Speech: HMCI's monthly commentary: March 2017

Two years ago, Ofsted said it would start testing inspection reliability. This was, in part, a response to sector voices, who quite reasonably thought we should know how consistent inspection judgements are. All our inspectors are thoroughly and repeatedly trained, and all our inspections are quality-assured, giving us some confidence that what are ultimately human judgements are made properly and consistently. Yet nothing beats hard evidence from a well designed trial.

At the same time, our short inspection framework was being developed. We did not want to miss the opportunity to evaluate this new type of inspection from the start. The study was therefore designed to answer a single question: were the decisions about whether short inspections should or should not convert to full inspections being made consistently by different inspectors? There were many more questions that could have been asked, but the study was a first step towards a more evidence-based approach to the development of inspection.

Today, I am pleased to set out the findings in this commentary, based on the full report, which is published today.

The basic design of the study was a comparison of the outcomes from 2 inspectors carrying out a short inspection of the same school independently, on the same day. So what did we learn?

First, it appears we are breaking new ground here. Some reliability studies have been done before, but they were usually looking at specific parts of inspection, such as lesson observation. They have not looked at the whole inspection process from start to finish. Our report contributes new findings to the research literature.

Secondly, carrying out this study was surprisingly difficult. The complexities included:

- getting the balance right between the live inspection and the study goal
- identifying ways to minimise bias and cross-contamination of inspector evidence gathering and thinking
- ensuring that inspectors and participating schools were fully prepared for simultaneous parallel inspections
- achieving a large enough sample of participating schools

Thirdly, and most importantly for everyone who is inspected, the study provides a welcome positive view of inspector consistency in the particular context studied. Of the 24 short inspections in our sample, inspectors agreed on the outcome in 22 cases. This indicates a high rate of agreement (92%) between these inspectors about the conversion decision.

Furthermore, in 1 of the 2 cases of disagreement, the disagreement was at the good/outstanding borderline and was resolved by the full inspection: 1 inspector's view was that conversion was unnecessary as the school remained good; the other had opted for conversion to collect further evidence to see if an outstanding judgement was justified. The outcome of the full inspection was that the school remained good. So in only 1 out of 24 cases might the final judgement have been different between the 2 inspectors, as both decided to convert to a full inspection for opposing reasons. Despite this, the outcome at the full inspection was that this school also remained good.

There are, of course, limitations to a small-scale exploratory study like this that need to be taken into account. The findings cannot be extrapolated across other types of inspections or all types of institution. For instance, the study looked only at short inspections of primary schools in a certain size range and it had a relatively small sample. Yet, as an initial attempt at evaluating reliability, these findings should provide some reassurance that the purpose of the short inspection model is being met and that inspectors made consistent judgements.

I suspect that, despite this encouraging result, most comment will be about the 2 cases where inspectors arrived at different decisions. We all know that there is low education system tolerance of variability in marking in exams. (See: 'The reliability programme: final report of the policy advisory group', Burslem, S. (2011). Coventry: Ofqual)

It is likely that this is the case with inspection, because of its high-stakes nature and, in particular, the consequences that can follow from a poor inspection outcome.

The imperative is rightly on Ofsted to ensure that our judgements are as reliable as possible. But a medical analogy may be helpful here: many kinds of clinical testing give both false positive results (where someone doesn't actually have the condition, but appears to) and false negatives (where someone has the condition but is not picked up by the test). Perfectly reliable tests are the exception, not the rule.

Turning back to education and social care, we know that inspection is a process based on human judgement to interpret and complement available data. We know a great deal about human judgement, and can work to minimise the impact of the limitations resulting from the various kinds of bias in human judgement, but we are unlikely ever to reach a position where perfect consistency can be guaranteed.

For one thing, we would not want to over-simplify inspection in the pursuit of consistency. A tick-box approach, for instance, might lead to improved reliability but would be a mechanistic approach to inspection that would

almost certainly undermine its validity. We need some degree of professional judgement to reflect the complexity and variety of institutions we inspect. This may well lead to experts disagreeing at times. It does not necessarily mean that 1 inspector or the other is wrong or that they made mistakes, as there are likely to be multiple decisions made on the areas to evaluate that can lead to legitimately different views.

So how can we increase reliability while recognising that inspectors cannot be clones?

The short inspection process attempts to do just that, as any disagreement between inspectors can be resolved once the short inspection converts to a full inspection. In the 2 cases in our sample where inspectors did not agree on the short inspection outcome, the follow-up inspection activity led to both schools remaining good. This is a small amount of evidence to suggest that the safety net at the end of the short inspection adds an extra layer of security to the final judgement. As such, it is likely that the conversion process is another mechanism that allows us to protect schools from the risk of unreliable inspector judgements. It certainly appears to be more secure than past attempts at light-touch inspection frameworks.

Of course, there are a number of assumptions here. While I have confidence that inspection frameworks, inspector training and quality assurance procedures mitigate the risks of inconsistency, we need to study the inspection judgements themselves, as well as the decisions around the conversion of short inspections.

As I have already mentioned, this study is just a first step towards a continuing programme of research into inspection. We should routinely be looking at issues of consistency and reliability. And even more importantly, we should be looking at the validity of inspection: is inspection succeeding in measuring what it is intended to measure? This is not an easy question, in part because validity is not an absolute: it depends on the purpose of the inspection.

We are beginning to shape up what this research programme should look like. But this is not a quick hit in which everything is sorted at once: rather, it will be a steady process in which questions are addressed systematically. Some of this may come through work on components of inspection rather than inspection in its entirety.

And as part of that process, we will continue to work with outside academics and other experts, as well as those at the receiving end of inspection, to help shape the approach we take. It is really valuable to have the right level of challenge in this kind of work, as well as specialist expertise.

And finally, in this context, I am very grateful to our own staff who have worked hard on this study, especially Alan Passingham and Matthew Purves. I am also extremely grateful to the members of our expert advisory panel, whose helpful advice contributed a great deal to the project. The panel has included, at various points: Professor Robert Coe, Dr Melanie Ehren, Lesley Duff, Dr Iftikhar Hussain, Danielle Mason, Stefano Pozzi, Rebecca Allen, Sam

Freedman and Jonathan Simons. We are very much looking forward to continuing to work with these and others as we develop this work in the future.