CEE DP 135

Subjective Performance Evaluation in the Public Sector: Evidence from School Inspections

Iftikhar Hussain

February 2012

Published by
Centre for the Economics of Education
London School of Economics
Houghton Street
London WC2A 2AE

Subjective Performance Evaluation in the Public Sector: Evidence from School Inspections

Iftikhar Hussain

Acknowledgments

Iftikhar Hussain is Research Associate at the Centre for Economic Performance, LSE, and a lecturer in the Department of Economics, University of Sussex.

The author would like to thank Orazio Attanasio, Oriana Bandiera, Tim Besley Steve Bond, Martin Browning, Ian Crawford, Avinash Dixit, Sergio Firpo, Caroline Hoxby, Andrea Ichino, Ian Jewitt, Kevin Lang, Valentino Larcinese, Clare Leaver, Steve Machin, Meg Meyer, Imran Rasul, David Ulph and John Van Reenen for their comments as well as seminar participants at Nottingham University, Oxford, Sussex and the Second Workshop On the Economics of Education, Barcelona.

1 Introduction

In an effort to make public organizations more efficient, governments around the world make use of 'hard' performance targets. Examples include student test scores for the schooling sectors in the US, England and Chile (see the survey by Figlio and Loeb, 2011) and patient waiting times in the English public health care system (Propper et al, 2010). 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 multifaceted outcome.[1]

Subjective evaluation, on the other hand, holds out the promise of 'measuring what matters'. 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. 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 empirically evaluates a subjective performance evaluation regime for schools. The setting is the English public (state) schooling system, where independent inspectors visit schools with a maximum of two days notice, assess the school's performance and disclose their findings on the internet. 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. Furthermore, schools rated 'fail' may be subject to sanctions, such as more frequent and intensive inspections.

Inspection systems exist in many countries around the world and in-class evaluations by external assessors have been proposed recently in the US for the K-12 sector as well as the pre-school Head Start program.[3] Yet, very little hard empirical evidence exists on the effects of inspection systems (see, for example, the review by Faubert, 2009). An evaluation of the English inspection regime, which is well-embedded and has been in place since the early 1990s, may provide valuable lessons for other countries.

I provide evidence on the effectiveness of inspections along the following dimensions. First, do inspector ratings provide any extra information on school quality, over and above that already available in the public sphere? Second, I examine the effects of a fail inspection on student test

[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." Whether complementing objective performance evaluation with subjective assessment is an effective strategy in such settings remains an open question.

[3] Haskins and Barnett ("Finally, the Obama Administration Is Putting Head Start to the Test", *The Washington Post*, October 11, 2010) note that the US administration has proposed a review of all Head Start centers, a central component of which may be in-class evaluation of the quality of teacher support and instruction by external assessors.

scores. If a school fails its inspection, the stakes - certainly for the school principal - are high.[4] Consequently the incentives to generate improvements in school performance are strong. Thus it seems plausible that the fail treatment will affect both the mean and the distribution of student test scores.

However, empirically 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 (Ashenfelter, 1978). As explained below, assignment to treatment, fail, is at least partly based on past realizations of the outcome variable, test scores. Any credible strategy must overcome the concern that poor performance prior to a fail inspection is simply due to bad luck and that test scores would have risen even in the absence of the inspection. Figure 1 illustrates the problem in the current setting.

Third, I investigate 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 established the empirical importance of such phenomena in the context of schools.[5] I test to what extent such behavior can be detected in the current context.

The findings of this study are as follows. In order to address whether inspection ratings provide any additional information on school quality, I ask whether these ratings are correlated with teenage student survey reports of teacher practices, conditional on standard observable school characteristics.[6] The results from this 'validity test' demonstrate that 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 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 survey measures of parent satisfaction.

Turning to the effect of a fail rating on test scores, I exploit a design feature of the English testing system to assess the causal effect of a fail inspection. As explained below, tests for age-11 students in England are administered in the second week of May in each year.[7] These tests are marked externally, and results are released to schools and parents in mid-July. The short window

[4] Hussain (2009) analyzes the effects of inspections on the teacher labor market. The evidence shows that when schools receive a 'severe fail' inspection rating, there is a significant rise in the probability that the school principal exits the teaching labour market. The next section provides further details.

[5] 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. Strategies include excluding low ability students from the test taking pool and targeting students on the margin of passing proficiency thresholds, see, for example, Figlio 2006, Jacob 2005 and Neal and Schanzenbach 2010.

[6] 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.

[7] Official tests are administered twice to students in the primary schooling phase, at ages seven and 11.

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.[8] 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.[9]

Results using student-level data from a panel of schools show that students at early failed schools (the treatment group) gain around 0.1 of a standard deviation on age-11 national standardized mathematics and English tests relative to students enrolled in late fail schools (the control group).

This overall finding masks substantial heterogeneity in treatment effects. In particular, the largest gains are for students scoring low on the prior (age seven) test; these gains cannot be explained by ceiling effects for the higher ability students.[10] For mathematics, students in the bottom quartile of the age-seven test score distribution gain 0.2 of a standard deviation on the age-eleven test, whilst for students in the top quartile the gain is 0.05 of a standard deviation.

The empirical strategy allows for tests of gaming behavior on a number of key margins. The results show that teachers do not exclude low ability students from the test-taking pool. Next, the evidence tends to reject the hypothesis that teachers target students on the margin of attaining the official proficiency level at the expense of students far above or below this threshold. Finally, there is evidence to suggest that for some students, gains last into the medium term, even after they have left the fail school. These findings are consistent with the notion that teachers inculcate real learning and not just test-taking skills in response to the fail rating.

Having ruled out important types of gaming behavior, I provide tentative evidence on what might be driving the main results. First, I differentiate between the two sub-categories of the fail rating - termed 'moderate' and 'severe' fail. As explained below, the former category involves increased oversight by the inspectors but does not entail other dramatic changes in inputs or school principal and teacher turnover. The results show that even for this category of moderate fail schools there are substantial gains in test scores following a fail inspection. Second, employing a survey of teachers, I provide evidence that a fail inspection does not lead to higher turnover (at the classroom teacher level), relative to a set of control schools. However, teachers at fail schools do appear to respond by improving classroom discipline. Piecing this evidence together suggests that greater effort by the current stock of classroom teachers at fail schools is an important mechanism behind the test score gains reported above.

The main contribution of this study is to offer new evidence on the effectiveness of subjective performance evaluation in the public sector. It sheds light on the usefulness of top-down moni-

[8] 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.

[9] 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.

[10] Quantile regression analysis reveals substantial gains across all quantiles of the test score distribution.

toring for schools by external assessors who exercise discretion in forming their judgements. In particular, the evidence suggests that inspectors can identify schools which improve under pressure of a fail rating. The finding that such intervention by bureaucrats, or so-called 'experts', is especially helpful in improving outcomes for students from poorer households conforms with emerging evidence suggesting that such families may face especially severe information constraints.[11]

This study also contributes to a small but growing literature investigating the validity of novel measures of school and teacher effectiveness. These include school principals' subjective assessment of teacher effectiveness (Jacob and Lefgren, 2008) as well as in-class evaluations of teacher performance by mentors and peers external to the school (Rockoff and Speroni, 2010, Kane et al, 2010, and Taylor and Tyler, 2011). As far as I am aware, this is the first study investigating the validity of inspector ratings.

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 adopted 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 Background

2.1 Institutional Context

The English public schooling system combines centralized testing with school inspections. Tests take place at ages 7, 11, 14 and 16; these are known as the Key Stage 1 to Key Stage 4 tests, respectively[12]. Successive governments have used these Key Stage tests, especially Key Stages 2 and 4, as pivotal measures of performance in holding schools to account. For further details see, for example, Machin and Vignoles (2005).

Since the early 1990s all English public schools have been inspected by a government agency called the Office for Standards in Education, or Ofsted. As noted by Johnson (2004), Ofsted has three primary functions: (i) offer feedback 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'.[13] 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 organiza-

[11] Hastings and Weinstein (2008) provide evidence on the importance of information constraints for poor households when choosing among schools.

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

[13] 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).

tions via a competitive tendering process.[14] Setting overall strategic goals and objectives, putting in place an inspection framework which guides the process of inspection, as well as 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.[15] 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.[16] Inspectors arrive at the school at very short notice (maximum of two to three days), which should limit disruptive 'window dressing' in preparation for the inspections.[17] Importantly for the empirical strategy employed in the current study, inspections take place throughout the academic year, September to July.

During the on-site 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). 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 inspection report. The school is given an explicit headline grade, ranging between 1 ('Outstanding') and 4 ('Unsatisfactory'; also known as a fail rating). The report is made available to students and parents and is posted on the internet.[18]

There are two categories of fail, a moderate fail (known as 'Notice to Improve') and a more severe fail category ('Special Measures'). Sanctions for these two treatments vary as follows. For the moderate fail category, schools are subject to additional inspections, with an implicit threat of downgrade to the severe fail category if inspectors judge improvements to be inadequate. Schools receiving the severe fail rating, however, may experience more dramatic intervention: these can include changes in the school leadership team and the school's governing board, increased resources, as well as increased oversight from the inspectors.[19]

Over the relevant period for this study (September 2006 to July 2009) 13 percent of schools received the best rating, 'Outstanding' (grade 1); 48 percent received a 'Good' (grade 2) rating; 33

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

[15] Inspection cycles typically lasted between three and six years. From September 2009 schools judged to be good or better are subject to fewer inspections than those judges to be mediocre or worse (Ofsted, 2011a).

[16] English primary schools cater for students between the age of 5 and 11.

[17] 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). Allen and Burgess (2012) provide suggestive evidence that students gain from a fail inspection.

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

[19] For evidence of the negligible effects of a moderate fail on principal turnover and the contrasting large effects for severe fail schools, see Hussain (2009).

percent received a 'Satisfactory' (grade 3) rating; and 6 percent received a 'Fail' (grade 4) rating. The Fail rating can be decomposed into 4.5 percent of schools receiving the moderate fail and 1.5 percent of schools receiving the severe fail rating.

Inspectors clearly place substantial weight on test scores: this is borne out by analysis of the data as well as official policy statements.[20] Regression analysis (not reported here in full for brevity) demonstrates a strong association between the likelihood of being failed and test scores. Conditional on the proportion of students eligible for free lunch and local authority fixed effects, a decline of ten national percentile points on a school's test performance in the year before inspection is associated with a 3 percentage point rise in the probability of being rated fail. Nevertheless, as the above discussion indicates, test scores are not the only signal used by inspectors to rate schools. This is demonstrated by the fact that around 25 percent of schools in the bottom quarter of the test score distribution were rated either Outstanding or Good during the 2006 to 2009 period.

2.2 Theoretical Background

In order to arrive at the overall school rating, inspectors must combine two signals of underlying school quality - the objective measure (the school's test rank) and the subjective assessment from the school visit. Suppose the principal (a policymaker acting on behalf of parents, say) can instruct the inspector how these two measures should be combined. Then the key question is what are the optimal weights to attach to each of these two measures.

Some guidance on this issue is provided by the personnel economics literature investigating optimal contract design for firms where workers are subjectively assessed by supervisors. These studies emphasize the potential for biased supervisor reports arising from, for example, influence activities (Milgrom and Roberts, 1988) and favoritism (Prendergast and Topel, 1996). Two key implications of this literature are as follows. First, supervisors may compress worker performance ratings. Compressed ratings can arise through leniency bias, where supervisors are reluctant to give bad ratings to underperforming workers, as well as centrality bias, where all or most workers are rated adequate or good according to prevailing social norms. Second, this literature notes that the principal will make use of bureaucratic rules, such as using seniority in determining promotions and job assignment (Prendergast, 1999). Consequently, incentives may be muted and talent may not be optimally allocated, but on the upside, such bureaucracy will limit rent seeking behavior and unproductive activities, such as time spent ingratiating with the supervisor.

In the current setting, where there is limited scope for long-term relations and repeated interaction between the assessor (inspector) and the worker (school principal and teachers), arguably mechanisms such as influence activities and favoritism are less important. Thus, ratings compres-

[20] The government views test scores as an 'anchor' in the English accountability system. See, for example, the statement by the secretary of state for education in oral evidence to the House of Commons education select committee (Uncorrected Transcript of Oral Evidence, 31 January 2012, http://www.publications.parliament.uk/pa/cm201012/cmselect/cmeduc/uc1786-i/uc178601.htm, accessed February 10, 2012.)

sion may not be as important an issue as it is in the private firm context. This is borne out by the spread of inspector ratings observed in the data (see discussion in the previous section).

However, a number of potential concerns remain with subjective evaluation even in this one-shot setting. First, there is the possibility that the inspector is misled into thinking the school is of higher quality than it really is if, as seems likely, teachers respond to the inspection visit by delivering lessons of a higher quality than is the norm.[21]

Perhaps even more importantly, it can be argued that relative to teachers and parents, inspectors have limited knowledge of both the student's education production function as well as the best use of inputs to maximize social welfare. But the incentives for teachers under an inspection regime may be distorted such that they seek to satisfy inspectors rather than serve the interests of students or their parents. Relatedly, if the inspection body takes a particular stance on pedagogical practice, there is also the danger that such a top-down approach to accountability drives out variety and experimentation in the production of education, leading to a loss of efficiency.

This logic suggests that, just as in the case of firms, the policymaker may want to limit the weight placed on the subjective component of the inspection process. As before, assessor bias is a concern. But in the case of school inspections, the discussion above suggests that an additional concern is that inspectors impose a 'cooky cutter' or 'one size fits all' approach when assessing school quality. Overall, these two mechanisms will tend to reinforce the reliance on test scores.[22] The trade-off is the potential for increased gaming behavior associated with the objective performance measure.

The empirical analysis in the next section assesses whether inspector ratings are valid, i.e. related to underlying school quality measures not observed by the inspectors. This relates to the issues of inspector bias discussed above. In particular, the question addressed is whether ratings add any value in terms of providing additional information on school quality over and above that already available in the public sphere, such as test scores.

Teachers' Behavioral Response to a Fail Rating

As noted above, there are clear sanctions for schools inspectors judge to be failing. Teachers may respond directly to such incentives by increasing the supply of effort. These incentives may also be mediated through the school principal, who arguably faces the strongest sanctions.[23]

[21] Inspectors may of course apply some discount factor when evaluating quality of observed lessons. Nevertheless, there is evidence to suggest that the performance measurement technology is indeed imperfect. When inspections moved from many weeks notice to a few days notice in 2005, there was a dramatic rise in the proportion of schools failing the inspection. One explanation for this rise is that under the longer notice period teachers were able to put in place processes which artificially boost measured quality in time for the inspection visit. The possibility remains that such strategies are employed even under very short notice inspections.

[22] One policy rule might be for inspectors to concentrate their efforts on schools falling below a given threshold on the objective performance measure. As discussed in the previous section, such an approach may be in line with the one adopted by the English inspectorate starting in September 2009.

[23] In a private sector setting, Bandiera et al (2007) show that when managers are incentivized on the average productivity of lower tier workers they target their effort to particular workers and select out the least able workers, raising overall productivity.

However, strong incentives to perform on a particular metric (test scores) may also lead teachers to try to game the system. Such gaming behavior may have distributional consequences and may also lead to misallocation of resources. 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 (Baker, 1992; Holmstrom and Milgrom, 1991).

In the schooling setting there are a number of dimensions along which such strategic 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). 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 (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 (Neal and Schanzenbach 2010). Fourth, there may be outright cheating by teachers (Jacob and Levitt 2003). In the empirical analysis below, I outline the approach I adopt to detect these types of responses.

Aside from these strategic concerns, at a theoretical level there is another reason to expect heterogeneity in student test score gains in response to the fail treatment. This arises from the hypothesis that parents may have differential ability to monitor their child's teacher and the progress the child makes in school. If teachers are able to slack when monitoring by parents is less effective then it is possible that students even in the same classroom receive differential levels of attention and effort from the teacher. If teachers raise effort levels in response to a fail inspection, they may do so where the marginal gain is greatest. Arguably, this may be with respect to students who received the least attention prior to the inspection. In the empirical analysis below, I investigate whether the evidence supports the hypothesis that gains from a fail inspection fall disproportionately in favor of students whose parents may be the least effective in monitoring the quality of education provided by the school.

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.[24] This speaks to the issues of inspector bias discussed in section 2.2. Furthermore, if inspectors rely mostly or

[24]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.

exclusively on test scores to arrive at the overall rating then, conditional on test scores, these ratings will have limited power to forecast measures of underlying school quality not observed by the inspectors.

This test may be compared to that proposed by Jacob and Lefgren (2005). They find that the *school principal's* subjective assessment of teacher effectiveness can help forecast a measure of parental satisfaction (parent requests for a teacher). In addition, in on-going work, researchers have found that student perceptions of a teacher's strengths and weaknesses are related to achievement gains (including in other classes taught by the same teacher).[25]

The data employed to construct the two measures of underlying school quality (student perceptions and parent satisfaction) come from the Longitudinal Survey of Young People in England (LSYPE). This is a major survey supported by the Department for Education. The survey includes the following six questions asked of students aged 14 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. Five questions asked of parents relate to satisfaction with the following aspects: interest shown by teachers in the child, school discipline, child's school progress, and feedback from teachers.

A composite student-level score is computed by taking the mean of the responses to the six questions relating to student perceptions. These student-level means are then converted into z-scores by normalizing them to mean zero and standard deviation one.[26] 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:

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

where i indicates individual survey respondent (the unit of observation) at school j in local authority k. X_{ij} captures school- and student-level variables and λ_k represents local authority fixed effects. School-level variables include the school's national percentile test rank, the proportion of students eligible for a free lunch and whether the school is secular or religious. Detailed student background controls include prior test score, ethnic background, parents' education, income, economic activity and whether in receipt of government benefits. 'Rating_j' is the school's inspection rating.[27] Similar regressions are run for the parent-level satisfaction z-score outcome, computed

[25] Bill and Melinda Gates Foundation (December 2010), *Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project*, report, available at: http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf (accessed 15 December 2011).

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

[27] 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. Thus, 'Rating_j' corresponds to the rating from one of these three years. (Over this period schools 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, some of 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

using the five parent satisfaction questions.

The inspection ratings are then 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 conditional association between the inspection rating and the measure of quality, 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. Such feedback information from consumers is typically not observed in the public sector.[28] Inspection ratings may play an important role in helping to fill this gap if they can be used to forecast consumers' views of quality.[29]

Results

Column 1 of Table 1 shows the unconditional relationship between the teacher practices z-score and the inspection rating. The results suggest 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.[30] Thus, the gap in the z-scores between an Outstanding (Grade 1) and a Fail (Grade 4) school is around 0.7 of a standard deviation.

Controlling for the school's test percentile and the proportion of students receiving a free school meal in column 2 leads to a 40% reduction in the association between the inspection rating and the teacher practices z-score. Given that inspection ratings and test scores are correlated, this is not surprising. The noteworthy finding is that the coefficient on the inspection rating remains large, in relative terms, and is highly significant. (Also included as controls in column 2 are the size and type of school as well as local education authority fixed effects.)

Column 3 includes detailed controls on students' family background and prior test scores.[31] These lead to a minor fall in the absolute size of the coefficient on inspection ratings, which remains statistically significant at the 1 percent level. Column 4 includes the inspection rating prior to the year of the student interview, 2003/04. This addresses the concern that students' survey responses may be influenced by past inspection ratings. If a school's inspection ratings are correlated over time then the results in Table 1 may simply be capturing the effect of past

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.

[28] Heckman (2000) notes that for 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" (p. 24).

[29] For example, if inspection ratings can be used to predict student perceptions of the quality of teaching, then for parents currently engaged in choosing schools, it may be optimal to place some weight on these ratings when making their decisions.

[30] Note that although the R-squared value of 0.035 for the regression in column (1) is small, this should not come as a surprise given the noisy nature of the left hand side variable. 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.

[31] 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. Including detailed student background controls should address some of these concerns.

inspections on respondents' views. The results in column 4 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 4 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 the school's position in the national test score distribution: the results show that a fall of 50 percentile points is associated with a fall of 0.15 (0.0029 x 50) of one standard deviation in the teacher practices z-score. Compare this with a two-unit deterioration - e.g. from Outstanding to Satisfactory - in the inspection rating: this is associated with a decline of 0.21 of one standard deviation in the teacher practices z-score.

Results for the parent satisfaction outcome are very similar to those reported for the teacher practices outcome. For brevity, column 5 in Table 1 shows the results for the parent satisfaction outcome with the full set of controls. These results demonstrate that the association between inspection ratings and parental satisfaction is also strong. A two-unit deterioration in the inspection rating is associated with a decline of 0.17 of one standard deviation in the parent satisfaction z-score.[32]

In summary, this analysis reveals that inspection ratings can help detect good and poor teacher practices (or high and low parental satisfaction) 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. 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 local authority fixed effects as well as observable school and detailed 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.

[32] Further analysis, not reported here, allows for non-linear effects. For the teacher practices outcome, there is some evidence suggesting a mildly concave relationship between teacher practices and inspection ratings: the gap is largest when we move from a Grade 1 school to Grade 2; it is smallest between Grade 3 and Grade 4 (Fail) schools. For the parental satisfaction outcome, the linearity assumption appears to be justified.

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? As described earlier, selection into the fail treatment is based at least partly on past test performance. Therefore, a simple school fixed-effect analysis using pre- and post-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 arguably any test score rise following a fail inspection would have occurred even in the absence of treatment.

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 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.

This insight enables me to identify credible causal estimates of the short-term effects of a fail inspection. 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.[33] 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 primary schools inspected and failed in England in the four years 2005/06 to 2008/09.[34] For each year the first two columns show means for schools failed early in the academic year (September to

[33] 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.

[34] 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.

November[35]) 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, respectively.

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. For example, for schools failed in 2005/06 the first two columns show that the mean year of inspection for late inspected schools is 2000.6; for early inspected schools it is 2000.1.[36] This suggests that schools inspected slightly earlier in the previous inspection round are also inspected slightly earlier in 2005/06. Table 2 shows that this pattern is typical across all four fail years.[37] Thus Table 2 demonstrates that over the period relevant to this study, the timing of inspection is exogenous: within a given year, the month of inspection is determined by the timing of the previous inspection.[38]

The third, fourth and fifth rows report the proportion of students receiving a free school meal (lunch), the proportion of students who are white British and the school's inspection rating from the previous inspection round. The table demonstrates that for each of the four inspection years the differences in means between the treatment and control schools are small and are statistically insignificant. (The only exception is the previous inspection rating 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 fail schools perform between 0.4 and 0.5 of one standard deviation below the national mean. This is in line with the idea that inspectors select schools for the fail treatment at least in part on the basis of past performance.[39]

[35] The early inspection category is expanded to three months in order to increase the sample of treated schools.

[36] 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.

[37] Further analysis of the data on timing of inspections 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.

[38] 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 failed in 2005/06 (columns 2 and 3 in Table 2).

[39] One remaining concern is that the worst schools may be closed down after a fail inspection. If this happens immediately after an inspection then any test score gains posted by the early fail schools may be due to these selection effects. In fact, three pieces of evidence suggest that this is not the case. Although such a 'weeding out' process may be important in the medium term, the evidence in Table 2 demonstrating the comparability of the early and late inspected group of schools suggests that such a process does not take place immediately, i.e. in the year of inspection.

Second, an examination of the data shows that the probability of dropping schools from the estimation sample because of a lack of test data from the May tests in the year of inspection - perhaps because the school is closed down - when test data are available in previous years, is similar for treated and control schools. In total, for the years 2005/06 to 2008/09, 4 per cent of schools (6 schools) from the control group and 5 per cent (14 schools) from the treatment group are dropped because of lack of test score data in the year of inspection. These schools appear to be comparable to the treatment and control schools on characteristics such as student attainment in the year before inspection. For example, the proportion of students attaining the mandated attainment level for age 11 students in the year before inspection is 62 per cent for the 14 treated (early inspected) schools dropped from the estimation sample; the corresponding mean is 63 per cent for the 258 treated schools included in the estimation

In sum, the evidence in Table 2 demonstrates that there is little difference between control and treatment schools on observable characteristics. This, combined with the fact that timing is determined by a mechanical rule, suggests that unobservable differences are also unlikely to exist between the control and treatment groups.

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 2005 (the treatment group) and June 2006 (the control group). 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.

OLS models of the following form are estimated:

$$y_{is} = \alpha + \delta D_s + X_{is}\beta_1 + W_s\beta_2 + u_{is}, \tag{2}$$

where y_{is} is the May 2006 test score outcome on the age-11 (Key Stage 2) test for student i attending school s. The treatment dummy is defined as follows: $D_s = 1$ if school s is failed in September 2005 and $D_s = 0$ if the school is failed in June 2006. X_{is} is a vector of student demographic controls and W_s is a vector of pre-treatment school characteristics. Given the evidence on assignment to a September inspection versus a June inspection presented in the previous sub-section, it can be credibly argued that treatment status D_s is uncorrelated with the error term, u_{is}.[40]

Results are also presented using difference-in-differences (DID) models. Continuing with the example of schools failed in 2005/06, data are taken from 2004/05 (the 'pre' year) and 2005/06 (the 'post' year). The following DID model is estimated:

$$y_{ist} = \alpha + \eta post_{06} + \delta D_{st} + X_{is}\beta_1 + \lambda_s + u_{ist}, \tag{3}$$

where $t = 2005$ or 2006, corresponding to the academic years 2004/05 and 2005/06, respectively. λ_s is a school fixed effect and $post_{06}$ is a dummy indicator, switched on when $t = 2006$. D_{st} is a time-varying treatment dummy, switched on in the post period ($t = 2006$) for schools inspected early in the academic year 2005/06.[41]

---

sample.

Finally, a comparison of results for the moderate fail schools versus the severe fail ones also sheds light on this issue. Note that a moderate fail is unlikely to lead to changes in school leadership and school closure. If test gains following a fail are observed for moderate fail schools then selection effects arising from school closures are unlikely to be driving the results. I report on these findings are below.

[40] In a heterogenous treatment effect setting where δ_i is the student-specific gain from treatment, the key assumption is that D_s is uncorrelated with both u_{is} and δ_i. The evidence presented above suggests that this assumption is satisfied. In this case a comparison of means for the treatment and control outcomes yields the Average Effect of Treatment on the Treated. This is the effect of a fail inspection rating for schools inspectors judge to be failing. Another parameter of policy interest - not estimated - is the Marginal Treatment Effect, i.e. the test score gain for students in schools on the margin of being failed.

[41] The key DID assumption, which embodies the assumption of common trends across treatment and control groups, is that conditional on the school fixed effect (λ_s) and year ($post_{06}$) the treatment dummy D_{st} is uncorrelated with the error, i.e. $E(u_{ist} \mid \lambda_s, post_{06}, D_{st}) = 0$.

4.1 Testing for Strategic Behavior

As highlighted in section 2.2 above, a growing body of evidence has demonstrated that when schools face strong incentives to perform on test scores they game the system. These stratgeies include the removal of low ability students from the testing pool, teaching to the test and targeting students close to the mandated proficiency threshold.

In the analysis below, I test for the presence 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.[42] This involves checking whether the estimated effect of treatment in the OLS and DID regressions (δ 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 test scores.[43]

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, students are enrolled in primary school at the time of the fail inspection. 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 have 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.[44] The key headline measure of performance used by the government and commonly employed to rank schools is the percentage of students attaining 'Level 4' proficiency on the age-11 Key Stage 2 test. Following a fail inspection the incentives to maximize students passing over the threshold are 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 teachers may target resources towards students on the margin of attaining this threshold, to the detriment of students far below and far above this critical level.

[42] 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 incentives to improve test score performance following a fail inspection are indeed very strong.

[43] 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.

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

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 neglect students at the bottom of the prior ability distribution in response to the introduction of performance thresholds (see Neal and Schanzenbach, 2010).

Appendix Table A1 shows the distribution of Year 6 students achieving the 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 expected, Table A1 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 percent at the top quartile of prior ability. As the final rows of Table A1 show, in the year before the fail inspection the average number of students achieving the Level 4 threshold is 67 and 72 percent for mathematics and English, respectively. One implication of the evidence presented in Table A1 is that students in the lowest ability quartile are the least likely to attain the official threshold, and so at fail schools teachers may substitute effort away from them towards students in, for example, the second quartile. The analysis below tests this prediction.

A second approach to analyzing the distributional effects of a fail rating is to employ quantile regression analysis. In particular, I investigate distributional effects *within* prior ability quartiles. 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 absence of a fail treatment 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, labeled 'T0'. For illustrative purposes, suppose that 20 percent of students from the bottom quartile attain proficiency; 50, 75 and 90 percent do so in the second, third and top quartiles, respectively. (These numbers correspond roughly to actual data for fail schools.) Following a fail inspection, and on the assumption that 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 simple theory.

5 Results

5.1 Basic Results

Table 3 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.[45] The top panel reports results from the OLS model and the bottom panel reports results from the difference-in-differences model.

I pool the four inspection years together. Pooling over the four years is justified because over this period schools were inspected and rated in a consistent manner.[46] The evidence presented in Table 2 shows that schools are indeed comparable on observable characteristics across the different years. As a robustness check, results from regression analysis conducted for each year separately are also reported (in the Appendix Tables A2 and A3). As indicated below, these show that results for the pooled sample and for individual years produce a consistent picture of the effects of a fail inspection.

In Table 3, as well as the following tables, the comparison is between students enrolled in schools failed in the early part of the academic year, September to November - the treatment group - with those attending schools failed late in the academic year, mid-May to mid-June - the control group.[47]

Turning first to mathematics test scores, the row 'early fail' in Panel A of Table 3 corresponds to the estimate of the treatment effect δ in equation (2). Column (1) reports estimates from the simplest mode with only school-level controls.[48] The result in column (1) suggests that the effect of a fail rating is to raise standardized test scores by 0.12 of a standard deviation. This effect is highly 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

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

[46] 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 a notice period of many months. This regime has remained in place since September 2005.

[47] Treatment effects, not reported here to conserve space, estimated for schools failed in each individual month from September through to April yield a pattern which suggests a steadily declining effect as we move closer to May. For example, a comparison of April failed schools with the control group of schools (i.e. schools failed mid-May to mid-June) reveals effects which are both small and not statistically significantly different from zero.

[48] The following school-level controls are included in all the regressions reported in Panel A of Table 3: pre-inspection math and English attainment; percent of students eligible for free lunch; and percent of students who are non-white. Dropping these from the regressions makes very little difference to the estimates. For example, without any controls at all, the estimated effect for mathematics is 0.10 of a standard deviation.

background; and census information on the home neighborhood deprivation index. The model in column (3) also includes the student's age seven (Key Stage 1) test scores.

The rise in the R-squared statistics as we move from columns (1) to (2) and then to (3) clearly indicates that student background characteristics and early test scores are powerful predictors of students' test outcomes. However, the addition of these controls has 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 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 whilst the 'post' year corresponds to test scores from the year of inspection. The estimate of mean reversion is provided by the gain in test scores between the pre-inspection year and the year of inspection for schools failed late in the academic year (i.e. the control group). This estimate is indicated in the row labeled 'post'. The DID estimate of the effect of a fail inspection 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 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 earlier contention that a fail inspection raises student test scores and, 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 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 0.09 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. As before, the results for English provide no evidence of gaming behavior: 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 shows that 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.[49]

Falsification Test and the 'Pre-Program Dip'

Table 4 presents analysis from a falsification exercise. This makes use of the fact that data are available one 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 in fact there was none?

Table 4 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 close to zero and statistically insignificant 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.[50]

Finally, Table 4 also provides 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 sheds some light on the selection rule employed by inspectors: for English at least, this evidence suggests that inspectors are more likely to fail schools which have had a recent dip in test score performance.

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.

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. In order to test the prediction that low ability students are adversely affected when incentives to attain the performance threshold are strengthened, I test whether the effect of treatment varies with prior ability.[51] The following model incorporating the

[49] 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.

[50] Results for individual inspection years confirm the finding that the placebo treatment produces no discernible effect.

[51] Appendix Table A1, discussed in section 4.1, highlighted the relationship between prior ability - measured by the age-seven Key Stage 1 test score - and the probability of attaining the target level on the age-11 Key Stage 2 test. This showed that only around a quarter to one third of students in the lowest prior ability quartile attain the stipulated Level 4.

interaction between the treatment dummy and prior ability is estimated:

$$y_{is} = \alpha + \delta D_s + \gamma.Percentile_{is} * D_s + X_{is}\beta_1 + W_s\beta_2 + \beta_3 Percentile_{is} + u_{is}, \quad (4)$$

where the treatment dummy D_s is turned on for schools inspected early in the academic year. $Percentile_{is}$ is student i's percentile, within the the selected sample of fail schools, in the prior test score distribution (the age-seven Key Stage 1 test). Thus, the coefficient on the interaction between the treatment dummy and the test percentile, γ, estimates how the effect of treatment varies by prior ability.

The effect 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:

$$y_{is} = \alpha + \delta D_s + \sum_{k=2}^{4} \gamma_k Q_{isk} D_s + X_{is}\beta_1 + W_s\beta_2 + \sum_{k=2}^{4} \beta_{3k} Q_{isk} + u_{is}, \quad (5)$$

where the dummy variable Q_{isk} is switched on for student i if her percentile on the prior test score lies in quartile k. Thus, γ_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 5, columns (1) and (3), presents estimates of the main (δ) and interaction (γ) effects for mathematics and English, respectively, for the linear interactions model (4). The row 'early fail' corresponds to the estimate of δ and 'early fail x prior ability percentile' corresponds to the estimate of γ. 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 0.15 of a standard deviation for mathematics and English, respectively.

The estimates for the nonlinear interactions model, equation (5), are reported in columns (2) and (4).[52] 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 0.03 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.

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 percent attainment. However, it should be noted that even for students in the highest ability quartile at fail schools, the mean test scores in the year before treatment are some way below the 100 percent mark (76 percent and

[52] 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.

68 percent for mathematics and English, respectively). This hypothesis that ceiling effects bite is explored further (and rejected) in the quantile treatment effect analysis below.

In summary, the results presented in Table 5 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.[53] 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

This section further explores distributional effects by examining how the conditional distribution of test scores is affected by treatment at each quantile $\tau \in [0, 1]$. The following model is estimated:

$$Q_\tau(y_{is} \mid .) = \alpha_\tau + \delta_\tau D_s + X_{is}\beta_{1\tau} + W_s\beta_{2\tau}, \tag{6}$$

where $Q_\tau(. \mid .)$ is the τ^{th} conditional quantile function and δ_τ is the quantile treatment effect at quantile τ. Figure 4 plots δ_τ as well as the associated 95 percent confidence interval. For mathematics, 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 at all quantiles. For English (Panel B) the effect varies between 0.05 and 0.1 of a standard deviation, with the largest effects recorded for quantiles below 0.7. In addition, the evidence from Figure 4 tends to reject the hypothesis that teachers act strategically to raise performance of students on the margin of attaining the official government target.[54]

Importantly, the pattern of treatment effects across quantiles reported in Figure 4 tends to reject 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, certainly for mathematics.

Next, I investigate quantile treatment effects within prior ability subgroups. I focus on quantile treatment effects within each prior ability quartile. There are two justifications for this. First, the evidence in Table 5 points to heterogenous gains across these four ability groups. It is possible that within these subgroups, teachers target students who are on the *margin* of attaining the performance threshold rather than the *average* student. Quantile regression analysis will provide some evidence on this issue. Second, looking for heterogeneous effects within prior ability subgroups accords with the notion that teachers may set (track) students within (among) classes by ability.

[53] 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.

[54] Recall that the evidence in Appendix Table A1 shows that at fail schools, around 70 per cent of students attain the government's target for mathematics and English in the year prior to the inspection. Thus, if teachers can identify and strategically target the marginal students we would expect the treatment effect to peak at around quantile 0.3. Broadly speaking, this does not appear to be the case: for mathematics the treatment effects are relatively stable across most of the test score distribution; for English the treatment effect is stable up to quantile 0.7, with evidence of some decline at higher quantiles.

The incentives they face suggest that they may target effort towards marginal students within these subgroups.

Figures 5 and 6 plot the quantile treatment effects within each prior ability quartile, for mathematics and English, respectively. These demonstrate that there is a great deal of heterogeneity in estimated effects within each quartile. Perhaps the most marked heterogeneity is for students in the bottom prior ability quartile, where 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 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 performance threshold.[55] The evidence does not generally support this hypothesis. First, it should be noted that Panel A, Figure 5 and Panel A, Figure 6 demonstrate that there are substantial gains for students even at low quantiles, i.e. students quite far from the performance threshold post large gains from the treatment. In addition, the evidence from the remaining three panels (prior ability quartiles 2, 3 and 4) in each of Figures 5 and 6 does not generally support the view that teachers target the marginal students.[56]

Further Subgroup Analysis

Table 6 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 standard deviation (though only marginally significant) for the first group of students and 0.08 of a standard deviation for the latter group.[57]

Discussion: Explaining the Gains for Low Ability Students

[55] As indicated in Appendix Table A1, these students are likely to be in the higher quantiles of the test score distribution: Appendix Table A1 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).

[56] For example, for the second quartile prior ability subgroup the evidence in Table A1 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 A1 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.

[57] Dustmann et al (2010) show that even though students from most minority groups lag behind white British students upon entry to compulsory schooling, they catch up strongly in the subsequent years. The relatively large gains for this subgroup reported in Table 6 suggest that one mechanism driving the results reported in Dustmann et al may be the school inspection system.

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. 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 in Table 5 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 5.

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.[58]

5.3 Medium-Term Effects

The results reported above show that a fail inspection leads to test score gains for age-11 (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 would provide indirect evidence 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.[59]

[58] This interpretation of the results is also supported by the subgroup analysis of Table 6, 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 6 boost the conclusion that it is children from these sorts of families who are helped most by the fail inspection.

[59] 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 7 reports results for the Key Stage 3 test score outcome for students aged 14 (Year 9), 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 analysis above) and are currently only available for students at primary schools failed in 2006 (Key Stage 3 test taken in 2009) and 2007 (Key Stage 3 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 7 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 percent 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.

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.

5.4 Mechanisms

Having ruled out certain types of gaming behvaior, this section provides tentative evidence on what might be driving test score improvements at failed schools. First, I investigate whether moderate and severe fail ratings - each of which entails different degrees of intervention following the fail inspection - yield markedly different outcomes. Second, using teacher survey data, I examine whether fail schools experience changes in teacher tenure, curriculum and classroom discipline.

Effects by Severity of Fail

As discussed in Section 2, the overall fail rating can be sub-divided into a *moderate* fail and a *severe* fail: the 'Notice to Improve' and 'Special Measures' sub-categories, respectively. It was noted above that the moderate fail rating leads to increased oversight by the inspectors but does not entail other dramatic changes in inputs or school principal and teacher turnover. Schools subject to the severe fail category, on the other hand, may well experience higher resources as well as changes in the school leadership team and the school's governing board.

Table 8 shows the effects on test scores separately for schools receiving a moderate fail (columns 1 and 2) and severe fail (columns 3 and 4). For moderate fail schools, the OLS (difference-in-difference) estimates suggest gains of 0.16 (0.11) and 0.07 (0.03) of a standard deviation for mathematics and English, respectively. For the severe fail treatment, the OLS (difference-in-difference) estimates show gains of 0.10 (0.11) and 0.13 (0.15) of a standard deviation for mathematics and English, respectively.

The finding that there are test score gains - certainly for mathematics - at both moderate and severe fail schools is noteworthy. Given that large injections of additional resources and personnel

changes are less likely at moderate fail schools than at sever fail schools, the findings in Table 8 point to greater effort and increased efficiency as the key mechanism behind gains for the former group.[60]

Classroom Discipline, School Curriculum and Teacher Tenure

In this section I use a survey of teachers to investigate whether a fail rating leads to changes in the following outcomes (i) classroom discipline, (ii) hours of the curriculum allocated to high- and low-stakes subjects and (iii) teacher tenure. In one set of regression results, schools rated 'Satisfactory', the rating just above the fail category, are used to construct the control group. A second control group is constructed using schools which are 'yet-to-be-failed'.

The teacher survey data are part of a major longitudinal study, the Millennium Cohort Study, which follows children born in the UK in or just after the year 2000. In the fourth wave of the study the primary school teacher of the study child was contacted and surveyed in one of the academic years 2007/08 or 2008/09. A small set of questions in the survey relate to the teacher's tenure at the school, years of experience, school curriculum and classroom discipline.

Table 9 reports results for the classroom discipline outcome.[61] For each of the columns, (1) to (4), the 'Fail 2004 - 2007' dummy is switched on if the teacher is at a school which was failed in one of the academic years 2003/04 to 2006/07 (i.e. before the survey date).

In columns (1) and (2) the control group consists of teachers at schools which were rated 'Satisfactory' between 2003/04 and 2006/07. In columns (3) and (4) a different control group is constructed using teachers at schools which are failed *after* the interview, namely in 2009/10 or 2010/11. These are the 'yet-to-be-failed' schools.[62]

Column (1) shows that the raw difference in teacher-reported discipline problems are lower at schools which experienced a fail rating in the recent past than at schools which received a Satisfactory rating. This gap is substantial (5.5 percentage points) but not statistically significant. When school controls are included in column (2), this gap widens to 8.0 percentage points and is significant at the 10 percent level.[63] This is a large effect, representing a 20 percent decline in this measure of indiscipline. One interpretation of this evidence is that teachers in treated schools places greater emphasis on classroom discipline and ensure that fewer students behave in a way

[60] One other notable feature of the results in Table 8 worth highlighting is the contrasting nature of mean reversion in the moderate and severe fail categories. The extent of reversion-to-mean for the control groups is depicted in the 'post' row of Table 8. For the moderate fail schools, there appears to be substantial mean reversion: there is bounce back, compared to the prior year, in mathematics and English test scores for the moderate fail schools of 0.06 and 0.14 of a standard deviation, respectively. For the severe fail schools, however, there is *no* bounce back. One interpretation of this evidence is that inspectors are able to distinguish between poorly performing schools and the very worst performing schools, with the latter category exhibiting no improvements in the absence of treatment.

[61] The precise question in the survey is: 'Are there any children in the study child's class whose behaviour in class prevents other children from learning?' (check Yes or No).

[62] Any schools which were also failed in one of the previous inspection cycles between 2003/04 and 2008/09 are dropped from this control group.

[63] School controls included in the regression: percentage of students eligible for a free lunch, the school's national test score percentile and type of school.

that impedes other children from learning.[64] The second set of results, reported in columns (3) and (4) are consistent with the previous results, although the standard errors are now substantially larger.[65]

Appendix Table A4 shows the effect of a fail inspection on teacher tenure and experience (columns 1 to 4) as well as number of hours devoted each week to English, mathematics and physical education (columns 5 to 10). The point estimates for teacher tenure and experience suggest very small differences between teachers in treatment and control schools. For example, lower tenure of 0.198 of a year for the treatment group represents a decline in tenure of less that 3 percent (mean tenure at control schools is 7.3 years). This suggests that higher teacher turnover is unlikely to be the mechanism behind the positive test score gains at fail schools. Differences in experience are even smaller. However, due to the relative small sample size, standard errors are large and hence larger effects cannot be ruled out. Similarly, the point estimates for the curriculum outcomes suggest small effects from the treatment: there appears to be a 2 (5) percent decline (increase) in hours devoted to math (English) and a 4 percent decline in hours devoted to physical education. In part, this finding may be a consequence of the fact that the national curriculum in England is set nationally.

Overall, the analysis presented here further boosts the hypothesis that greater effort on the part of the current stock of teachers at the fail school is at least part of the explanation for the test score gains reported in this study. However, without more detailed survey information on school practices, hiring policies and teacher turnover it is not possible to go beyond the tentative evidence provided here.

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. A key concern under such a regime is that it is open to manipulation by the bureaucrats charged with oversight.

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. The evidence suggests that inspectors are able to discriminate between more and less effective schools, and, significantly, report on their findings in a setting where the stakes are high. Thus, this aspect of school inspections - simply disseminating inspection ratings and reports - may help

[64] This result is unlikely to be driven by student selection into schools. Hussain (2009) shows that enrolment declines following a fail inspection. More motivated parents are more likely to switch schools, in which case discipline should *deteriorate* following a fail inspection.

[65] As a robustness check, using different sample selection rules, e.g. 2005/06 and 2006/07 Fail and Satisfactory schools, yields qualitatively similar results, though the reduction in sample size results in larger standard errors.

relax information constraints facing consumers and other decision makers. Such constraints are thought to be pervasive in the public sector.[66]

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 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 artificially inflate test performance by gaming the system. Given the prior evidence on strategic behavior in similar high stakes contexts, the fact that I find little evidence of dysfunctional response is revealing. If inspectors are able to successfully evaluate actual practices and quality of instruction in place at the school, both before and after a fail inspection, inspections may well have a mitigating effect on such distortionary responses.

Finally, examining treatment heterogeneity reveals that gains are especially large for students scoring low on the prior (age seven) test. The gains are large when compared to other possible policy interventions, such as the effects of 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.[67]

These findings are noteworthy given the prior empirical evidence suggesting that subjective assessments may give rise to various kinds of biases. For example, subjective evaluations of workers may lead to 'leniency' and 'centrality' bias in private sector firms (Prendergast, 1999) . Evidence from the public sector points to bureaucrats indulging their preferences when allowed to exercise discretion (Heckman et al, 1996).[68] 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 those charged with allocating training in the

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

[67] 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 robust evidence in the current literature usually focuses on children of motivated parents who *are* active in such programs. The most credible studies tend to make use of lottery assignment to determine, for example, the effects of attending a higher performing school (Hastings et al, 2009; Abdulkadiroglu et al, 2011).

[68] Heckman et al (1996) show that in the context of a job training program, case workers ignore the overall performance standard set for the training center, and instead enroll the least advantaged and least employable applicants into the program.

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.

Allen, Rebecca and Simon Burgess (2012), "How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England", CMPO Working Paper No. 12/287, Bristol.

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

Baker, George P. (1992), "Incentive Contracts and Performance Measurement", *Journal of Political Economy*, 100(3), pp. 598-614.

Bandiera, Oriana, Iwan Barankay and Imran Rasul (2007), "Incentives for Managers and Inequality among Workers: Evidence from a Firm-Level Experiment", *The Quarterly Journal of Economics*, 122(2), pp. 729-773.

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.

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.

Dustmann, Christian, Stephen Machin and Uta Schönberg (2010) "Ethnicity and Educational Achievement in Compulsory Schooling", *The Economic Journal*, 120: F272–F297.

Faubert, Violaine (2009), "School Evaluation: Current Practices in OECD Countries", OECD Education Working Papers, No. 42.

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

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.

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., 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.

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 (2005), "Principals as Agents: Subjective Performance Measurement in Education", NBER Working Paper No. 11463.

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., Eric S. Taylor, John H. Tyler and Amy L. Wooten (2010) "Identifying Effective Classroom Practices Using Student Achievement Data", NBER Working Paper No. 15803.

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.

Edward P. Lazear (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 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.

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.

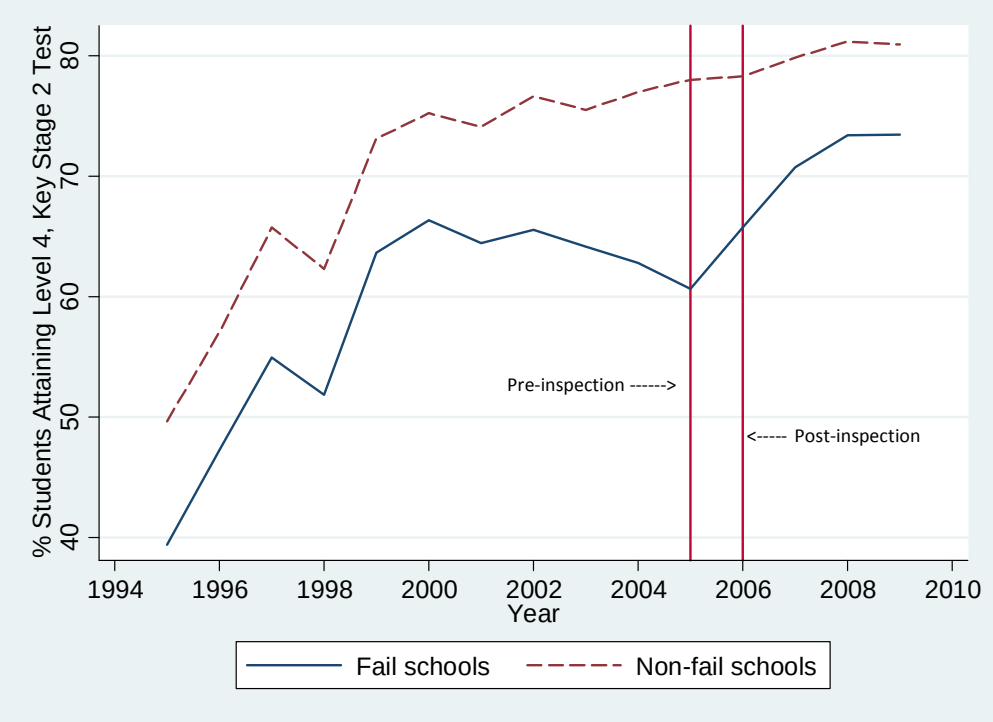
Propper, Carol, Sutton, Matt, Whitnall, Carolyn and Windmeijer, Frank (2010) "Incentives and Targets in Hospital Care: Evidence from a Natural Experiment", *Journal of Public Economics*, 94, pp. 318-335.

Rockoff, Jonah E., and Cecilia Speroni (2010), "Subjective and Objective Evaluations of Teacher Effectiveness", *American Economic Review*, 100(2): 261–66.

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.

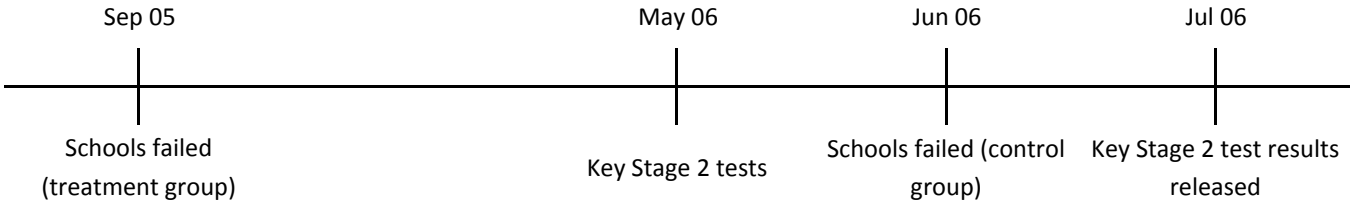
Taylor, Eric S. and John H. Tyler (2011), "The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-career Teachers", NBER Working Paper 16877.

Figure 1 - Relative Test Score Performance at Fail Schools, 2005/06 Inspections: Causal Effect or Mean Reversion?



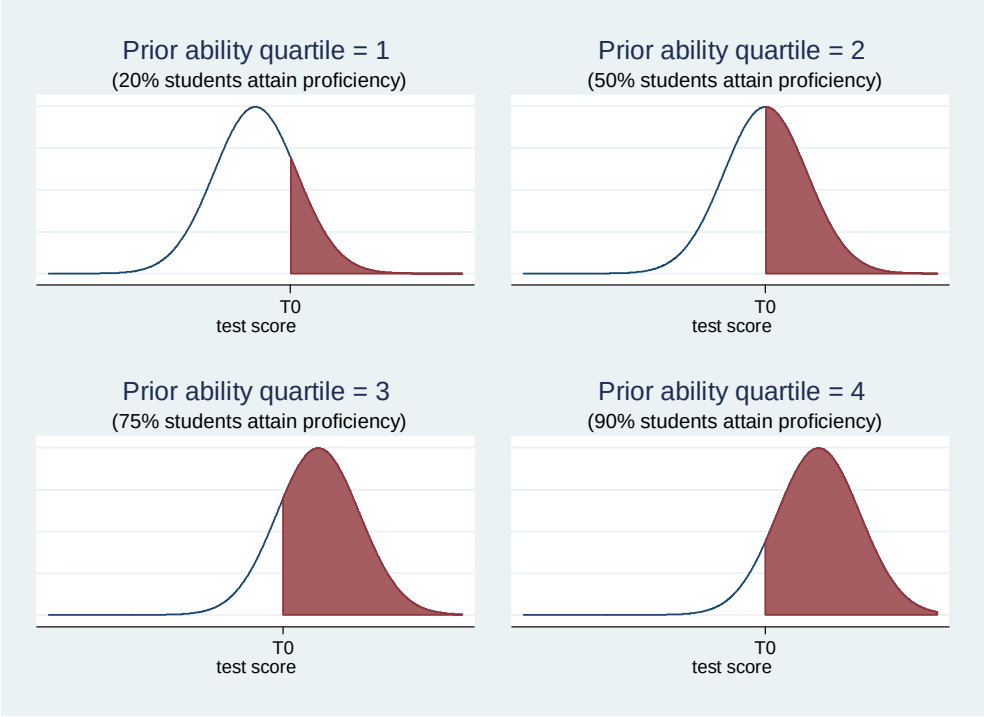
Notes: Figure shows the percentage of students attaining the government's performance threshold, 'Level 4', on the age-11 Key Stage 2 mathematics test for fail and non-fail schools. Each year represents an academic year (e.g. 2002 corresponds to 2001/02). The figure shows that between 2000 and 2005, test score performance declined at schools inspected and 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. (Note that inspections are undertaken throughout the academic year, September through to July. The 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.)

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). See section III for details.

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

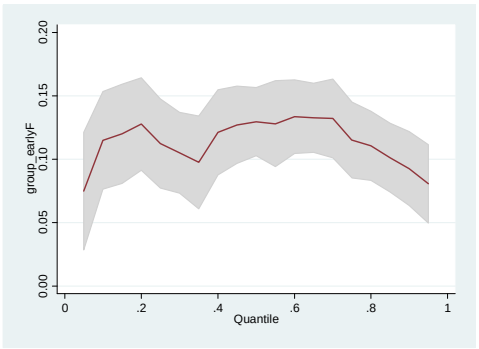


Note: 'T0' denotes the official proficiency threshold.

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

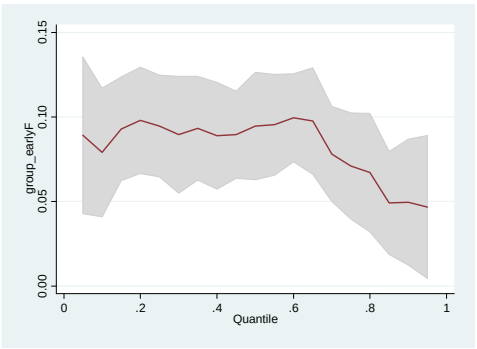
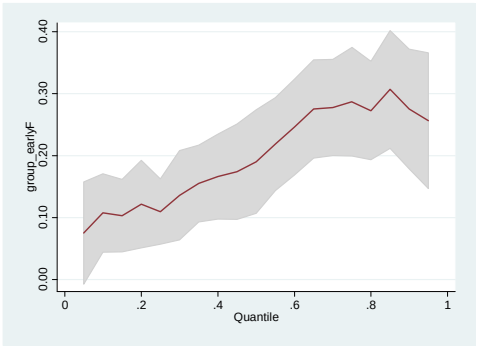


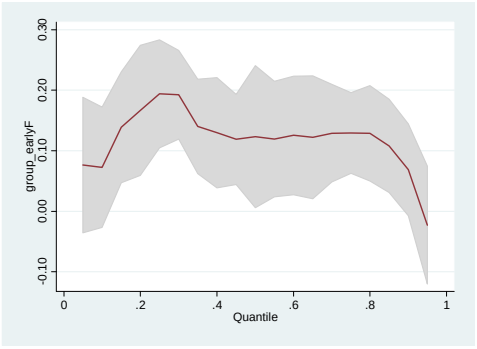
Figure 5: Quantile Regression Estimates: by Prior Ability Quartile (Mathematics)

Outcome variable: age 11 (Key Stage 2) national standardised test score; Prior ability = age 7 test

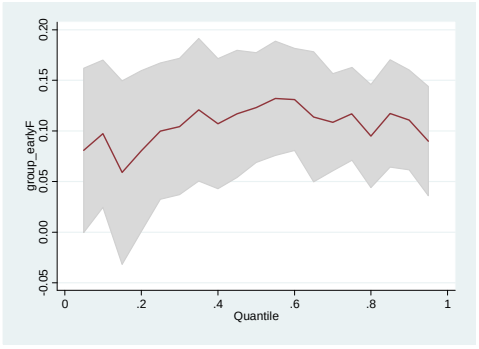
Panel A: Prior ability quartile = 1



Panel B: Prior ability quartile = 2



Panel C: Prior ability quartile = 3



Panel D: Prior ability quartile = 4

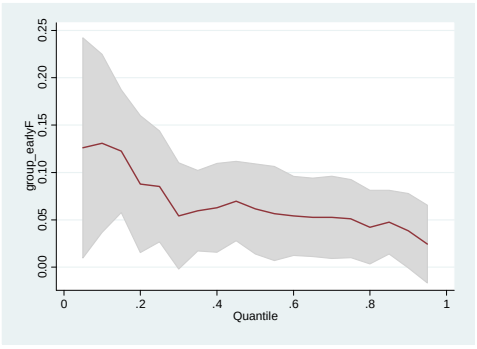
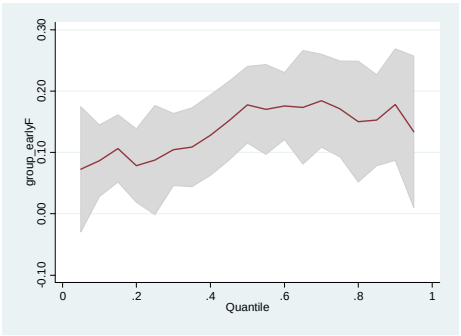


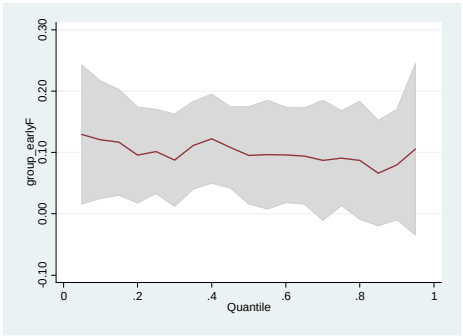
Figure 6: Quantile Regression Estimates: by Prior Ability Quartile (English)

Outcome variable: age 11 (Key Stage 2) national standardised test score; Prior ability = age 7 test

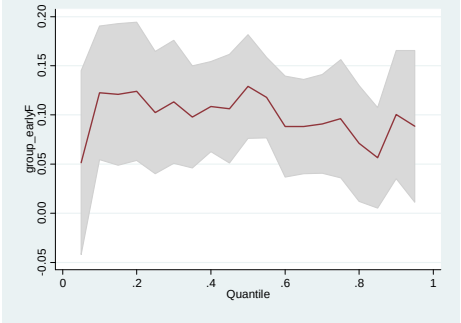
Panel A: Prior ability quartile = 1



Panel B: Prior ability quartile = 2



Panel C: Prior ability quartile = 3



Panel D: Prior ability quartile = 4

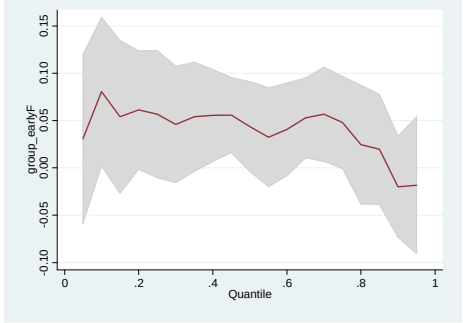


Table 1: Relationship Between Inspection Ratings and Stakeholder Perceptions and Satisfaction

(Outcomes: Student perceptions / Parental satisfaction; z-scores)

	Student perceptions of teacher practices				Parental satisfaction
	(1)	(2)	(3)	(4)	(5)
Inspection rating	-0.219**	-0.134**	-0.126**	-0.103**	-0.085**
(range: 1 = Oustanding, 4 = Fail)	(0.0191)	(0.0201)	(0.0194)	(0.0196)	(0.0210)
Test score percentile		0.0050**	0.0038**	0.0029**	0.0008
		(0.0008)	(0.0008)	(0.0008)	(0.0009)
Fraction of students eligible for		0.323	-0.049	-0.073	-0.299
free lunch		(0.1904)	(0.1855)	(0.1840)	(0.1965)
Additional school controls and Local Authority fixed effects	No	Yes	Yes	Yes	Yes
Respondent b/g controls	No	No	Yes	Yes	Yes
Previous insp. rating	No	No	No	Yes	Yes
Observations	10012	10012	10012	10012	9841
R-squared	0.035	0.084	0.146	0.149	0.152

Notes: Standard errors are reported in brackets; clustered at the school-level. ** and * denote significance at 1% and 5% level, respectively. The dependent variable is a z-score computed from: the student-level mean across six questions relating to teacher practices (columns 1 to 4); and parent-level mean across five questions relating to parental satisfaction (column 5). The inspection rating corresponds to the first inspection after the year of the survey (2004). Inspection ratings prior to the survey are included as additional controls in columns (4) and (5). School's national percentile on test score calculated using the proportion of students achieving five A*-C grades on the age 16 GCSE exams in 2004. Additional school controls (columns 2-5): dummies for type of school (Community, Voluntary Controlled, Voluntary Aided, Foundation) and log of total enrolment. Respondents' background controls in columns (3) to (5): student's prior test score (Key Stage 2 test, taken at age 11 in primary school); female dummy; whether has a long-standing illness or disability; ethnic background; parents' education, income, economic activity 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.

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 and *before* mid-July, when results are released. Mathematics and English standardised test scores are from the academic year immediately preceding the inspection year.

Table 3: OLS and DID Estimates of the Effect of a Fail Inspection on Test Scores

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

Panel A: OLS	**Mathematics**			**English**		
	(1)	(2)	(3)	(4)	(5)	(6)
early Fail	0.120**	0.127**	0.130**	0.082*	0.090**	0.090**
	(0.028)	(0.028)	(0.027)	(0.038)	(0.029)	(0.029)
Student characteristics	No	Yes	Yes	No	Yes	Yes
Age-7 test scores	No	No	Yes	No	No	Yes
R-squared	0.04	0.27	0.49	0.04	0.32	0.53
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*
	(0.033)	(0.032)	(0.030)	(0.038)	(0.036)	(0.036)
post	0.014	0.038	0.028	0.061	0.085**	0.078**
	(0.026)	(0.025)	(0.024)	(0.032)	(0.029)	(0.030)
Student characteristics	No	Yes	Yes	No	Yes	Yes
Age 7 test scores	No	No	Yes	No	No	Yes
R-squared	0.07	0.24	0.49	0.08	0.31	0.54
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. OLS and DID models estimated for schools rated Fail in the years 2006 to 2009. 'Early Fail' dummy switched on for schools failed September to November; switched off for 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: dummies for female; eligibility for free lunch; special education needs; month of birth; first language is English; ethnic group; and census information on local neighborhood deprivation (IDACI score). Missing dummies included for student characteristics and age-7 test scores. All OLS regressions in Panel A include school controls (math and English attainment; per cent free lunch; per cent non-white; all from year before inspection); DID regressions in Panel B include school fixed effects.

Table 4: The Effect of a Fail Inspection on Test Scores in the Pre-Treatment Year (Falsification Test)

(Outcome variable: age 11 (Key Stage 2) national standardized test score, *in year before inspection*)

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. OLS and DID models estimated for schools rated Fail in the years 2006 to 2009. 'Early Fail' dummy switched on for schools failed September to November; switched off for 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. See notes to Table 3 for student- and school-level controls. All DID regressions in Panel B include school fixed effects.

Table 5: Ability Interactions

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

	Mathematics		English	
	(1)	(2)	(3)	(4)
	Linear	Non-linear	Linear	Non-linear
early Fail	0.193**	0.198**	0.147**	0.141**
	(0.045)	(0.043)	(0.044)	(0.042)
early Fail x prior ability percentile	-0.00193**		-0.00155*	
	(.00069)		(0.00075)	
early Fail x prior ability quartile=2		-0.074		-0.046
		(0.046)		(0.045)
early Fail x prior ability quartile=3		-0.095*		-0.058
		(0.042)		(0.043)
early Fail x prior ability quartile=4		-0.144**		-0.110*
		(0.047)		(0.051)
Full set of controls	Yes	Yes	Yes	Yes
Observations	14387	14387	14429	14429
R-squared	0.579	0.580	0.545	0.545

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 standardized test scores. Each column reports results from an OLS model estimated for schools rated Fail in the years 2006 to 2009. Prior ability percentile calculated using age 7 (Key Stage 1) mathematics and writing tests. Students with missing age 7 test scores dropped from the sample. All regressions include full set of student- and school-level controls; see notes to Table 3 for details. Regressions in columns (1) and (3) also include prior ability percentile as control; regressions in columns (2) and (4) include prior ability quartile.

Table 6: 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)
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
Mean standardized test score	-0.41	-0.30	-0.79	-0.39	-0.53
	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)
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
Mean standardized test score	-0.42	-0.30	-0.85	-0.39	-0.58

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 standardized test scores. Each column reports results from an OLS model estimated for schools rated Fail in the years 2006 to 2009. All regressions include full set of student- and school-level controls; see notes to Table 3 for details. Mean standardized score reported in the final row of each panel is from the year before inspection.

Table 7: Medium-Term Effects

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

	Basic	Ability interactions		Subgroup analysis			
				Free lunch	Free lunch	First language	First language
		Linear	Non-linear	= 0	= 1	English	NOT English
	(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		-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. Combined mathematics and English outcome 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). All regressions include full set of student- and school-level controls; see notes to Table 3 for details. For columns (2) and (3) students with missing age 7 test scores are dropped from the sample.

Table 8: Effects for Moderate and Severe Fail Schools

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

	Moderate Fail Treatment		Severe Fail Treatment	
	(1)	(2)	(3)	(4)
OLS	Math	English	Math	English
early Fail	0.157**	0.074^{+}	0.097*	0.126**
	(0.038)	(0.040)	(0.042)	(0.040)
R-squared	0.49	0.53	0.48	0.54
Observations	8695	8596	7150	7136
No. of schools	211	211	175	175
Diff-in-diff	Math	English	Math	English
post x early Fail	0.111*	0.029	0.110**	0.147**
	(0.045)	(0.051)	(0.042)	(0.050)
post	0.063^{+}	0.135**	-0.001	-0.001
	(0.036)	(0.042)	(0.032)	(0.042)
R-squared	0.50	0.54	0.49	0.54
Observations	17642	17449	14529	14392
No. of schools	211	211	175	175

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. Moderate fail corresponds to the 'Notice to improve' fail rating. Severe fail corresponds to the 'Special measures' fail rating. See text for details. All regressions include full set of student background controls. OLS estimates in top half of table also include school-level controls; see notes to Table 3 for details. Diff-in-diff estimates include school fixed effects. 8 fail schools included in Table 3 but with no information on category of failure (moderate versus severe) are excluded from this estimation sample.

Table 9: Effect of a Fail inspection on Classroom Discipline

(Outcome: Are there any children in the class whose behavior prevents other children from learning? Yes = 1; No = 0; mean = 0.42)

	Experiment 1: Control group = 'Satisfactory' schools		Experiment 2: Control group = Later failed schools	
	(1)	(2)	(3)	(4)
Fail (2004 - 2007)	-0.055	-0.80^{+}	-0.067	-0.085
	(0.050)	(0.050)	(0.069)	(0.070)
Percent free lunch		0.004**		0.002
		(0.001)		(0.003)
Percent attaining English		-0.002		-0.004
and math target		(0.002)		(0.004)
Observations	872	872	204	204
R-squared	0.001	0.041	0.005	0.062

Notes: Robust standard errors reported in brackets; + , * and ** indicate significance at the 10%, 5% and 1% levels, respectively. Control group in columns (1) and (2) consists of teachers at schools attaining a 'Satisfactory' rating in the years 2004 – 2007; control group in columns (3) and (4) consists of teachers at schools failed in the years 2009 or 2010. 'Fail (2004-2007)' dummy turned on for schools failed 2004 – 2007. School-level controls (percent students eligible for free lunch and percent attaining English and math target at age 11) from 2004.

Appendix Table A1: 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.

Appendix Table A2: 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. 'Early Fail' dummy switched on for schools failed in the early part of the academic year (September to November); 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). Missing dummies included for student characteristics and age 7 test scores. All OLS regressions in Panel A include school controls (math and English attainment; per cent free lunch; per cent non-white; all from year before inspection); DID regressions (Panel B) include school fixed effects.

Appendix Table A3: 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 A2

Appendix Table A4: Effect of a Fail Inspection on Teacher Tenure, Years of Experience and School Curriculum

	Teacher tenure (years)		Teacher experience (years)		Curriculum: Math (hrs/week)		Curriculum: Literacy (hrs/week)		Curriculum: Phys Ed (hrs/week)	
Control group:	'Satisfactory' schools	Later failed schools	'Satisfactory' schools	Later failed schools	'Satisfactory' schools	Later failed schools	'Satisfactory' schools	Later failed schools	'Satisfactory' schools	Later failed schools
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Fail (2004 - 2007)	-0.198	-0.110	-0.045	0.099	-0.104	-0.059	0.294	0.372	-0.080	-0.036
	(0.825)	(1.016)	(1.077)	(1.473)	(0.084)	(0.083)	(0.267)	(0.319)	(0.071)	(0.108)
Percent free lunch	0.031	0.025	-0.018	0.004	-0.001	0.000	-0.000	0.002	-0.004	0.002
	(0.023)	(0.049)	(0.029)	(0.065)	(0.002)	(0.003)	(0.004)	(0.010)	(0.002)	(0.003)
Percent attaining English and math	0.052*	0.039	0.045	0.084	-0.007	-0.007	-0.007	0.009	-0.001	0.009
	(0.026)	(0.044)	(0.036)	(0.087)	(0.005)	(0.004)	(0.007)	(0.014)	(0.004)	(0.008)
Observations	834	201	823	195	671	157	658	154	642	150
R-squared	0.010	0.016	0.019	0.022	0.017	0.050	0.016	0.035	0.007	0.024

Notes: Robust standard errors reported in brackets; + , * and ** indicate significance at the 10%, 5% and 1% levels, respectively. The control group for the column labelled 'Satisfactory schools' consists of teachers at schools attaining a 'Satisfactory' rating in the years 2004 – 2007; control group for the column labelled 'Later failed schools' consists of teachers at schools failed in the years 2009 or 2010. 'Fail (2004-2007)' dummy turned on for schools failed 2004 – 2007. School-level controls (percent students eligible for free lunch and percent attaining English and math target) from 2004.