School inspections: can we trust Ofsted reports?

Ofsted inspections of schools have been a central feature of state education in England for nearly 20 years. Research by **Iftikhar Hussain** *explores the validity of the school ratings that Ofsted produces, the impact of a fail rating on subsequent pupil performance and the extent to which teachers can 'game' the system.*

How to ensure that teachers, doctors and other public sector employees act in the best interests of the users of public services is a challenge facing governments around the world. Many have responded by making increased use of objective performance targets, such as pupil test scores for schools and patient waiting times for hospitals. Empirical evidence suggests that organisations given incentives to perform on such measures often 'game' the system, subverting the intention behind the target. For example, schools might try to raise their average test score by excluding lower ability pupils from the test. Or when given a target for pupils to attain a certain level in key stage tests, teachers might focus on pupils on the margin of this achievement level, at the expense of both high and low ability pupils.

In England, and in the UK more broadly, top-down performance targets have often been complemented by inspection regimes. Examples include the school inspection system run by Ofsted (the Office for Standards in Education), and the Care Quality Commission for the healthcare sector. In fact, the UK is something of a world leader in the area of inspections.

The subjective nature of the performance evaluation undertaken by inspectors holds the promise of 'measuring what matters'. But inspections may open up a whole new can of worms. In particular, just like targets, inspections are top-down and, what's more, the officials who lead them may be prone to biases and prejudices. Left to their own devices to exercise judgement as they see fit, it is unclear whether inspectors will act in the best interest of pupils and parents. Neither pro- nor anti- inspection arguments have been backed by hard empirical evidence and this research seeks to remedy this gap.

New evidence on inspection systems

In the absence of previous empirical evidence – from the UK or elsewhere – this study uses the case of Ofsted inspections of state schools in England to provide some evidence on the effectiveness of such subjective performance evaluation systems. The first question addressed is whether inspection ratings provide any extra information on school quality, over and above what is already available in the public sphere.

This 'validity test' is implemented as follows: are inspection ratings correlated with underlying school quality measures – constructed from teenage pupils' survey reports of teacher practices as well as parental satisfaction – after controlling for standard observable school characteristics, such as test score rankings and the proportion of pupils eligible for free school meals? If they are, then inspection ratings have the potential to play an important role in providing information for parents considering which school they would like their children to attend.

The next question addressed is whether pupil test scores improve following a fail inspection. This is a thorny empirical problem because it suffers from the classic problem of 'mean reversion' – the idea that a failed school's test performance would have improved (reverted to the mean) even in the absence of a fail inspection. Relatedly, whether schools are able to game the system and artificially boost performance following a fail inspection is also addressed. The post-fail incentives to game are strong and the prior evidence – from England and elsewhere – suggests that when teachers are put under pressure to raise pupil test scores, they may well resort to such strategies.

The nature of Ofsted school inspections

Broadly speaking, Ofsted inspections of schools have three main objectives: first, to offer feedback and advice to the head and other teachers; second, to provide information to parents to aid their decision-making; and third, to identify schools that suffer from 'serious weakness'. Over the period covered by my research (2005/6 to 2008/9), schools were generally inspected once during an inspection cycle. 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.

For the visit, inspectors arrive at the school at very short notice (no more than a few days), which in theory should enable them to see the school as it 'really is' and limit disruptive 'window dressing' in preparation for the inspections. During the inspection, the inspectors collect qualitative evidence on performance and practices at the school. This involves classroom observations, in-depth interviews with the school leadership, examination of pupils' work as well as discussions with pupils and parents.

At the end of this process, the school is given an explicit headline inspection rating – 1 (outstanding), 2 (good), 3 (satisfactory) or 4 (unsatisfactory, also known as a fail rating) – and the inspection report is made available to parents and posted on the internet.

Do inspection ratings convey any new information on quality?

Previous research suggests that inspectors' findings are reliable, in that two inspectors independently observing the same lesson will come to very similar judgements about the quality of classroom teaching. The issue 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. There is almost no empirical evidence on this question.

I construct a measure of underlying school quality from 14-year-old pupils' survey responses to questions about teacher behaviour and practices. These data come from the Longitudinal Survey of Young People in England, a major survey supported by the Department for Education. The survey asks six questions on how likely teachers are to: take action when a pupil breaks rules; make pupils work to their full capacity; keep order in class; set homework; check that any homework that is set is done; and mark pupils' work. Further analysis using pupils' reports of school discipline as well as parents' reports of satisfaction levels yields very similar results to those reported here for the teacher practices outcome.

A composite pupil-level score of teacher practices at the school is computed by taking the average of the responses to these six questions. The validity test is then undertaken by asking the following question: can inspection ratings help distinguish between (or forecast) schools with good and poor teacher practices? The critical issue is whether inspection ratings summarise information about underlying school quality that is not already available in the public sphere. My findings demonstrate that on this measure at least, inspectors appear to be doing a reasonable job. Even after controlling for things like test scores and the socio-economic background of pupils, inspection ratings appear to have substantial power in predicting underlying quality.

These results suggest that parents who are looking for a good school ought to place at least some weight on inspection ratings. The actual weight will depend in part on how much information parents already have about the relevant schools, beyond publicly available information such as test scores.

The effect of a fail inspection

Turning to the effects of a fail rating on a school's future test scores, the data show that pupils' performance on key stage tests improves following a fail inspection. But whether this is a consequence of the fail inspection or simply a bounce back after a year or two's bad luck is a difficult question to answer. To make progress on this issue, I exploit a design feature of the English primary schools' testing system to solve this mean reversion problem. Key stage 2 tests for Year 6 (age 11) pupils are administered in the second week of May in each year. These tests are marked externally, and results are released to schools and parents in mid-July.

The 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 inspectors know the outcome of the tests. Thus the May test outcome for these schools is not affected by the subsequent fail; neither do inspectors select them for failure on the basis of this outcome.

By comparing schools failed early in the academic year – September, say – with schools failed in June of the same academic year, I can isolate mean reversion from the effect of the fail inspection. If there is any rebound in test scores independent of the fail inspection, we should see it in the schools failed in June. As a check on this approach, a comparison of observable characteristics prior to the fail rating for schools failed in June (the 'control' group) and schools failed in the early part of the academic year (the 'treatment' group) reveals negligible differences between the two groups.

The results show that a fail inspection leads to significant improvements in test scores. Furthermore, the largest gains are for pupils scoring low on the prior (age 7) key stage 1 test. The gains are large compared with the effects of other policy interventions that pupils might experience, such as having higher quality teachers or attending a school with higher attainment levels.

These findings are consistent with the view that the children of low-income parents – who are, arguably, the least vocal in holding teachers to account – benefit the most from inspections. Such evidence 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.

Evidence on gaming behaviour

Finally, this research also looks into possible strategic behaviour by teachers. I find little evidence to suggest that schools failed by the inspectors are able to inflate their pupils' test performance by gaming the system. First, teachers do not exclude low ability pupils from the test-taking pool. Second, although the evidence on whether teachers target pupils on the margin of attaining the official proficiency level (level 4 on the key stage 2 test) is mixed, I find no evidence to suggest that some groups (say, very low or very high ability pupils) are adversely affected by the fail inspection. Third, although test gains fade somewhat over time, there is evidence to suggest that for some pupils, the gains last into the medium term, even after they have left the failed primary school. This suggests that teachers inculcate real learning and not just test-taking skills in response to the fail rating.

These findings on strategic behaviour are in stark contrast to a significant body of evidence demonstrating dysfunctional responses to test-based performance evaluation in other settings. My interpretation of these results is that by subjecting schools to close scrutiny, inspectors may play an important role in limiting such distortionary activities.

Conclusions

What are the broader lessons from this study? The findings are particularly noteworthy given the indications from past research that subjective assessments may give rise to various biases. For example, there is evidence to suggest that subjective evaluations of workers may lead to 'leniency' and 'centrality' bias in private firms (Prendergast, 1999). And evidence from the public sector points to staff indulging their preferences when allowed to exercise discretion rather than following formal rules (Heckman et al, 1996).

Although such biases in school inspectors' behaviour cannot be ruled out, this research demonstrates that the inspection system appears to be effective along the following two dimensions: first, inspectors produce ratings that make it easier to distinguish between more and less effective schools; and second, they are able to identify poorly performing schools, leading to test score gains.

One important feature of the English school inspection system is that the key output produced by the inspectors – an inspection rating and report – is available for public consumption on the internet. Consequently, inspectors' decisions are themselves 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.

This article summarises 'Subjective Performance Evaluation in the Public Sector: Evidence from School

Inspections?' by Iftikhar Hussain, Centre for the Economics of Education Discussion Paper No. 135.

Note: This article gives the views of the author, and not the position of the British Politics and Policy blog, nor of the London School of Economics. Please read our comments policy before posting.

About the author

Iftikhar Hussain is a lecturer in economics at the University of Sussex and a research associate in CEP's programme on education and skills.

You may also be interested in the following posts (automatically generated):

1. Poor pupil performance is more about poverty than school quality. We must ensure our schools work for poor children in all places. (18.6)
2. Not everyone is Sir Alex Ferguson: systemic constraints, not just individual leadership, are responsible for persistent differences in school standards (18.6)
3. The 'NEET' category emphasises participation regardless of its content or value. Focusing on progression and achievement can help set clearer goals for school leavers. (16.5)
4. Failing before school: the gap between children in high and low income families has led to a dangerous disadvantage with those in poorer families more likely to suffer from serious social and emotional problems. (16.3)