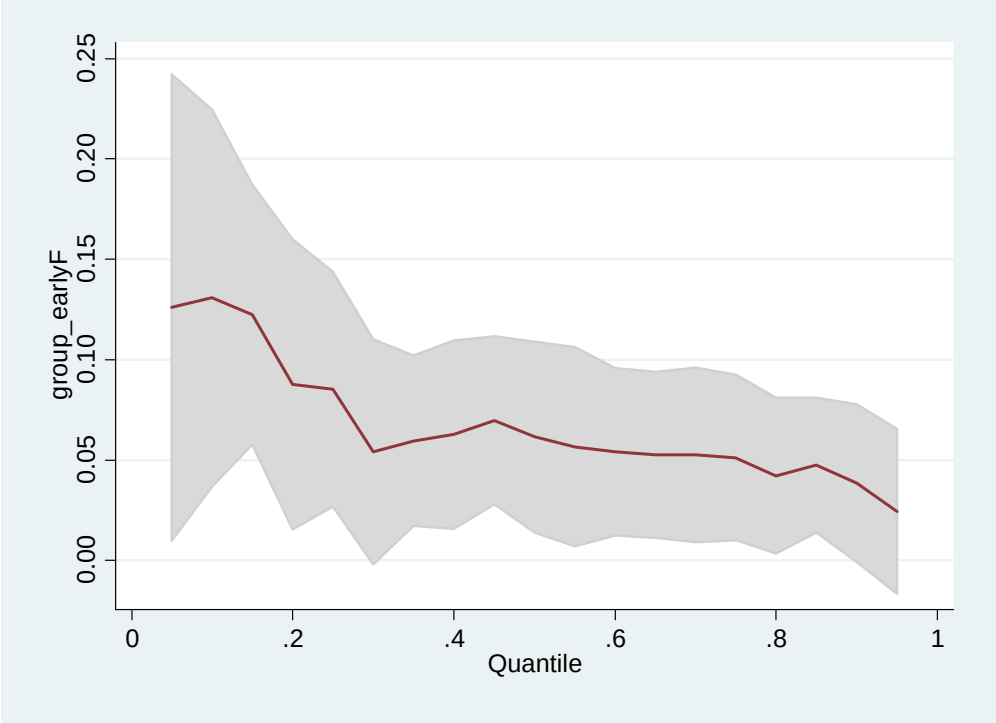
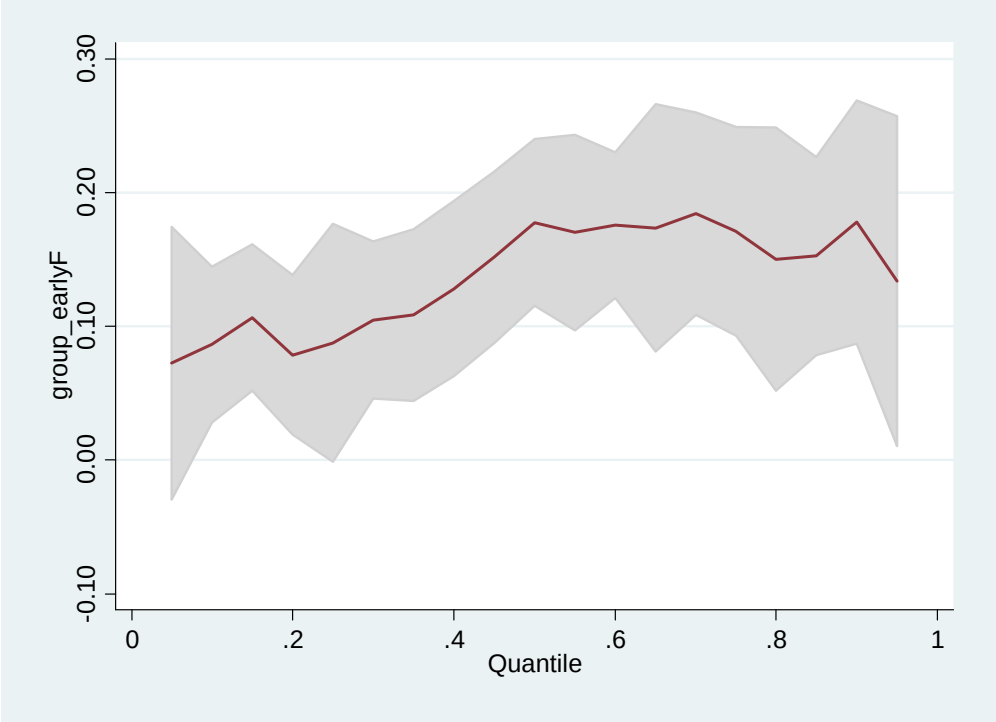


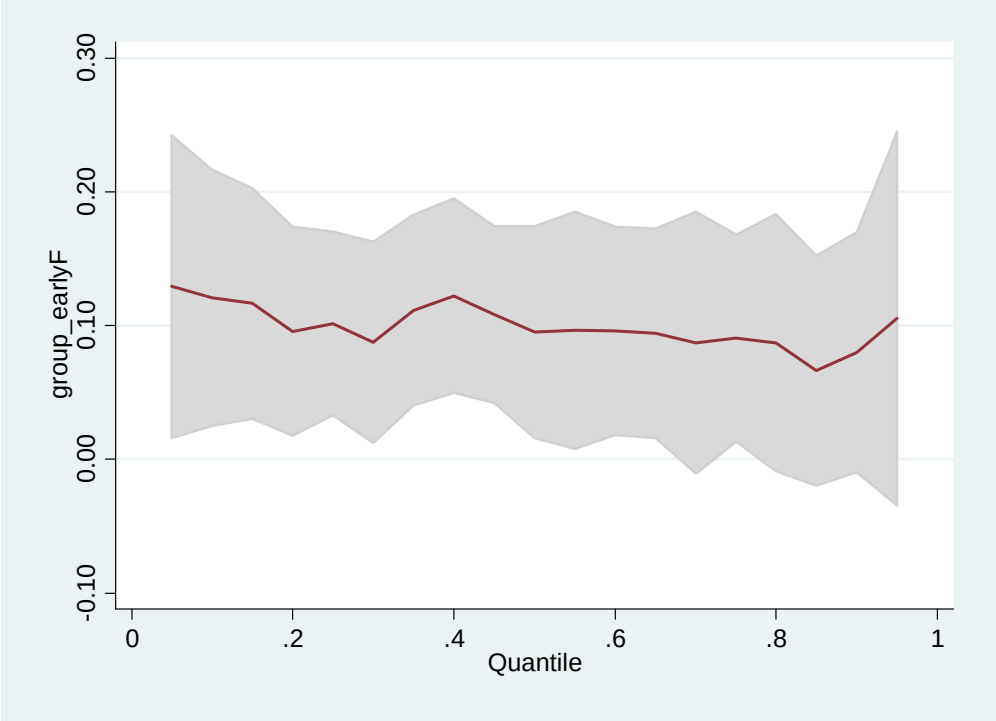
**Figure 6: Quantile Regression Estimates: by Prior Ability Quartile (English)**

Outcome variable: age 11 (Key Stage 2) national standardised test score

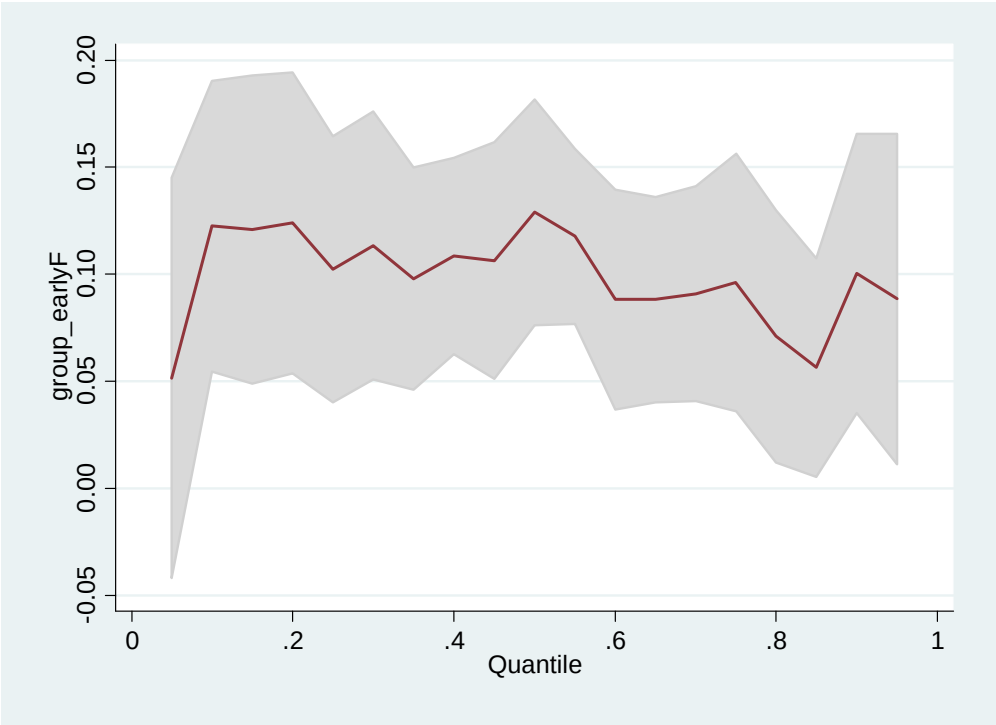
**Panel A: Prior ability (age 7 test) quartile = 1**




**Panel B: Prior ability (age 7 test) quartile = 2**



**Panel C: Prior ability (age 7 test) quartile = 3**




**Panel D: Prior ability (age 7 test) quartile = 4**

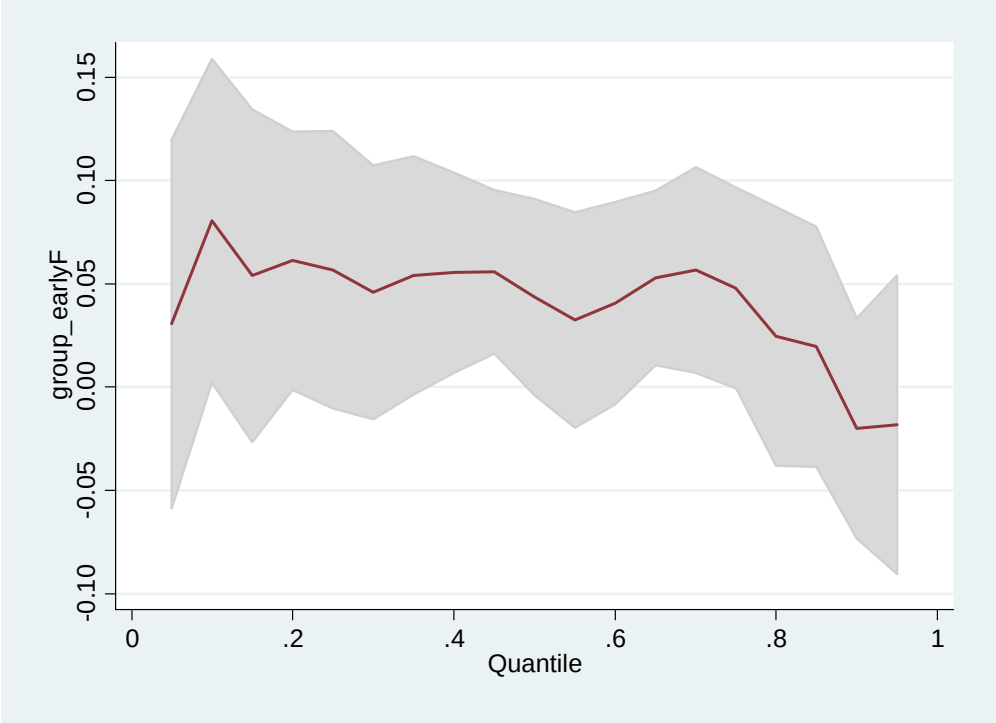


**Table 1: The relationship between inspection ratings and student and parent evaluations of school quality**

| **Panel A , Outcome: teacher practices and behaviour z-score** | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Inspection rating (range: 1-4) | | | -0.2190** | -0.1341** | -0.1257** | -0.1033** | |
| | | | (0.0191) | (0.0238) | (0.0228) | (0.0224) | |
| Test score percentile rank | 0.0064** | | | 0.0050** | 0.0038** | 0.0029** | 0.0030** |
| | (0.0006) | | | (0.0009) | (0.0009) | (0.0009) | (0.0009) |
| Fraction of students eligible for free lunch | | -0.2077 | | 0.3234 | -0.0491 | -0.0735 | -0.0770 |
| | | (0.1212) | | (0.1843) | (0.1912) | (0.1859) | (0.1869) |
| Insp. rating = 2 (Good) | | | | | | | -0.1337** |
| | | | | | | | (0.0419) |
| Insp. rating = 3 (Satisfactory) | | | | | | | -0.2349** |
| | | | | | | | (0.0498) |
| Insp. rating = 4 (Fail) | | | | | | | -0.2958** |
| | | | | | | | (0.0785) |
| Additional school controls and Local Authority fixed effects | No | No | No | Yes | Yes | Yes | Yes |
| Respondent background controls | No | No | No | No | Yes | Yes | Yes |
| Previous inspection rating | No | No | No | No | No | Yes | Yes |
| Observations | 10012 | 10012 | 10012 | 10012 | 10012 | 10012 | 10012 |
| R-squared | 0.032 | 0.001 | 0.035 | 0.031 | 0.096 | 0.100 | 0.100 |
| **Panel B, Outcome: school discipline z-score** | | | | | | | |
| Inspection rating (range: 1-4) | | | -0.2052** | -0.1146** | -0.1155** | -0.1060** | |
| | | | (0.0214) | (0.0200) | (0.0210) | (0.0215) | |
| Test score percentile rank | 0.0074** | | | 0.0049** | 0.0043** | 0.0038** | 0.0039** |
| | (0.0006) | | | (0.0008) | (0.0008) | (0.0008) | (0.0009) |
| Fraction of students eligible for free lunch | | -0.4790** | | -0.3143 | -0.3982* | -0.4137* | -0.4178* |
| | | (0.1203) | | (0.1742) | (0.1938) | (0.1916) | (0.1911) |
| Insp. rating = 2 (Good) | | | | | | | -0.1484** |
| | | | | | | | (0.0371) |
| Insp. rating = 3 (Satisfactory) | | | | | | | -0.2230** |
| | | | | | | | (0.0474) |
| Insp. rating = 4 (Fail) | | | | | | | -0.3366** |
| | | | | | | | (0.0882) |
| Additional school controls and Local Authority fixed effects | No | No | No | Yes | Yes | Yes | Yes |
| Respondent background controls | No | No | No | No | Yes | Yes | Yes |
| Previous inspection rating | No | No | No | No | No | Yes | Yes |
| Observations | 10011 | 10011 | 10011 | 10011 | 10011 | 10011 | 10011 |
| R-squared | 0.042 | 0.006 | 0.030 | 0.034 | 0.082 | 0.082 | 0.083 |

(continued on next page)

**Table 1 (cont.)**

| **Panel C, Outcome: parental satisfaction z-score** | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Inspection rating (range: 1-4) | | | -0.1913** | -0.1184** | -0.1143** | -0.0849** | |
| | | | (0.0200) | (0.0225) | (0.0215) | (0.0214) | |
| Test score percentile rank | 0.0054** | | | 0.0033** | 0.0021* | 0.0008 | 0.0009 |
| | (0.0006) | | | (0.0009) | (0.0010) | (0.0011) | (0.0011) |
| Fraction of students eligible for free lunch | | -0.1333 | | 0.0387 | -0.2621 | -0.2999 | -0.3018 |
| | | (0.1222) | | (0.1888) | (0.2110) | (0.2050) | (0.2059) |
| Insp. rating = 2 (Good) | | | | | | | -0.1045* |
| | | | | | | | (0.0480) |
| Insp. rating = 3 (Satisfactory) | | | | | | | -0.1798** |
| | | | | | | | (0.0584) |
| Insp. rating = 4 (Fail) | | | | | | | -0.2566** |
| | | | | | | | (0.0693) |
| Additional school controls and Local Authority fixed effects | No | No | No | Yes | Yes | Yes | Yes |
| Respondent background controls | No | No | No | No | Yes | Yes | Yes |
| Previous inspection rating | No | No | No | No | No | Yes | Yes |
| Observations | 9841 | 9841 | 9841 | 9841 | 9841 | 9841 | 9841 |
| R-squared | 0.023 | 0.000 | 0.027 | 0.023 | 0.095 | 0.100 | 0.100 |

Notes: The dependent variable for the regressions are z-scores computed from the student-level mean across six questions relating to teacher practices (Panel A); three questions relating to class and school discipline (Panel B); and four questions relating parental satisfaction (Panel C). See Appendix Table A1 for details of the survey questions. The inspection rating corresponds to the first inspection after the year of the survey (2003/04). Inspection ratings prior to the survey are included as additional controls in columns (6) and (7). The school's national test score percentile rank is calculated using the proportion of students achieving five A*-C grades on the age 16 GCSE exams in 2004. The proportion of students eligible for a free school meal (lunch) is also from 2004. Additional school controls (columns 4-7): dummies for type of school (Community, Voluntary Controlled, Voluntary Aided, Foundation) and log of total enrolment. Respondents' background controls in columns (5) to (7) include the following student and family characteristics: student's prior test score (Key Stage 2 test, taken at age 11 in primary school), whether has an illness or disability; parents' education, income, employment history and whether in receipt of various government benefits; whether a single parent household; and number of children in the household. Dummies included for any missing respondent background controls. ** and * denote significance at 1% and 5% level respectively. Standard errors are reported in brackets; clustered at the school-level (columns 1 - 3) or local authority-level (columns 4 - 7).

**Table 2: School Characteristics Prior to Fail Inspection, Treatment and Control Schools**

| | Schools Failed in 2005/06 | | | Schools Failed in 2006/07 | | | Schools Failed in 2007/08 | | | Schools Failed in 2008/09 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Late inspected schools (control) | Early inspected schools (treated) | t-test of difference (p-value) | Late inspected schools (control) | Early inspected schools (treated) | t-test of difference (p-value) | Late inspected schools (control) | Early inspected schools (treated) | t-test of difference (p-value) | Late inspected schools (control) | Early inspected schools (treated) | t-test of difference (p-value) |
| Month of inspection | 6.17 | 10.19 | 0.000** | 6.07 | 10.22 | 0.000** | 6.16 | 10.24 | 0.000** | 6.23 | 10.14 | 0.000** |
| | (0.10) | (0.09) | | (0.09) | (0.09) | | (0.10) | (0.10) | | (0.11) | (0.13) | |
| Year of previous inspection | 2000.6 | 2000.1 | 0.004** | 2002.2 | 2001.5 | 0.001** | 2004.1 | 2003.4 | 0.001** | 2006.0 | 2005.0 | 0.001** |
| | (0.12) | (0.11) | | (0.14) | (0.13) | | (0.13) | (0.13) | | (0.05) | (0.23) | |
| % students entitled to free school meal | 26.8 | 23.5 | 0.336 | 22.8 | 25.2 | 0.409 | 25.6 | 24.6 | 0.807 | 21.6 | 19.8 | 0.665 |
| | (2.99) | (1.92) | | (2.45) | (1.67) | | (3.36) | (2.06) | | (3.34) | (2.29) | |
| % students white British | 78.2 | 82.4 | 0.338 | 78.1 | 77.4 | 0.881 | 75.5 | 83.3 | 0.183 | 68.9 | 75.3 | 0.420 |
| | (4.23) | (2.26) | | (4.34) | (3.00) | | (5.56) | (3.06) | | (7.09) | (4.43) | |
| Previous inspection rating (Outstanding = 1; Good = 2; Satisfactory = 3) | 2.20 | 2.33 | 0.271 | 2.33 | 2.46 | 0.302 | 2.42 | 2.33 | 0.513 | 2.82 | 2.38 | 0.004** |
| | (0.10) | (0.07) | | (0.09) | (0.08) | | (0.11) | (0.08) | | (0.08) | (0.10) | |
| Age 11 standardised test scores, year before Fail | | | | | | | | | | | | |
| Mathematics | -0.43 | -0.39 | 0.667 | -0.40 | -0.43 | 0.636 | -0.47 | -0.49 | 0.785 | -0.36 | -0.45 | 0.354 |
| | (0.05) | (0.04) | | (0.05) | (0.04) | | (0.05) | (0.04) | | (0.09) | (0.05) | |
| English | -0.42 | -0.40 | 0.827 | -0.42 | -0.48 | 0.297 | -0.51 | -0.49 | 0.756 | -0.36 | -0.37 | 0.954 |
| | (0.06) | (0.04) | | (0.05) | (0.04) | | (0.06) | (0.05) | | (0.10) | (0.06) | |
| Number of schools | 41 | 83 | | 42 | 81 | | 31 | 59 | | 22 | 35 | |

Notes: Standard errors in brackets. Schools failed for the first time in the academic year indicated. 'Early inspected' schools are those failed in the early part of the academic year (September to November). 'Late inspected schools' are those inspected after the national age 11 (Key Satge 2) exam in the second week of May (i.e. schools inspected between May 18th and mid-July of the year of the fail inspection). Mathematics and English standardised test scores are from the academic year immediately preceding the inspection year.

**Table 3: Proportion of Students Attaining the Official Target, by Prior Ability**

| | Mathematics | English |
|---|---|---|
| Prior ability quartile: | | |
| 1 | 0.23 | 0.33 |
| | (0.42) | (0.46) |
| 2 | 0.58 | 0.60 |
| | (0.49) | (0.48) |
| 3 | 0.82 | 0.87 |
| | (0.38) | (0.33) |
| 4 | 0.96 | 0.98 |
| | (0.20) | (0.13) |
| All students | 0.67 | 0.72 |
| | (0.47) | (0.45) |
| Total number of students | 14,805 | 14,853 |

Notes: This table shows the proportion of students attaining the government attainment target - Level 4 - for Year 6 students on the Key Stage 2 test. Prior ability is measured by Year 2 (age 7) Mathematics and Writing test scores. The sample consists of all students in the year before the fail inspection at schools failed between 2006 and 2009. Students with missing age seven test scores are dropped, and so the total sample size is slightly smaller than in Table 2 (Table 2 includes missing dummies for these students in the regression analysis). Standard deviations in brackets.

**Table 4: OLS and DID Estimates of the Effect of a Fail Inspection on Test Scores**

(Outcome variable: age 11 (Key Stage 2) national standardised test score)

| **Panel A: OLS** | **Mathematics** | | | **English** | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| early Fail | 0.107** | 0.114** | 0.121** | 0.071 | 0.079** | 0.083** |
| | (0.036) | (0.030) | (0.028) | (0.038) | (0.030) | (0.030) |
| Student characteristics | No | Yes | Yes | No | Yes | Yes |
| Age 7 test scores | No | No | Yes | No | No | Yes |
| R-squared | 0.002 | 0.257 | 0.486 | 0.001 | 0.316 | 0.529 |
| Observations | 16617 | 16617 | 16617 | 16502 | 16502 | 16502 |
| Number of schools | 394 | 394 | 394 | 394 | 394 | 394 |
| **Panel B: Difference-in-differences** | | | | | | |
| | **Mathematics** | | | **English** | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| post x early Fail | 0.117** | 0.113** | 0.117** | 0.080* | 0.075* | 0.072* |
| | (0.032) | (0.032) | (0.030) | (0.038) | (0.036) | (0.036) |
| post | 0.014 | 0.038 | 0.028 | 0.061 | 0.085** | 0.078** |
| | (0.025) | (0.025) | (0.024) | (0.031) | (0.029) | (0.030) |
| Student characteristics | No | Yes | Yes | No | Yes | Yes |
| Age 7 test scores | No | No | Yes | No | No | Yes |
| R-squared | 0.003 | 0.243 | 0.494 | 0.004 | 0.307 | 0.543 |
| Observations | 33730 | 33730 | 33730 | 33386 | 33386 | 33386 |
| Number of schools | 394 | 394 | 394 | 394 | 394 | 394 |

Notes: Standard errors reported in brackets; * and ** indicate significance at the 5% and 1% levels, respectively. S.e.'s clustered at the school level. Each column reports results from difference-in-differences or OLS models estimated for schools rated Fail in the years 2006 to 2009. The dummy 'early Fail' is switched on for schools failed in the early part of the academic year (September to November) and switched off for schools failed after the Key Stage 2 test taken in early May (i.e. schools failed mid-May to mid-July). The dummy 'post' is turned on for the year of inspection and off for the previous year. Controls for student characteristics include dummies for: female; eligibility for free lunch; special education needs; month of birth; first language is English; twenty ethnic groups; and census information on local neighborhood deprivation (IDACI score). Controls for age 7 (Key Stage 1) test scores are dummy variables indicating one of four attainment levels. Missing dummies included for student characteristics and age 7 test scores. All DID regressions in Panel B include school fixed effects.

**Table 5: The Effect of a Fail Inspection on Test Scores in the Pre-Treatment Year (Falsification Test)**

(Data pooled over four inspection years)

| **Panel A: OLS** | **Mathematics** | | | **English** | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| early Fail | -0.018 | -0.006 | -0.001 | -0.016 | 0.000 | 0.007 |
| | (0.035) | (0.027) | (0.025) | (0.038) | (0.030) | (0.029) |
| Student characteristics | No | Yes | Yes | No | Yes | Yes |
| Age 7 test scores | No | No | Yes | No | No | Yes |
| R-squared | 0.000 | 0.252 | 0.501 | 0.000 | 0.316 | 0.549 |
| Observations | 17113 | 17113 | 17113 | 16884 | 16884 | 16884 |
| Number of schools | 394 | 394 | 394 | 394 | 394 | 394 |
| **Panel B: Difference-in-differences** | | | | | | |
| | **Mathematics** | | | **English** | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| post x early Fail | -0.000 | 0.002 | 0.003 | -0.012 | 0.004 | 0.008 |
| | (0.029) | (0.029) | (0.026) | (0.038) | (0.037) | (0.035) |
| post | -0.049* | -0.026 | -0.033 | -0.101** | -0.072* | -0.081** |
| | (0.022) | (0.023) | (0.022) | (0.031) | (0.031) | (0.030) |
| Student characteristics | No | Yes | Yes | No | Yes | Yes |
| Age 7 test scores | No | No | Yes | No | No | Yes |
| R-squared | 0.001 | 0.232 | 0.495 | 0.003 | 0.300 | 0.546 |
| Observations | 34838 | 34838 | 34838 | 34390 | 34390 | 34390 |
| Number of schools | 394 | 394 | 394 | 394 | 394 | 394 |

Notes: Standard errors reported in brackets; * and ** indicate significance at the 5% and 1% levels, respectively. S.e.'s clustered at the school level. Outcome variable: age 11 (Key Stage 2) national standardised test scores *in the year before inspection* . Each column reports results from difference-in-differences or OLS models estimated for schools rated Fail in the years 2006 to 2009. The dummy 'early Fail' is switched on for schools failed in the early part of the academic year (September to November) and switched off for schools failed after the Key Stage 2 test taken in early May (i.e. schools failed mid-May to mid-July). The dummy 'post' is turned on in the *year before inspection* and off two years before the inspection. Controls for age 7 (Key Stage 1) test scores are dummy variables indicating one of four attainment levels. All DID regressions in Panel B include school fixed effects.

**Table 6: Ability Interactions**

| | Mathematics | | English | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Linear | Non-linear | Linear | Non-linear |
| early Fail | 0.185** | 0.190** | 0.138** | 0.131** |
| | (0.044) | (0.042) | (0.044) | (0.042) |
| early Fail x prior ability percentile rank | -0.00191** | | -0.00146 | |
| | (0.00069) | | (0.00075) | |
| early Fail x prior ability quartile=2 | | -0.075 | | -0.040 |
| | | (0.045) | | (0.046) |
| early Fail x prior ability quartile=3 | | -0.093* | | -0.050 |
| | | (0.042) | | (0.043) |
| early Fail x prior ability quartile=4 | | -0.144** | | -0.105* |
| | | (0.047) | | (0.051) |
| Full set of controls | Yes | Yes | Yes | Yes |
| Observations | 14387 | 14387 | 14429 | 14429 |
| R-squared | 0.576 | 0.576 | 0.541 | 0.541 |

Notes: Standard errors reported in brackets; * and ** indicate significance at the 5% and 1% levels, respectively. Standard errors clustered at the school level. Outcome variable: age 11 (Key Stage 2) national standardised test scores. Prior ability percentile rank calculated using age 7 (Key Stage 1) mathematics and writing tests. The dummy 'early Fail' is switched on for schools failed in the early part of the academic year (September to November) and switched off for schools failed late (mid-May to mid-July). All regressions include full set of controls for student characteristics, i.e. dummies for: female; eligibility for free lunch; special education needs; month of birth; first language is English; twenty ethnic groups; and census information on local neighborhood deprivation (IDACI score). Missing dummies included for these student characteristics. Students with missing age 7 test scores dropped from the sample. Regressions in columns (1) and (3) also include prior ability percentile rank as control; regressions in columns (2) and (4) include prior ability quartile.

**Table 7: Subgroup Estimates of the Effect of a Fail Inspection on Test Scores**

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Full sample | Free lunch = 0 | Free lunch = 1 | First language English | First language NOT English |
| **Panel A: Mathematics** | | | | | |
| early Fail | 0.121** | 0.114** | 0.136** | 0.115** | 0.187** |
| | (0.028) | (0.030) | (0.039) | (0.029) | (0.066) |
| Mean standardized test score | -0.41 | -0.30 | -0.79 | -0.39 | -0.53 |
| Observations | 16617 | 12852 | 3705 | 14289 | 2268 |
| Number of schools | 394 | 394 | 384 | 392 | 296 |
| R-squared | 0.486 | 0.476 | 0.454 | 0.505 | 0.390 |
| | Full sample | Free lunch = 0 | Free lunch = 1 | First language English | First language NOT English |
| **Panel B: English** | | | | | |
| early Fail | 0.083** | 0.088** | 0.057 | 0.081** | 0.120 |
| | (0.030) | (0.030) | (0.043) | (0.030) | (0.074) |
| Mean standardized test score | -0.42 | -0.30 | -0.85 | -0.39 | -0.58 |
| Observations | 16502 | 12818 | 3628 | 14230 | 2216 |
| Number of schools | 394 | 394 | 384 | 392 | 294 |
| R-squared | 0.530 | 0.520 | 0.489 | 0.553 | 0.398 |

Notes: Standard errors reported in brackets; * and ** indicate significance at the 5% and 1% levels, respectively. S.e.'s clustered at the school level. Outcome variable: age 11 (Key Stage 2) national standardised test scores. Each column reports results from an OLS model estimated for schools rated Fail in the years 2006 to 2009. See footnote to Table 4 for details of student-level controls. Mean standardized score reported in the second row of each panel is from the year before inspection.

**Table 8: Medium-Term Effects**

Outcome: National standardised score on age 14 teacher assessments of mathematics and English attainment (combined)

| | Basic | Ability interactions | | Subgroup analysis | | | |
|---|---|---|---|---|---|---|---|
| | | | | Free lunch = 0 | Free lunch = 1 | First language English | First language NOT English |
| | | Linear | Non-linear | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| early Fail | 0.048+ | 0.056 | 0.069* | 0.053+ | 0.017 | 0.051+ | 0.060 |
| | (0.029) | (0.039) | (0.036) | (0.030) | (0.045) | (0.030) | (0.069) |
| early Fail x prior ability percentile rank | | -0.001 | | | | | |
| | | (0.001) | | | | | |
| early Fail x prior ability quartile=2 | | | -0.069 | | | | |
| | | | (0.043) | | | | |
| early Fail x prior ability quartile=3 | | | -0.048 | | | | |
| | | | (0.043) | | | | |
| early Fail x prior ability quartile=4 | | | -0.095* | | | | |
| | | | (0.046) | | | | |
| Full set of controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 10047 | 8948 | 8948 | 7685 | 2324 | 8594 | 1415 |
| R-squared | 0.344 | 0.538 | 0.539 | 0.303 | 0.330 | 0.369 | 0.248 |

Notes: Standard errors reported in brackets; + , * and ** indicate significance at the 10%, 5% and 1% levels, respectively. S.e.'s clustered at the school level. Mathematics and English attainment measured three years after leaving the (failed) primary school. Thus, age 14 (Key Stage 3) mathematics and English attainment is derived from secondary school teacher assessments in 2008/09 (for students who attended primary schools failed in 2005/06) and 2009/10 (students who attended primary schools failed in 2006/07). Controls for student characteristics include dummies for: female; eligibility for free lunch; special education needs; month of birth; first language is English; twenty ethnic groups; and census information on local neighborhood deprivation (IDACI score). Controls for age 7 (Key Stage 1) test scores are dummy variables indicating one of four attainment levels. Missing dummies included for student characteristics and age 7 test scores. For columns (2) and (3) students with missing age 7 test scores are dropped from the sample.

**Appendix Table A1: OLS and Difference-in-Differences Estimates of the Effect of a Fail Inspection on Mathematics Test Scores**

| **Panel A: OLS** | **2006 Fail** | | | **2007 Fail** | | | **2008 Fail** | | | **2009 Fail** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
| early Fail | 0.184* | 0.141** | 0.135** | 0.119* | 0.145** | 0.129** | -0.018 | -0.001 | 0.059 | 0.117 | 0.113 | 0.129 |
| | (0.075) | (0.053) | (0.051) | (0.056) | (0.052) | (0.047) | (0.069) | (0.058) | (0.053) | (0.086) | (0.079) | (0.075) |
| Student characteristics | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |
| Age 7 test scores | No | No | Yes | No | No | Yes | No | No | Yes | No | No | Yes |
| R-squared | 0.007 | 0.283 | 0.502 | 0.003 | 0.261 | 0.487 | 0.000 | 0.253 | 0.508 | 0.003 | 0.248 | 0.460 |
| Observations | 5117 | 5117 | 5117 | 5185 | 5185 | 5185 | 3851 | 3851 | 3851 | 2464 | 2464 | 2464 |
| Number of schools | 124 | 124 | 124 | 123 | 123 | 123 | 90 | 90 | 90 | 57 | 57 | 57 |

| **Panel B: Difference-in-differences** | **2006 Fail** | | | **2007 Fail** | | | **2008 Fail** | | | **2009 Fail** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
| post x early Fail | 0.111 | 0.131* | 0.101 | 0.168** | 0.169** | 0.150** | 0.010 | -0.020 | 0.068 | 0.186* | 0.158 | 0.146 |
| | (0.059) | (0.059) | (0.056) | (0.057) | (0.054) | (0.052) | (0.063) | (0.065) | (0.060) | (0.091) | (0.085) | (0.083) |
| post | 0.031 | 0.029 | 0.024 | -0.033 | -0.011 | 0.022 | 0.091* | 0.130** | 0.038 | -0.043 | 0.007 | 0.030 |
| | (0.049) | (0.049) | (0.047) | (0.046) | (0.043) | (0.042) | (0.044) | (0.047) | (0.045) | (0.065) | (0.056) | (0.058) |
| Student characteristics | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |
| Age 7 test scores | No | No | Yes | No | No | Yes | No | No | Yes | No | No | Yes |
| R-squared | 0.003 | 0.253 | 0.500 | 0.003 | 0.247 | 0.490 | 0.002 | 0.236 | 0.504 | 0.003 | 0.248 | 0.496 |
| Observations | 10532 | 10532 | 10532 | 10490 | 10490 | 10490 | 7657 | 7657 | 7657 | 5051 | 5051 | 5051 |
| Number of schools | 124 | 124 | 124 | 123 | 123 | 123 | 90 | 90 | 90 | 57 | 57 | 57 |

Notes: Standard errors reported in brackets; * and ** indicate significance at the 5% and 1% levels, respectively. S.e.'s clustered at the school level. Outcome variable: age 11 (Key Stage 2) national standardised test scores. '2006' refers to the academic year 2005/06 and so on for the other years. The dummy 'early Fail' is switched on for schools failed in the early part of the academic year (September to November) and switched off for schools failed after the Key Stage 2 test taken in early May (i.e. schools failed mid-May to mid-July). The dummy 'post' is turned on for the year of inspection and off for the previous year. Controls for student characteristics include dummies for: female; eligibility for free lunch; special education needs; month of birth; first language is English; twenty ethnic groups; and census information on local neighborhood deprivation (IDACI score). Controls for age 7 (Key Stage 1) test scores are dummy variables indicating one of four attainment levels. Missing dummies included for student characteristics and age 7 test scores. All the DID regressions (Panel B) include school fixed effects.

**Appendix Table A2: OLS and Difference-in-Differences Estimates of the Effect of a Fail Inspection on English Test Scores**

| **Panel A: OLS** | **2006 Fail** | | | **2007 Fail** | | | **2008 Fail** | | | **2009 Fail** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
| early Fail | 0.079 | 0.048 | 0.039 | 0.074 | 0.093 | 0.070 | -0.008 | 0.022 | 0.074 | 0.165 | 0.164* | 0.181* |
| | (0.071) | (0.055) | (0.052) | (0.069) | (0.056) | (0.052) | (0.069) | (0.057) | (0.060) | (0.087) | (0.071) | (0.076) |
| Student characteristics | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |
| Age 7 test scores | No | No | Yes | No | No | Yes | No | No | Yes | No | No | Yes |
| R-squared | 0.001 | 0.345 | 0.561 | 0.001 | 0.310 | 0.535 | 0.000 | 0.316 | 0.532 | 0.006 | 0.321 | 0.501 |
| Observations | 5153 | 5153 | 5153 | 5112 | 5112 | 5112 | 3795 | 3795 | 3795 | 2442 | 2442 | 2442 |
| Number of schools | 124 | 124 | 124 | 123 | 123 | 123 | 90 | 90 | 90 | 57 | 57 | 57 |

**Panel B: Difference-in-differences**

| | **2006 Fail** | | | **2007 Fail** | | | **2008 Fail** | | | **2009 Fail** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
| post x early Fail | 0.007 | 0.028 | -0.002 | 0.162* | 0.160* | 0.136* | -0.002 | -0.020 | 0.056 | 0.168 | 0.123 | 0.108 |
| | (0.069) | (0.064) | (0.063) | (0.070) | (0.066) | (0.067) | (0.065) | (0.070) | (0.068) | (0.094) | (0.089) | (0.098) |
| post | 0.140* | 0.136* | 0.137* | 0.012 | 0.031 | 0.057 | 0.095* | 0.131* | 0.056 | -0.047 | 0.019 | 0.039 |
| | (0.062) | (0.055) | (0.055) | (0.058) | (0.054) | (0.057) | (0.046) | (0.051) | (0.053) | (0.072) | (0.062) | (0.073) |
| Student characteristics | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |
| Age 7 test scores | No | No | Yes | No | No | Yes | No | No | Yes | No | No | Yes |
| R-squared | 0.005 | 0.327 | 0.564 | 0.005 | 0.305 | 0.541 | 0.002 | 0.291 | 0.540 | 0.003 | 0.321 | 0.532 |
| Observations | 10537 | 10537 | 10537 | 10316 | 10316 | 10316 | 7545 | 7545 | 7545 | 4988 | 4988 | 4988 |
| Number of schools | 124 | 124 | 124 | 123 | 123 | 123 | 90 | 90 | 90 | 57 | 57 | 57 |

Notes: See notes to Table A1