APPENDICES – NOT FOR PUBLICATION

Appendix A: 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 compression 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.[1]

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,

[1] 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.

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.[2] The trade-off is the potential for increased gaming behavior associated with the objective performance measure.

The empirical analysis in the section 3 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 in the main text, 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.[3]

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

[2] 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.

[3] 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. See also Griffith and Neely (2009) on the interaction between managerial experience and balanced score card incentives in allocating additional marginal effort.

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.

Appendix B: Further Robustness Tests

This Appendix investigates the distribution of inspections by month of inspection (B1); the effect of treatment by month of inspection (B2); whether the main results in the paper are driven by the possibility that inspectors respond to temporary dips in quality (B3); whether the effects of a Fail inspection persist beyond the actual year of inspection (B4); and reports on further results investigating strategic or dysfunctional response by teachers (B5).

Appendix B1: Distribution of inspections by month of inspection

Appendix Figure 1 shows the distribution of inspections across the school year. Apart from dips in the months with the longer school holidays (December, April and July) there appears to be a roughly even spread of inspections across the different months. Appendix Figure 2 shows the distribution of inspection grades by month of inspection. This shows that the quality of schools is more or less evenly balanced over the academic year. There is some tendency for proportionately fewer schools to be failed in the latter part of the year (e.g. around 6 per cent failure rate in September, October and November, versus around 5 per cent in June). However, for the main analysis in the paper, the key issue is whether *selection into treatment* (a Fail rating) is undertaken on a consistent basis, ensuring comparability of early and late treated schools. The descriptive analysis presented in section 3 suggests that this is indeed the case. (It is further boosted by the fact that the inspectorate rarely changes the 'rules of the game' midway through an academic year.)

Appendix B2: Effect of treatment by month of inspection

The main analysis in the text defines the treatment group (early failed schools) as those schools failed in the first three months of the academic year (September, October and November). This appendix undertakes this analysis month-by-month. Appendix Figure 3 shows the treatment effects, along with 95% confidence intervals. In each case schools inspected and failed in a given month – the treatment group – are compared with the control group of schools, i.e. those schools inspected and failed between mid-May and mid-July.

The general pattern from these charts is that of declining treatment effects, with effects peaking at some point between September and December and broadly declining thereafter. This pattern is consistent with the hypothesis that the longer a school has between the fail treatment and the May test, the larger the impact of treatment. Note additionally that the standard errors are quite large and effects are generally only significant in the first three months of the academic year.

Appendix B3: Robustness of identification strategy to temporary dips in quality

This appendix addresses the possibility that inspectors may be responding to temporary dips in quality, unobserved by the analyst, around the time of inspections. If this is so, then the concern might be that schools failed early in the academic year would have recovered, relative to late failed schools, by the time of the May test, even in the absence of a fail inspection.[4]

In order to shed some light on this issue I estimate the effect of a fail rating separately for schools which are more or less likely to be assigned a fail on the basis of the performance measure observed by the analyst, i.e. past test scores, as well as other observable characteristics. As highlighted in section 2, inspectors place substantial weight on test scores.

[4] Note, however, that this interpretation is hard to square with the heterogeneous effects – large gains for low ability students – reported in the main text.

If the decision rule inspectors use to rate schools places some weight on test scores and some on the qualitative findings during an inspection visit, then some schools may be failed largely on the basis of test scores, whilst another group may be failed largely on the basis of the qualitative evidence.[5] Thus if the latter group of schools are more likely to be selected for fail on the basis of unobservables, which confound the causal estimates of the fail rating, then the estimated effect for these schools ought to be higher than that for the former group of schools where, arguably, unobservables are less important for selection into the fail treatment.

Columns 1and 2 of Appendix Table 6 report the effects of a fail rating for the subset of schools which are predicted to – on the basis of observed test scores and other observable characteristics – have a relatively low probability of being failed; schools used for the analysis in columns 3 and 4 have a relatively high fail probability.[6] These results show that the estimated effects for those schools which are more likely to be failed on the basis of past test score performance (columns 3 and 4) are *not* significantly smaller than for those schools which are more likely to be failed on the basis of school quality attributes unobserved by the econometrician (columns 1 and 2).[7] Thus, on this evidence at least, the estimated effects of a fail rating using the early-late inspected contrast are not subject to bias arising from unobserved school quality.

[5] Such a decision rule also accords with the political economy of inspections: inspectors must be able to justify their judgements when teachers and possibly parents as well as local politicians react against a school being rated fail. Inspectors would then need to point to evidence supporting their view. This evidence may be in terms of the hard test score data, or the softer more qualitative evidence or perhaps both.

[6] A logit model of the probability of being failed is estimated for all schools (fail and non-fail) using the school's test score performance from the year before the inspection, percent of students eligible for free lunch, log enrollment and local education authority dummies. The coefficient on test scores is highly significant with an average marginal effect of -0.0028, implying that a decline of 10 percentiles in the school-level test performance distribution leads to around 3 percentage point rise in the probability of being failed. For fail schools, the estimated probability of fail varies between 0.2 percent and 49.2 percent. The mean estimated probability of fail for schools in columns 1 and 2 (columns 3 and 4) of Table 10 is 7.2 percent (24.1 percent). The results reported in Table 10 are little changed if test scores from the previous three years, instead of just the last year, are used for the model used to predict fail.

[7] For Mathematics, the OLS and DID estimates suggest that the estimates for schools with an above median probability of fail are around 20 percent smaller than for schools with a below median probability of fail; however, these differences are not statistically significantly different. For English the results suggest that, if anything, the magnitude of the effects is *larger* for the former group (OLS estimates in columns 2 and 4).

Appendix B4: Effects in the year after inspection

The key result in this paper points to large effects in the year of inspection for schools failed early in the academic year. A natural question is whether these treatment effects persist beyond this initial period. In this appendix I investigate this question by comparing outcomes for early and late failed schools in the year after inspection. The models in Appendix Table 4, Panel A are estimated using data from the year after inspection. These show that although the gap between early and late failed schools closes somewhat, early failed schools continue to outperform later failed schools. The DID models in Panel B are estimated using data from the year before inspection and from the year after inspection. The post dummy indicates that there is a large pick up in performance for the 'control' group (compare these with the much smaller gains relative to the pre-inspection year reported in Table 2). One interpretation of this finding is that the late failed schools improve in the twelve months following the fail rating. The additional effect for the treatment group (as indicated by the 'post x early Fail' row) suggests that strong performance is sustained at early failed schools in the year after inspection.

Appendix B5: Further tests for dysfunctional response by teachers

A further test for the presence of dysfunctional teacher response is to investigate whether there is a response in terms of the number (or proportion) of students taking the test. If teachers strategically remove weaker students from the testing pool, then this may be detected

in this outcome.[8] Appendix Table 5 reports results from school-level regressions comparing early and late inspected schools for the outcome: number of students taking the year-11 KS2 test in the year of inspection. These demonstrate that the effects of this outcome are small and statistically insignificant. This finding lends further support to the idea that inspectors help ameliorate the effects of strategic behaviour in this high stakes setting.

[8] Note that Hussain (2013) shows that there is little effect in the total number of students *enrolled* in school in the year of inspection.

Appendix C: Quantile Treatment Effects

One 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 particular students within these ability groups.

Appendix Figure A1 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 A1.

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.

Results

In order to assess how the conditional distribution of test scores is affected by treatment at each quantile $\tau \in [0,1]$, ehe following model is estimated:

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

where $Q_\tau(y_{is}|\,.)$ is the τ^{th} conditional quantile function and δ_τ is the quantile treatment effect at quantile τ. Appendix Figure A2 plots δ_τ as well as the associated 95 percent confidence interval. For mathematics, Panel A of Figure A2 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 A2 tends to reject the hypothesis that teachers act strategically to raise performance of students on the margin of attaining the official government target.[9]

Importantly, the pattern of treatment effects across quantiles reported in Figure A2 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 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. The incentives they face suggest that they may target effort towards marginal students within these subgroups.

Figures A3 and A4 plot the quantile treatment effects within each prior ability quartile, for mathematics and English, respectively. These demonstrate that there is a great

[9] 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.

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 A3, Panel A). For English, Panel A of Figure A4 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 A3 and A4 is that teachers target the students on the margin of attaining the performance threshold.[10] The evidence does not generally support this hypothesis. First, it should be noted that Panel A, Figure A3 and Panel A, Figure A4 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.[11]

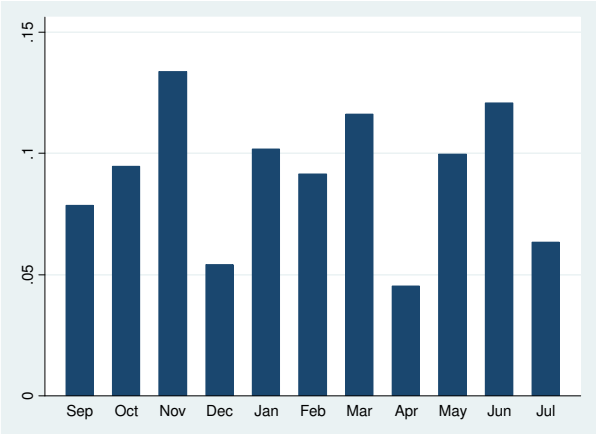
[10] As indicated in Appendix Table A1, these students are likely to be in the higher quantiles of the test score distribution: Appendix Table A1 shows that in the year before the fail inspection 23 per cent and 33 per cent of students reach the mathematics and English threshold, respectively. Thus, if teachers successfully target the marginal students, we would expect to see the largest gains at quantiles 0.77 (mathematics) and 0.67 (English).

[11] 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.

Appendix D: Teacher Tenure and Curriculum

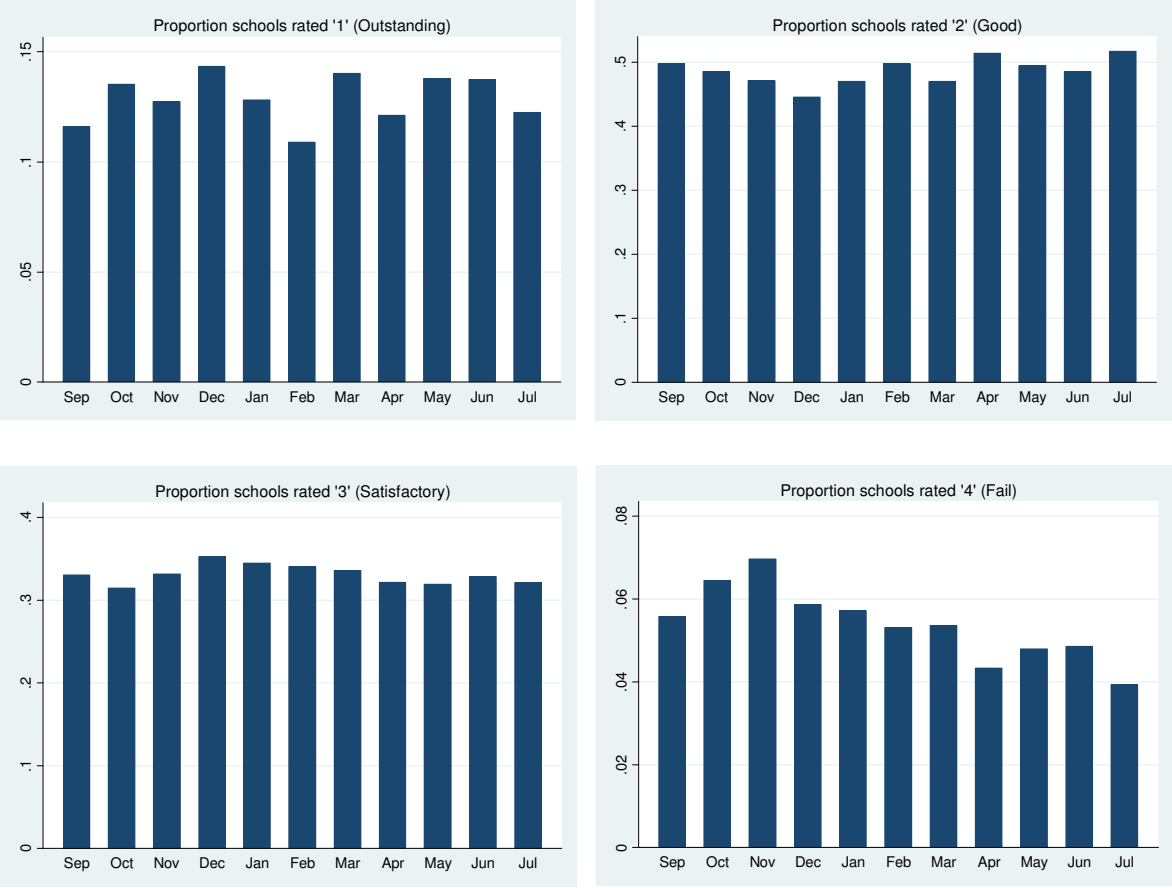
Appendix Table 7 employs the same empirical strategy and data used to estimate the effect of a fail rating on classroom discipline (section 5.4) in order to shed light on a number of related outcomes: teacher tenure and experience (columns 1 to 4) as well as number of hours devoted each week to English, 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 maths (English) and a 4 percent decline in hours devoted to physical education.

Appendix Figure 1: Distribution of inspections in the academic year



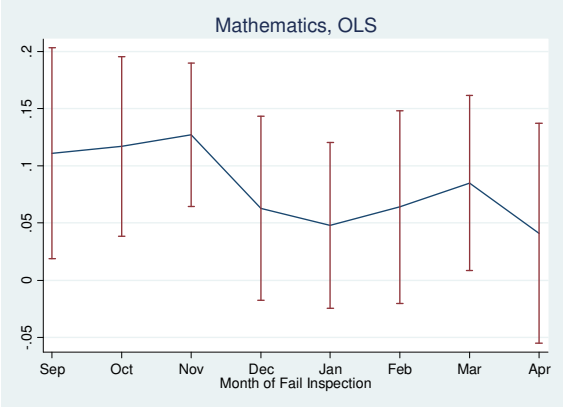
Note: Based on inspections over the period September 2006 - July 2009.

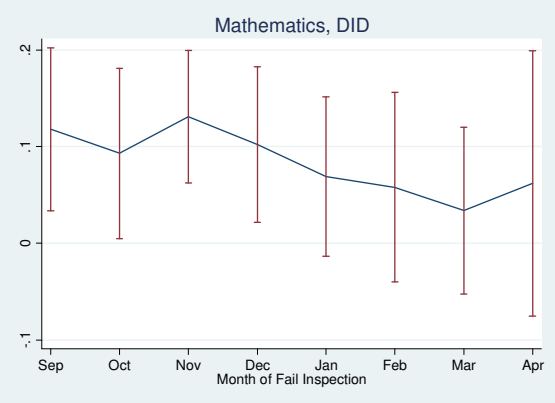
Appendix Figure 2: Fraction of schools receiving a given rating, by month of inspection

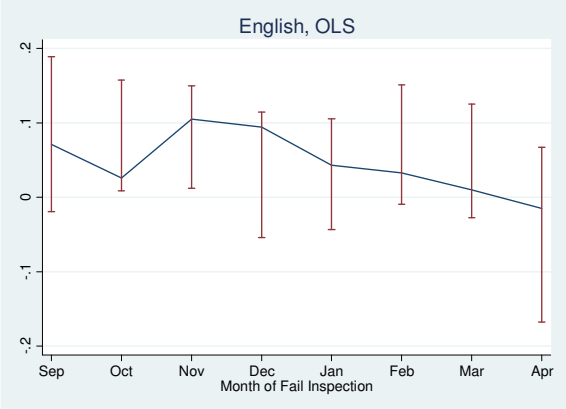


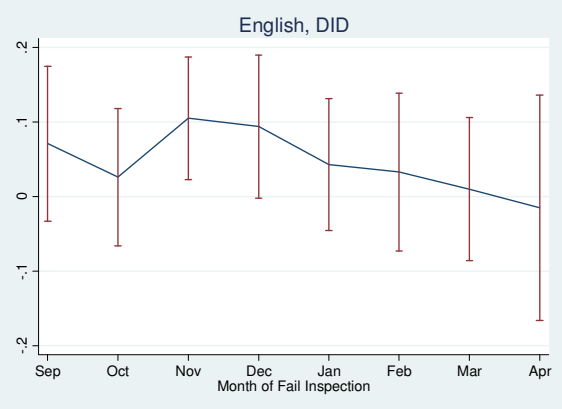
Note: Based on inspections over the period September 2006 - July 2009. Each bar represents the fraction receiving a given rating in that month. E.g. 11% (50%) of all schools inspected in September received an Outstanding (Good) rating.

Appendix Figure 3: Plot of treatment effects, by month of inspection

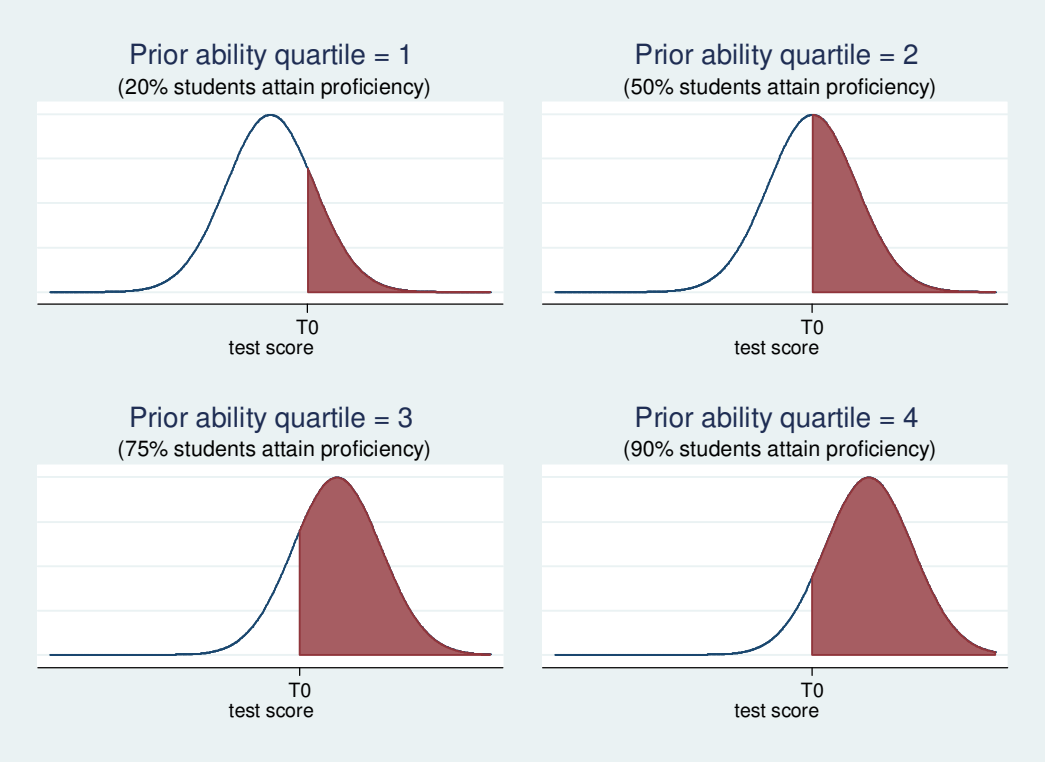








Appendix Figure 4: Stylized Example of Distribution of Students Passing Proficiency Thresholds, by Quartile of Prior Ability

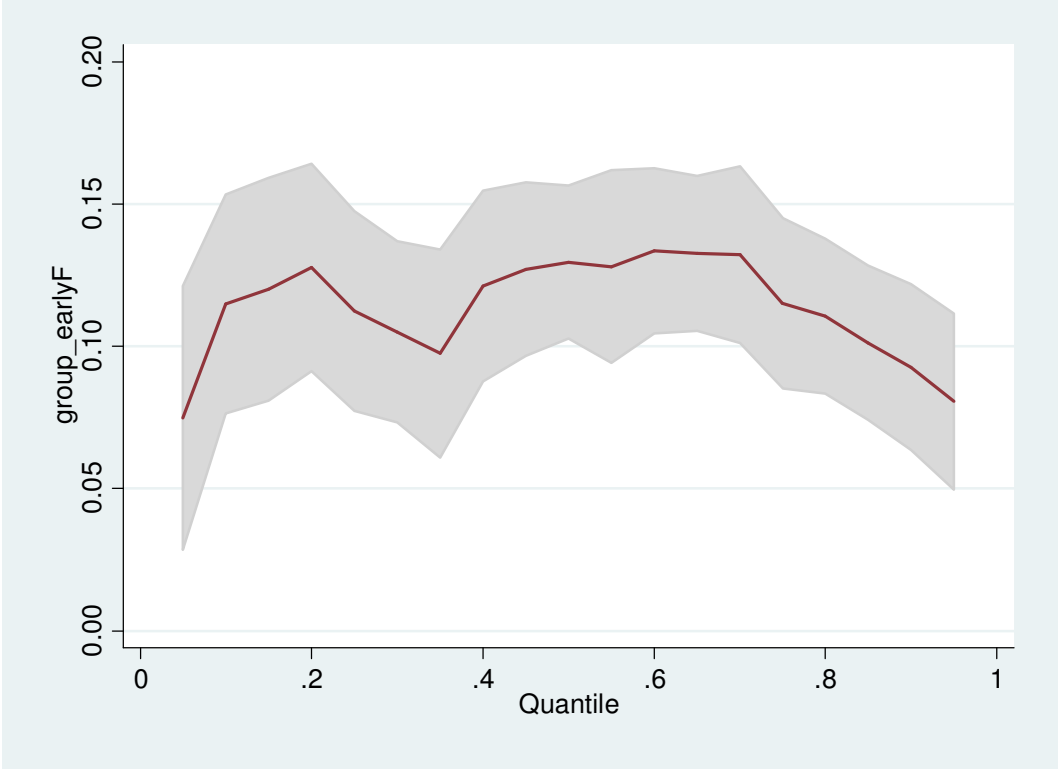


Note: 'T0' denotes the official proficiency threshold.

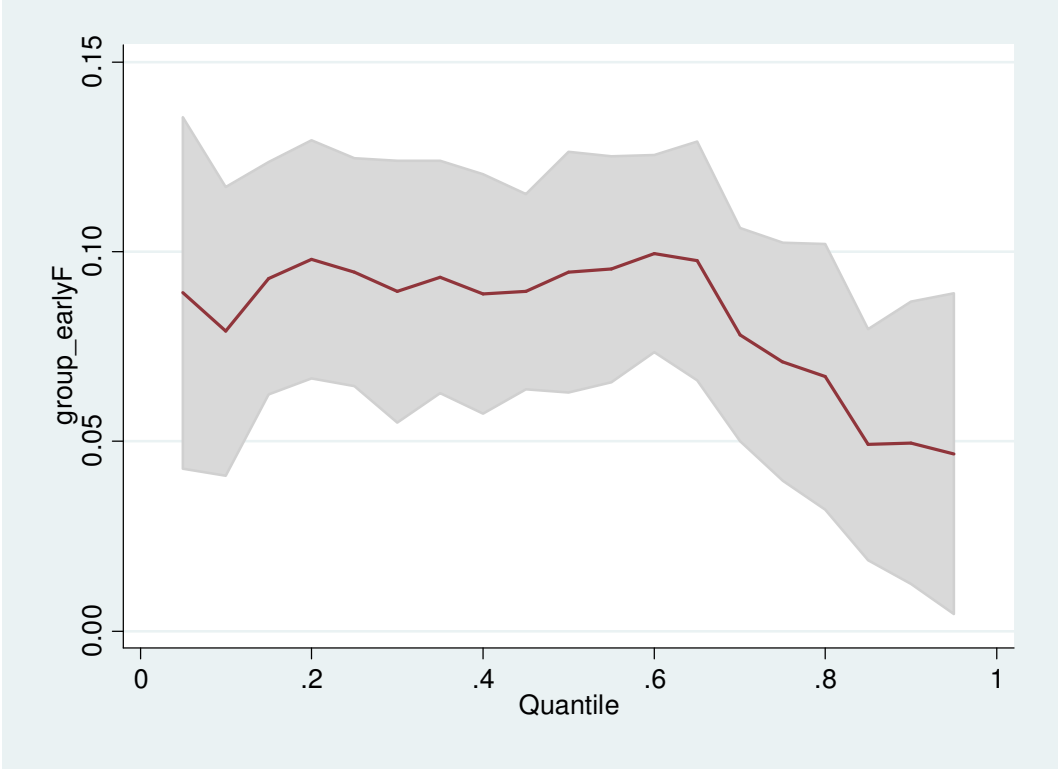
Appendix Figure 5: Quantile Regression Estimates of the Effect of a Fail Inspection

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

Panel A: Mathematics



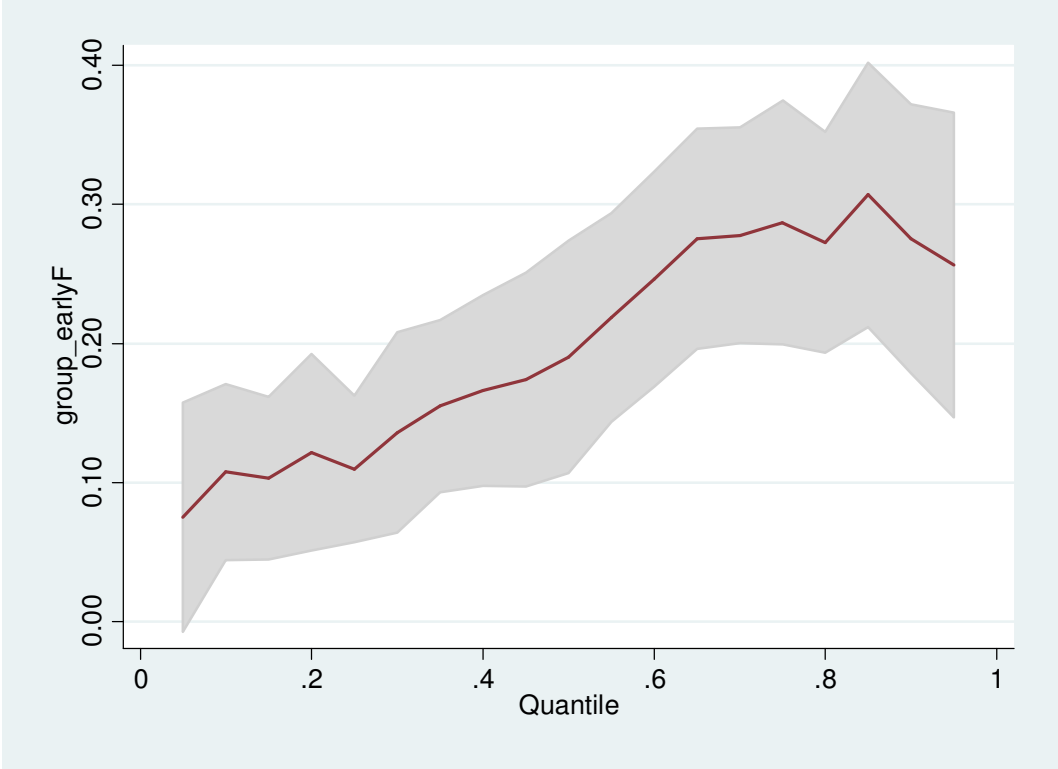
Panel B: English



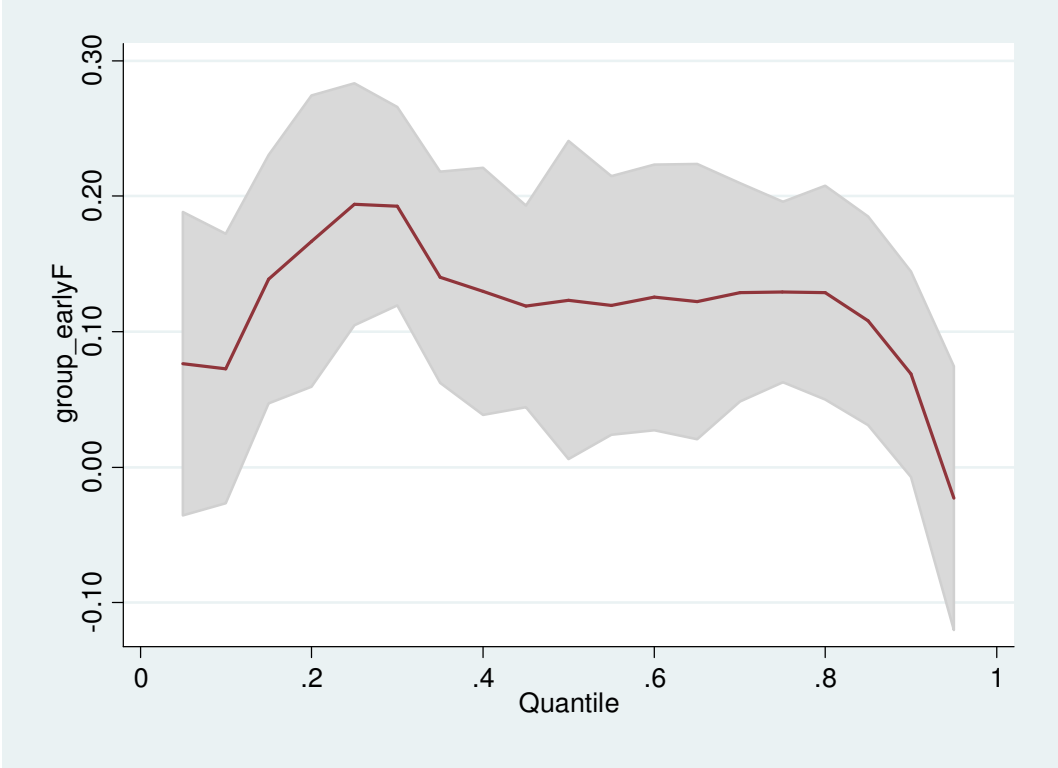
Appendix Figure 6: Quantile Regression Estimates: by Prior Ability Quartile (Mathematics)

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

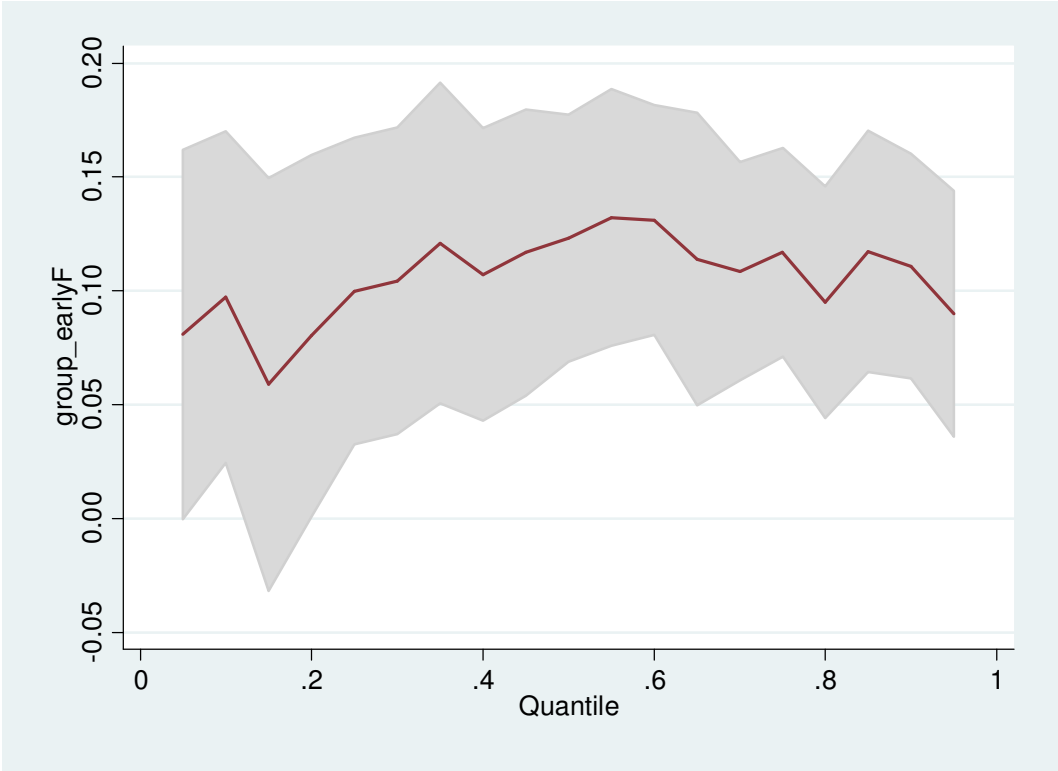
Panel A: Prior ability (age 7 test) quartile = 1



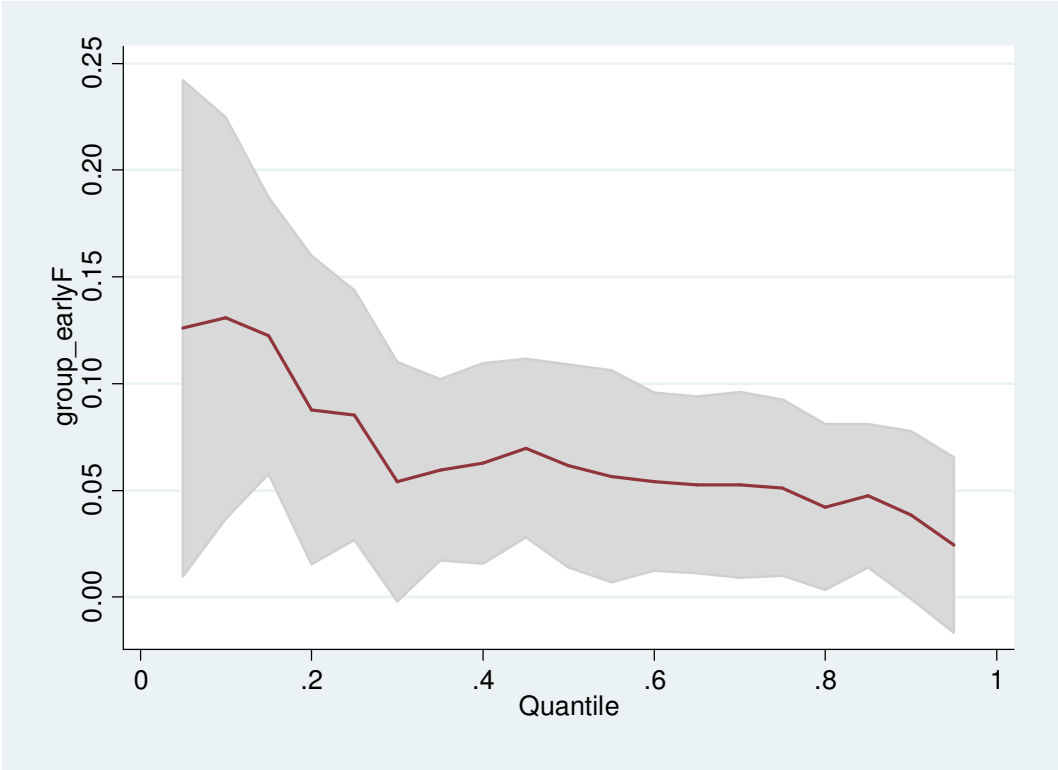
Panel B: Prior ability (age 7 test) quartile = 2



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



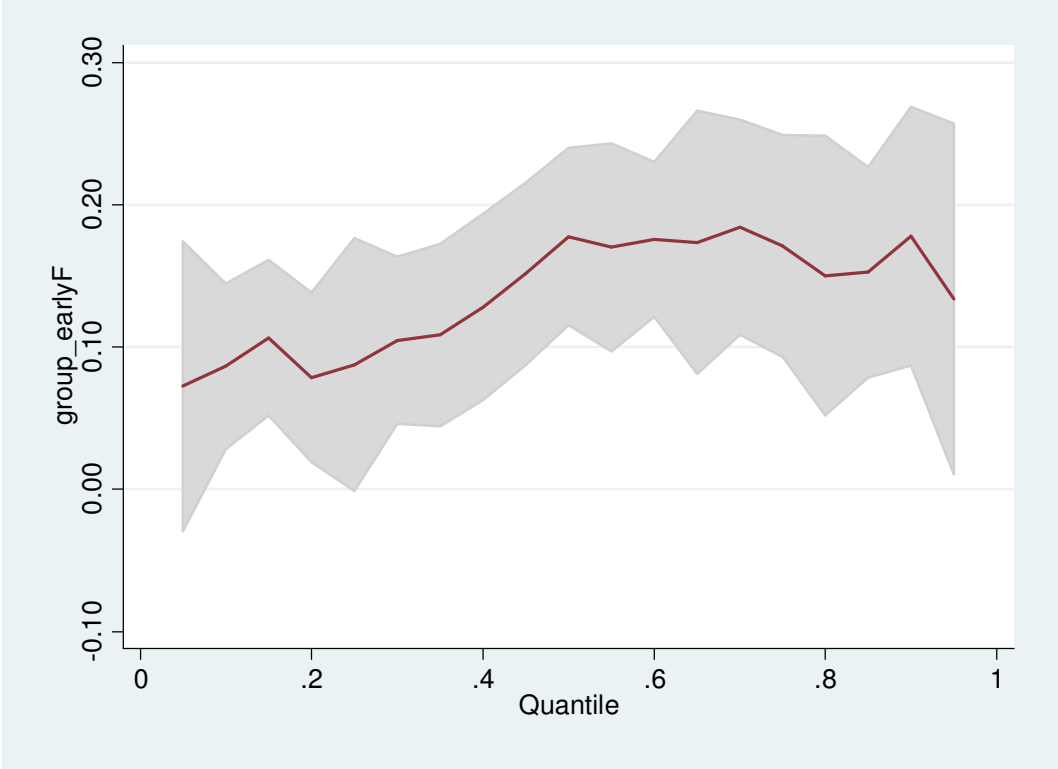
Panel D: Prior ability (age 7 test) quartile = 4



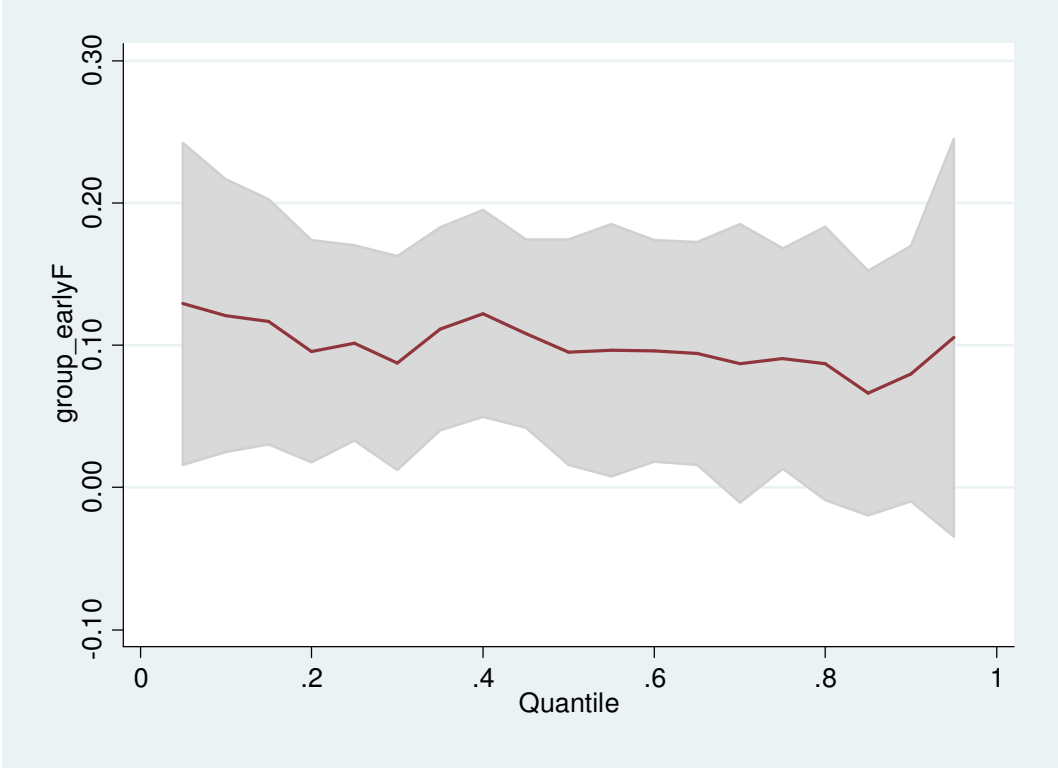
Appendix Figure 7: Quantile Regression Estimates: by Prior Ability Quartile (English)

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

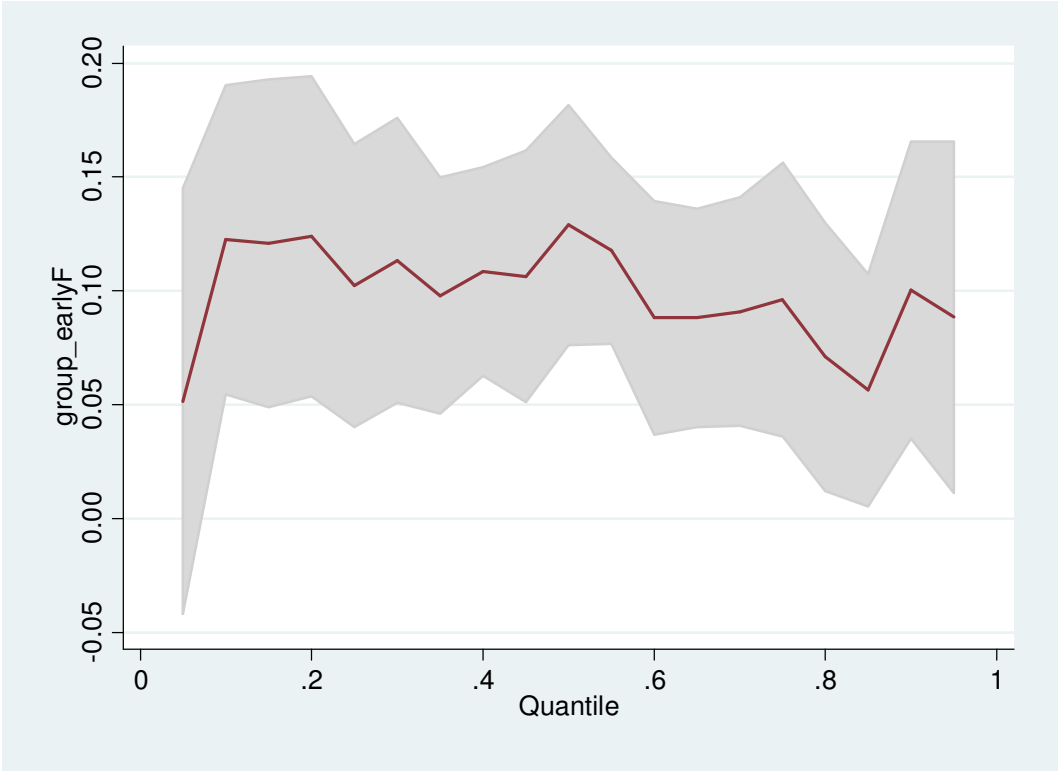
Panel A: Prior ability (age 7 test) quartile = 1



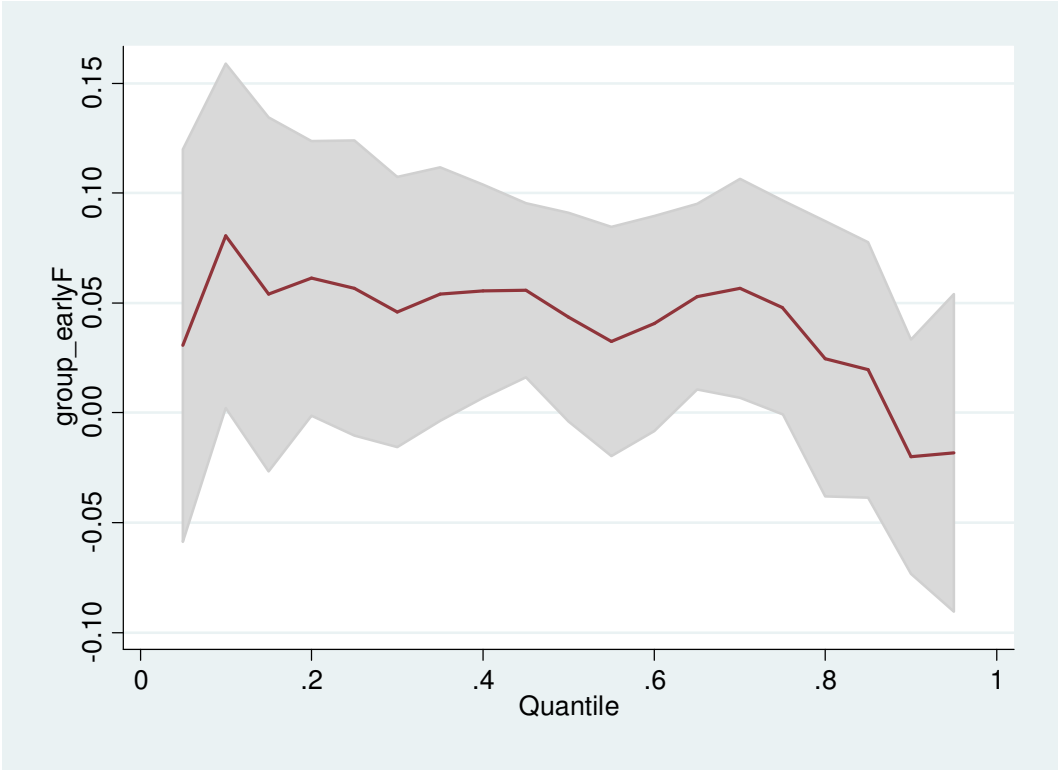
Panel B: Prior ability (age 7 test) quartile = 2



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



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



Appendix Table 1: OLS and Difference-in-Differences Estimates of the Effect of a Fail Inspection: Mathematics

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 2: OLS and Difference-in-Differences Estimates of the Effect of a Fail Inspection: English

,,

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 A3: 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 4: Test Scores in the Year after Inspection

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

Panel A: OLS	**Mathematics**			**English**		
	(1)	(2)	(3)	(4)	(5)	(6)
early Fail	0.074*	0.081*	0.081*	0.063	0.064	0.062*
	(0.033)	(0.034)	(0.032)	(0.033)	(0.034)	(0.031)
Student characteristics	No	Yes	Yes	No	Yes	Yes
Age-7 test scores	No	No	Yes	No	No	Yes
R-squared	0.032	0.258	0.441	0.044	0.329	0.502
Observations	15,475	15,475	15,475	15,142	15,142	15,142
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	0.068	0.067	0.083*	0.052	0.040	0.041
	(0.036)	(0.036)	(0.034)	(0.040)	(0.040)	(0.040)
post	0.111**	0.130**	0.132**	0.117**	0.148**	0.189**
	(0.033)	(0.028)	(0.031)	(0.032)	(0.032)	(0.033)
Student characteristics	No	Yes	Yes	No	Yes	Yes
Age 7 test scores	No	No	Yes	No	No	Yes
R-squared	0.082	0.245	0.474	0.097	0.313	0.526
Observations	32,588	32,588	32,588	32,026	32,026	32,026
Number of schools	394	394	394	394	394	394

Notes: Standard errors reported in parantheses; * 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 after inspection and off for the year before inspection. 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.

Appendix Table 5: Number of students tested

	OLS		DID	
	(1)	(2)	(3)	(4)
early Fail	1.37	1.42	0.32	0.36
	(2.254)	(2.243)	(2.356)	(2.325)
post			-2.13	-1.57
			(2.736)	(2.709)
post x early Fail			1.05	1.03
			(3.259)	(3.216)
School characteristics	No	Yes	No	Yes
Mean dep. variable	42.2	42.2	42.8	42.8
R-squared	0.002	0.029	0.002	0.032
Observations	394	394	788	788

Notes: Standard errors reported in parantheses; * and ** indicate significance at the 5% and 1% levels, respectively. S.e.'s clustered at the school level. Dependent variable: number of students taking the Key Stage 2 test. 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. School controls: per cent students receiving free lunch; per cent non-white.

Appendix Table 6: Effects for schools with low and high predicted chance of failure

(Outcome variable: age 11 national standardized test score)

	Below median probability of fail		Above median probability of fail	
	(1)	(2)	(3)	(4)
OLS	Math	English	Math	English
early Fail	0.139**	0.082*	0.111**	0.087*
	(0.035)	(0.038)	(0.040)	(0.042)
R-squared	0.498	0.551	0.478	0.518
Observations	8,213	8,137	8,404	8,365
No. of schools	197	197	197	197
Diff-in-diff	Math	English	Math	English
post x early Fail	0.114**	0.063	0.095*	0.062
	(0.040)	(0.049)	(0.044)	(0.051)
post	-0.036	0.010	0.107**	0.155**
	(0.031)	(0.040)	(0.035)	(0.041)
R-squared	0.506	0.557	0.489	0.538
Observations	16,629	16,491	17,101	16,895
No. of schools	197	197	197	197

Notes: Standard errors reported in parantheses; * and ** indicate significance at the 5% and 1% levels, respectively. S.e.'s clustered at the school level. All regressions include full set of student background controls. See notes to Table 3 for details. Sample stratified by probability of school failing the inspection. Columns 1 and 2: predicted probability of failure is below median for all 394 schools. Columns 3 and 4: predicted probability of failure is above this median. School-level regressors used for logit model predicting failure: previous year's test performance and proportion of students receiving free lunch as well as local education authority dummies.

Appendix Table 7: 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.

Appendix Table 8: School Characteristics Prior to Fail Inspection, Treatment and Control Schools, Inspections 2005/06 to 2008/09

	Late inspected schools (control)	Early inspected schools (treated)	t-test of difference (p-value)
Month of inspection	6.14	10.21	0.000**
	(0.05)	(0.05)	
Year of previous inspection	2002.8	2000.1	0.000**
	(0.17)	(0.13)	
% students entitled to free school meal	24.4	23.8	0.720
	(1.50)	(0.99)	
% students white British	76.0	80.1	0.146
	(2.51)	(1.51)	
Previous inspection rating (Outstanding = 1; Good = 2; Satisfactory = 3)	2.39	2.38	0.271
	(0.05)	(0.04)	
Age 11 standardised test scores, year before Fail			
Mathematics	-0.42	-0.44	0.627
	(0.03)	(0.02)	
English	-0.43	-0.44	0.747
	(0.03)	(0.02)	
Number of schools	136	258	

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