‘Do two inspectors inspecting the same school make consistent decisions?’

A study of the reliability of Ofsted’s new short inspections

Published: March 2017

Reference no: 170004

Contents

Executive summary

1. In January 2015, Ofsted announced that it would carry out a study into the reliability of short inspections when it introduced them in September 2015, evaluating them from the outset.[1] This was in response to criticism that the inspectorate had not done enough in the past to reassure the sector that school inspection judgements were consistent and reliable. This study was designed to be the first step towards collecting a body of evidence on the reliability of inspection practice. It aimed to evaluate how frequently two inspectors independently conducting a short inspection of the same school on the same day agreed whether the school remained good or whether they needed further evidence to reach a secure decision. It therefore tested reliability, not validity. The study also looked at the factors that drive reliability in short inspections.

2. Short inspections were introduced as a more proportionate approach to reduce the burden of inspection on good schools. A short inspection begins with the assumption that a school is still good. The purpose of it is to determine whether the school continues to provide a good standard of education and whether safeguarding is effective. Good schools have a short inspection, conducted by one or two inspectors and lasting one day, approximately every three years. This approach, which is more timely than the five-year cycle, is intended to identify schools that may be declining. Equally, schools that have improved can also be recognised and acknowledged earlier. Unlike a full section 5 inspection,[2] under short inspections inspectors do not make a full set of judgements under the common inspection framework or change the overall effectiveness judgement of the school. At the end of the short inspection, either the school will remain good or, if the inspector believes that more evidence needs to be gathered, the inspection will be 'converted' into a full section 5 inspection within 48 hours.

3. Following pilot inspections to test the design of the reliability methodology in the summer term 2015, the study was carried out during live short inspections in the academic year 2015/16. In total, we carried out 26 inspections of primary schools of above average size from across Ofsted's regions using the reliability methodology. We analysed the results from 24 completed inspections. All of these inspections were carried out by Her Majesty's Inspectors (HMI), with one identified as the 'lead inspector' and the other as the 'methodology inspector' per school. Along with the independent decision reached by each inspector, evidence was also collected from reflective evidence forms completed by inspectors during the inspection and post-inspection interviews with participants. Independent observers monitored whether there was any

[1] http://schoolsweek.co.uk/coming-soon-ofsted-double-inspections.
[2] Section 5 of the Education and Inspections Act 2006; http://www.legislation.gov.uk/ukpga/2006/40/contents

interaction between the inspectors that could invalidate the test at four of these inspections.

4. The results of the study indicate that inter-observer agreement between the lead and methodology inspector was strong. In 22 of the 24 short inspections, the inspectors agreed on their independent decision about whether the school remained good or the inspection should convert to gather more evidence. In two inspections, the inspectors arrived at different conclusions about whether the school remained good or needed to convert to a full inspection to collect further evidence.

5. The variation in one of the schools where inspectors disagreed overall came from the inspectors' subjective interpretation of the same evidence. While subjectivity was also observed in other inspections in the sample, the differences in these instances were not significant enough to affect the common view reached by both inspectors at the end of the inspection. This could suggest that Ofsted's protocols and inspection guidance for short inspections help to increase reliability by minimising the influence of subjectivity in the inspection process.

6. In the other school where inspectors disagreed on the decision about whether to convert the inspection, this was linked to issues with the study design rather than differing interpretations of the evidence. Some of the inspectors spoken to about the process said that, while the aim was to replicate the inspection carried out by the lead inspector, this was not always possible. For a few, the methodology inspectors' role was necessarily an artificial experience of a true inspection, which led to some unexpected variation.

7. In general, it was easier for inspectors to establish similar inspection practices in schools that were judged to have remained good by the end of the short inspection. This was due to the inspection being less of a burden for these schools. Additional barriers, particularly the time available for both inspectors to carry out similar activities with the same individuals, were more constrained in those inspections which converted. Critically, however, variation in inspection approaches did not commonly lead to disagreement at the end of the inspection. Methodology inspectors tended to secure a sufficient quantity and quality of evidence that led them independently to reach the same decision as the lead inspector, no matter how different the inspection pathways they followed were (sometimes substantially so) from those of the lead inspector.

8. The validity of the study depended largely on how successfully the inspections were conducted without one inspector unintentionally influencing the views of the other. In one case, it was clear that the lead inspector had influenced the approach of the methodology inspector. This inspection was therefore removed from the sample results. However, there is reasonable security that the 24 completed inspections were carried out independently. The two inspections where inspectors disagreed on their final decision are indicative of independence, as were the findings from three of the inspections where an

independent observer was involved. The fourth independent observer did identify minor infringements by inspectors that could impact on their colleague's independence. Minor infringements were also self-reported by inspectors from other inspections in the sample. However, rather than looking to intentionally bias decisions, these incidents were often about minimising the burden of the study on the school while maintaining the integrity of the live inspection. As such, we are reasonably confident that these inspections were conducted with minimal overlap between inspectors.

Factors associated with reliability

9. The findings from the methodology inspections lead us to hypothesise that there are four important factors associated with reliability. The first two relate to evidence gathered in this study:

   - **Triangulation** of the headteacher or senior leadership team's views from the initial leadership meeting against other evidence collected from the inspection is an influential driver of reliability. Agreement on judgements appears to be the result of aggregation of multiple pieces of different evidence supporting the perception that focused lines of enquiry and the collection of different types of evidence leads to greater consistency.
   - The **inspection framework** and the detailed grade descriptors the inspection handbook provides to support inspection judgements are important components in reducing subjectivity. The short inspection framework provides a fail-safe mechanism at the end of the inspection as it allows inspectors to convert to a full inspection if they need more evidence. This adds an additional layer of security that the final judgement given is reliable.

10. The next two hypotheses were not tested by the methodology study and are therefore offered more tentatively:

   - We hypothesise that Ofsted's **quality assurance** procedures provide further assurance that accurate judgements are made by inspectors, although this has proved impossible to test under this methodology and requires further study.
   - Similarly, we hypothesise that **inspector training** has a considerable effect on the consistency of inspector practice and judgements, leading to greater adherence to the inspection framework and therefore greater reliability. This factor was similarly not tested in the current methodology inspections and requires further study.

11. Overall, the evidence provides moderate security that the outcome agreed between the inspectors involved in these inspections were reliable and

consistent.[3] However, some caution is required in interpreting these results. The small number of inspections and the specific context of the sample mean the results of the study should not be generalised more broadly, particularly to reflect the reliability of all short inspections conducted by Ofsted. Further research with a larger sample and the involvement of more independent observers in these inspections (for greater assurance of independence) would likely strengthen the current findings and provide the means of testing the hypotheses set out in (1) to (4) above.

12. A logical next step from this study would be to ask which components of the short inspection methodology are most effective in driving consistent, accurate inspection decisions. While this study can suggest conclusions about reliability, it is less able to provide evidence on the validity of inspection. This is an area that Ofsted is interested in pursuing further and we will continue to engage with those in the sector to help shape future approaches to our evaluation work.

Key findings

- The inter-observer agreement between the independent inspectors was relatively strong. In 22 of the 24 completed inspections, inspectors agreed on their final decisions at the end of the short inspection. A full table of outcomes can be found on pages 24 to 26.
- In these 22 inspections, differences in the evidence collected between inspectors or in their interpretation of the same evidence were not sufficient to influence the overall inspection outcome. Inspectors reached the same view of the school overall regardless of whether they had conducted different inspection activities, or indeed the same activities but in a different order.
- The pre-inspection lines of enquiry that were formed independently by both inspectors tended to be similar.
- In one of the two inspections where inspectors disagreed on the final outcome, this was in part due to the inspectors interpreting the school's self-evaluation document and the initial discussion with the senior leadership team in different ways. This led to each inspector independently forming different perceptions of the capacity of the senior staff.
- In the other inspection where there was disagreement, inspector views were more clearly associated with inspectors undertaking different inspection activities with different people. The artificial nature of the inspection pathway that the methodology inspector followed was linked to these differences. In this instance,

[3] This assessment is arrived at on the basis of the qualitative evidence and the robustness of the method. The method has weaknesses and limitations, but it has not significantly affected inspectors' ability to arrive at similar decisions. The qualitative evidence provides assurance that no deliberate discussion between inspectors took place and that any unintentional bias introduced had limited impact on the findings.

the methodology inspector was unable to speak to some individuals they would normally interview as lead inspector, such as the chair of governors.

- The fact that inspectors disagreed on the outcomes of two inspections show that these inspectors kept to the conditions of the study and arrived at their decisions independently. This, alongside the evidence collected from three of the independent observers and the spontaneous style of completed reflective evidence forms, suggests independence was generally maintained across the sample.
- Some small infractions were identified where inspectors either spoke to each other or participated in activities together beyond the agreed method design. This was often due to the school not being prepared sufficiently for the methodology test or was agreed on by the inspectors to reduce the burden of inspection on school leaders or to avoid rehearsal bias. As such, unintentional influence by inspectors cannot be completely ruled out.
- Despite attempts to make the study design as equitable as possible for the methodology inspector as for the lead inspector, some participating inspectors suggested that the process was still too artificial. Whereas the lead inspector had priority throughout the inspection in how she/he established and followed inspection trails, sometimes it was difficult for methodology inspectors to decide on and carry out inspection activities as they would on a routine inspection. This rarely affected the overall decision reached though. Methodology inspectors were generally seeing enough quality evidence that led them to arrive at similar conclusions as the lead inspectors.
- The inspectors interviewed indicated that the methodology approach tended to work best and more equitably in the schools that were judged to have remained good. It was less of a burden to apply the methodology in these schools. More practical issues with implementing the methodology design were found in schools where the short inspection converted to a full inspection, particularly around having enough time for both inspectors to carry out similar activities.
- Inspector views on the methodology test varied. Some saw the process as good professional development and an opportunity to reflect on their own practice. A few inspectors mentioned that they found it reassuring to reach the same decision as a colleague independently on the same inspection. They felt that it validated their own inspection practice. Conversely, some methodology inspectors found not having ownership of the inspection frustrating. In these cases, the methodology approach was something that detracted from how they would normally conduct an inspection.
- Overall, 11 of the inspections converted to a full section 5 inspection. This includes the two inspections where inspectors disagreed: in both, the full inspection found these schools remained good. Two other schools with weaknesses identified on the short inspection were also found to be good after converting to a full inspection. A further three schools declined to requires improvement and another three were judged inadequate for overall effectiveness. One school with considerable strengths that converted was subsequently judged outstanding.

- The outcome of these 11 conversions suggests the short inspection methodology acts as a fail-safe mechanism that ensures accurate judgements are routinely provided. Rather than making a final decision based on incomplete evidence, the additional time given by the conversion process to acquire more relevant evidence at a full inspection adds an additional layer of security that the final judgement given is reliable.
- Agreement was generally reached in the reflective evidence forms completed by the inspectors, although there was greater variation in the forms at the first reflection point. The variation was often due to different interpretations of the evidence presented at the initial meeting with school leaders. By the end of the short inspection, initial differences between inspectors tended to converge as wider first-hand inspection evidence was gathered. This suggests that the leadership meeting alone is not sufficient for inspectors to consistently agree and that it is the triangulation of different sorts of inspection activity across the day that secures the level of reliability observed in this study.
- Along with the short inspection framework and the triangulation of evidence, it is our hypothesis that Ofsted's quality assurance procedures and inspector training make up four factors that appear to be associated with attaining greater reliability on short inspections.

Introduction

13. There is a perception among some stakeholders that assessments made by inspectors of school quality are too often unreliable or at least that there is no measure of inspectors' reliability in coming to their judgements.[4] That is, if different inspectors had inspected the school, how likely is it that they would have arrived at the same overall conclusion about the schools' effectiveness? This concern carries additional weight considering the uses to which inspection outcomes are put by those accountable for the quality of education in England.

14. Questions about the reliability and validity of school inspection in England are not new. Concerns following the formation of Ofsted and the prevailing untested methodology of classroom observation were at the time highly contested.[5] To date, there remains little empirical evidence about the validity of inspection judgements.[6] Subsequent research looking at inter-rater reliability between inspectors has, however, found that inspectors' findings are reliable in that two inspectors independently observing the same lesson will generally come to similar outcomes about the quality of the lesson.[7] The recent Measures of Effective Teaching (MET) project in the US has also indicated that teacher observation becomes more reliable when more than one observer watches the class.[8] Substantial training in observation was provided for this study, however, and some commentators have suggested that training in observation carried out by Ofsted inspectors or professional colleagues is generally not of the quality and scale used in the MET study.[9] Other recent research has posed that observations lasting 20 minutes may be sufficient time for raters to assess lesson quality reliably and evidence from the health sector has also suggested that groups of inspectors produce more reliable assessments than individual inspectors alone.[10,11]

15. The introduction of each new inspection framework in England has been met with limited research, whether by Ofsted or by external parties, into either

---

[4] H Waldegrave and J Simons, 'Watching the watchmen: the future of school inspections in England', London, Policy Exchange, 2014.
[5] CT Fitz-Gibbon and NJ Stephenson, 'Inspecting Her Majesty's Inspectors: should social science and social policy adhere?', The European Conference on Educational Research, 1996; www.leeds.ac.uk/educol/documents/000000048.htm.
[6] G Holger and HA Pant, 'How valid are school inspections? Problems and strategies for validating processes and results', Studies in Educational Evaluation, 37:2/3, 2011.
[7] P Matthews, JR Holmes, P Vickers and B Corporaal, 'Aspects of the reliability and validity of school inspection judgements of teaching quality', Educational Research and Evaluation, 4:2, 1998.
[8] MET project: 'Gathering feedback for teaching: combining high quality observations with student surveys and achievement gains', Bill & Melinda Gates Foundation, 2012.
[9] R Coe, 'Classroom observation: it's harder than you think', CEMblog, 2014; http://cem.org/blog/414.
[10] AJ Mashburn, JP Meyer, JP Allen and RC Pianta, 'The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality', Educational and Psychological Measurement, 74:3, 2013.
[11] Boyd, A., Addicott, R., Robertson, R., Ross, S., & Walshe, K. (2016). Are inspectors' assessments reliable? Ratings of NHS acute hospital trust services in England, Journal of Health Services Research & Policy, Early access

reliability or validity. Since the formation of Ofsted, the approach to inspection has continued to evolve. As such, some of the existing literature has less relevance in the current context. The short inspection methodology introduced in September 2015, for instance, has a greater focus on the impact of leadership on overall school effectiveness than previous frameworks, yet there have been few studies of approaches to evidence gathering outside of those relating to lesson rating, which is no longer part of the school inspection methodology. Some international research has looked at whole-school inspection processes, albeit across a very small number of schools, to isolate the legitimising constructs of inspector judgement, but no evaluation of reliability has been conducted on a real-time inspection process in the English context.[12]

16. In response to these criticisms, Sean Harford HMI, Ofsted's National Director of Education, announced in January 2015 that Ofsted would undertake a study into the reliability of inspection. This would focus on the reliability of inspectors' decisions in the new short inspection framework. The following report describes how this study was designed and implemented and sets out the results and main findings. It also tries to identify what the next steps are in the process of developing a measure for the reliability of school inspections.

Purpose and design of short inspections

17. In September 2015, Ofsted implemented short inspections for maintained schools and academies (and further education and skills providers) that were judged good at their previous inspection. Short inspections were introduced to be a more proportionate approach to reduce the burden of inspection on good schools, as the considerable majority of these schools tend to remain good at their next inspection.[13] This type of inspection was designed to encourage greater, constructive, honest and professional dialogue between inspectors and school leaders. A feature of short inspections is that the lead inspector gives ongoing feedback to school leaders throughout the day.

18. Short inspections begin with the assumption that a school is still good and that the purpose of the inspection is to determine whether the school continues to provide a good standard of education and whether safeguarding is effective. Instead of a good school receiving a section 5 inspection with a full inspection team that make a full range of judgements up to every five years, good schools now have a short inspection approximately every three years conducted by one or two inspectors (depending on the phase and size of the school) that lasts one day. This more timely approach is intended to identify schools that may be in decline earlier, before failure sets in; schools that have improved can also be

[12] J Lindgren, 'The front and back stages of Swedish school inspection: opening the black box of judgment', Scandinavian Journal of Educational Research, 59:1, 2015.
[13] In 2015/16, 73% of primary and secondary schools that received a short inspection remained good. Eighty-three per cent remained good or improved to outstanding.

recognised and acknowledged sooner. The short inspections are also intended to make the process less high-stakes for schools.

19. A short inspection does not change the overall effectiveness judgement of the school and inspectors do not make a full set of judgements under the common inspection framework. However, if inspectors are unsure of whether a school remains good, they will convert the inspection into a full (section 5) inspection. A short inspection, therefore, has three possible outcomes:

    - outcome 1 – the school continues to be a good school
    - outcome 2 – the school appears to be at least good and there is sufficient evidence of improved performance to suggest that the school may be judged outstanding, leading to the short inspection being converted to a section 5 inspection
    - outcome 3 – inspectors have insufficient evidence to satisfy themselves that the school remains good; this leads to the short inspection being converted to a section 5 inspection; if safeguarding is not effective, the inspection will always be converted to a section 5 inspection.

Reliability and inter-rater agreement

20. As already mentioned, reliability research on school inspection is an under-developed area of investigation. This study will be one of the first published that focuses on reliability across a full inspection process. The lack of previous research, however, makes it difficult to predict the degree of reliability that is acceptable within an inspection context. Owing to the consequences of a failed inspection, the imperative is for Ofsted to ensure that decisions by inspectors are as consistent as possible. Indeed, researchers generally agree that the greater the consequences resulting from the evaluation, the greater the need for high inter-rater agreement. Yet this view is tempered by the knowledge gained from current literature on assessment that removing all unreliability caused in marking examination papers is probably impossible.[14] As such, the aim must be to get as close as possible to perfect reliability given the constraints.[15]

21. Reliability between assessors judging the same item has been studied for many decades, across a range of disciplines. This has led to a variety of different methods being developed for analysing reliability.[16] A simple measure of reliability can be determined by the percentage of absolute agreement. This is a straightforward calculation of the number of times raters agree on a rating,

[14] J Tisi, G Whitehouse, S Maughan and N Burdett, 'A review of literature on marking reliability', National Foundation for Educational Research (NFER), 2013.
[15] J Wilmut, R Wood and R Murphy, 'A review of research into the reliability of examinations', University of Nottingham, School of Education, 1996.
[16] T Bramley, 'Quantifying marker agreement: terminology, statistics and issues', Research Matters - Cambridge Assessment, Issue 4, 2007.

divided by the total number of ratings. This measure can vary between 0 and 100%. Alternatively, a more stringent measure of reliability can be applied. Cohen's kappa coefficient is a statistic which measures inter-rater agreement for categorical items. It is generally thought to be a more robust measure than a simple percentage agreement calculation, as it takes into account whether the agreement reached has occurred by chance. The values of the coefficient can range from 1, where there is exact agreement, to 0 where there is no agreement. Early research in the use of kappa coefficients to measure inter-rater agreement described the relative strength of results as ranging from 'fair' (a coefficient of less than 0.21) to 'almost perfect' (a coefficient between 0.81 and 1).[17]

22. Many researchers have recognised that reliability should be considered within the context of the field of study. The lack of reliability research in the inspection space means we have to look further afield for comparable measures. Assessment practice provides a comprehensive body of literature on reliability and is similar to inspection in that both operate in high-stakes educational environments. This evidence is useful for providing a baseline with which to determine an acceptable level of reliability of inspector agreement, although some caution is required in seeing this as a plausible comparator. Even within educational assessment, two different approaches to assessing the same subject can be very different in terms of reliability.

23. The following studies estimate kappa coefficients for the inter-rater agreement of long or complex essay questions where secondary assessors have been blinded to the outcome of the original markers. Correlation coefficients include 0.72 and 0.73 for questions in English and business examinations respectively;[18] an average correlation of 0.87 in A-level sociology and 0.75 in A-level economics;[19] and correlations between 0.89 and 0.97 for components of assessments with relatively complex mark schemes.[20] The existing assessment literature suggests that an acceptable level of reliability is reached where coefficients are found to be above a minimum threshold of 0.7.[21] Research on the reliability of teacher observation in the United States also indicates benchmarks of high agreement of around 0.75 and 0.80.[22]

24. Where percentage agreement is concerned, studies have found that there is between 41.7% and 67.1% agreement within the grade bandwidth for long questions[23] and 78.9% and 97.4% agreement within a range of 10% of the

[17] J Landis and G Koch, 'The measurement of observer agreement for categorical data', Biometrics, 33:1, 1977.
[18] A Fearnley, 'An investigation of targeted double marking for GCSE and GCE', AQA Online, 2005.
[19] AJ Massey and N Raikes, 'Item-level examiner agreement', Cambridge Assessments, 2006.
[20] V Dhawan and T Bramley, 'Estimation of inter-rater reliability', Cambridge Assessment, Ofqual, 2013.
[21] V Brooks, 'Double marking revisited', British Journal of Educational Studies, 52:1, 2004.
[22] M Graham, A Milanowski, and J Miller, 'Measuring and promoting inter-rater agreement of teacher and principal performance ratings', Center for Educator Compensation Reform, 2012.
[23] V Dhawan and T Bramley, 'Estimation of inter-rater reliability', Cambridge Assessment, Ofqual, 2013.

total mark for components of English GCSE higher tier exams.[24] Indeed, in the United States, examiners for the National Assessment of Educational Progress must pass a test before they are allowed to mark; the examiner must secure a level of agreement above 70% of the pre-assigned scores.[25] From the teacher observation literature, various experts have suggested that percentage of absolute agreement values above 75% demonstrate an acceptable level of agreement.[26] With these results in mind and juxtaposed against the context of inspection reliability an estimate of a Cohen's kappa coefficient of 0.7 or percentage agreement of 80% will likely suggest a high level of agreement between inspectors.[27]

25. While more agreement is obviously always better than less agreement, perfect agreement is neither possible without limitations nor cost-effective to achieve. A number of steps to move rater agreement closer to perfect agreement are possible, but this may lead to oversimplifying the performance measures to a point where the approach to inspection is mechanistic and would undermine its validity. The assessment literature provides a good example here. Multiple-choice questions or assessments in mathematics where there is an unambiguous correct answer have often been shown to have perfect inter-rater agreement. However, longer essay questions are more likely to have lower reliability.[28]

26. Thus, some degree of professional judgement is necessary if decisions are expected to reflect different levels of complex behaviour. This will likely lead to experts disagreeing at times, although this does not mean one or the other is wrong in terms of outcomes or that mistakes have been made. For instance, research on vocational assessment has suggested the reproducibility of decision pathways is more relevant than reliability[29] and that teacher assessment should be characterised not by reliability but by whether the amount of information used in forming a decision was adequate.[30] In any complex evaluation, there are likely to be areas where decisions on which aspects to investigate can lead to a legitimately different view on the overall result.

27. Additionally, other sources of unreliability may also be present in any systematic error that leads to less reliable decisions. For instance, unreliable assessors

[24] Diana Fowles, 'How reliable is marking in GCSE English?', English in Education, 43:1, 2009.
[25] AlphaPlus Consultancy Ltd, 'Standardisation methods, mark schemes, and their impact on marking reliability', Ofqual, 2014.
[26] M Graham, A Milanowski, and J Miller, 'Measuring and promoting inter-rater agreement of teacher and principal performance ratings', Center for Educator Compensation Reform, 2012.
[27] Since the value of the kappa coefficient depends in part on how ratings are distributed across levels, high values should not be expected if most of the ratings are at one level. As there are only three categories of rating available for this study, there is a possibility that most outcomes will be at one level (the school remains good).
[28] V Dhawan and T Bramley, 'Estimation of inter-rater reliability', Cambridge Assessment, Ofqual, 2013.
[29] LKJ Baartman, TH Bastiaens, PA Kirschner and CPM van der Vleuten, 'The wheel of competency assessment: presenting quality criteria for competency assessment programs', Studies in educational evaluation, 32:2, 2006.
[30] KJ Smith, 'Reconsidering reliability in classroom assessment and grading', Educational measurement: Issues and practice, 22:4, 2003.

appear to account for less error than other factors, such as test-related variability or the varying performance of the candidates.[31]

28. The research evidence available suggests a number of mechanisms that could be used to avoid punishing schools for unreliable inspector judgements, such as additional observations or opportunities to submit further evidence post-evaluation, high-quality training and agreement meetings during the inspection. The conversion process of the short inspection methodology investigated in this study is potentially one such mechanism that helps to increase the reliability of school inspection in a cost-effective way.

Purpose of this study

29. The purpose of this study was two-fold. First, the short inspections introduced in September 2015 were a new type of inspection that required inspectors to take a different approach to inspection compared with previous frameworks.[32] While pilot inspections had been carried out in the lead-up to the new framework, these focused on the implementation of the short inspection process. This study, therefore, was designed to help identify which elements of short inspections most drive consistency and reliability. Second, with regards to the challenges raised by commentators in the previous section, this study also focused on testing the reliability of inspection outcomes. Combining these purposes together provided two testable key research questions to pursue:

- Are inspection outcomes reliable for short inspections?
- Is the current reliability testing method an effective approach for establishing reliability in short inspections?

30. Secondary research questions that we wanted to answer included the following:

- How frequently do inspectors agree on their overall outcomes?
- Where inspectors arrive at different decisions, what are the reasons for this?
- Are differences in the evidence base due to variation in inspection practices or the subjective interpretation of evidence?
- Do variations in subjectivity or process matter? Do they actually lead to inspectors making different decisions?

[31] J Baird, M Hayes, R Johnson, S Johnson and I Lamprianou, 'Marker effects and examination reliability: a comparative exploration from the perspectives of generalizability theory Rasch modelling and multilevel modelling', University of Oxford, Ofqual, 2013.
[32] The common inspection framework: education, skills and early years, Ofsted, 2015; www.gov.uk/government/publications/common-inspection-framework-education-skills-and-early-years-from-september-2015.

Research design

31. This study was designed to evaluate how frequently two inspectors conducting an independent short inspection of the same school on the same day agree on whether the school remains good or whether the inspection should convert.[33] Some consideration was given to how independence between the two inspectors could be assured. For example, discussion was had about whether inspectors could inspect the same school on different days, but issues with the second inspector's knowledge of the first inspection (and possibly its outcomes) and rehearsal bias suggested that independence would likely be compromised.[34]

32. Additionally, discussions also covered whether the study should be conducted on a live inspection or under conditions that replicate a live scenario. There were some concerns from the sector that this would lead to an unnecessary burden on schools, as seen through the lens of a 'double inspection', if the methodology approach was conducted during live inspections. Owing to the artificiality of a non-live environment, particularly when the pressure associated with reaching a decision is removed, it was felt this would not provide an accurate evaluation of inter-observer agreement. It was therefore decided that the study would test inspector consistency under as realistic conditions as possible on the same day, that is, during live short inspections that would result in a published inspection report. To counter the issue of additional burden, schools selected for methodology testing could opt out of participating in the tests.

33. Participating HMI were provided with additional guidance on how to carry out these methodology test inspections. The key measurement used to identify differences in inter-observer agreement replicated the three outcomes inspectors make on a routine short inspection:

    - convert to a section 5 inspection as there is potential for an outstanding overall effectiveness judgement
    - the school remains good
    - convert to a section 5 inspection as insufficient evidence has so far been gathered to confirm that the school remains good.

34. Evidence forms were also completed by each of the participating inspectors to inform the secondary research questions on inspector subjectivity and consistency.

---

[33] The data table shows that this school did not convert until after the third reflection point. This was due to the mis-application of the conversion process for the methodological inspections by the region. Both inspectors agreed they would have converted after the discussion with the headteacher, although both were advised to continue collecting evidence by regional colleagues.

[34] This means that as the school has had the chance to run through the inspection with the first inspector they would likely be better prepared for the inspection of the second inspector, leading to biased observations.

Pilot inspections

35. Initially, the feasibility of the study was trialled in seven pilot inspections during the 2015 summer term. This proved valuable in determining what worked in the design and where changes were needed.

36. In particular, the pilots revealed that the HMI who participated were commonly failing to understand the purpose of the study and how to conduct the inspections independently. A few declared in post-inspection discussions that they had spoken frequently with the other inspector throughout the process. One inspector also identified that bumping into the other independent inspector was a regular occurrence in the very small primary school they had visited. On this basis, there appeared to be little security that the outcomes reached on the pilot inspections had been arrived at by each inspector independently of the other.

37. A further issue was identified by leaders of the schools participating in the pilot around the additional burden this type of inspection had on the school, particularly where the school was very small. Inspectors would likely observe the same teachers separately, which could be a burden for particular individuals or groups within the school. Additionally, the inspectors talking to the same leader about the same issue could give rise to rehearsal bias. Headteachers were also concerned about the clarity of the inspection process. It was often difficult for them to know which inspector was the lead inspector, and having the same conversation twice was seen as unhelpful. As such, the approach appeared to be compromising the effectiveness of the live inspection.

38. Finally, there were also issues with the decision to convert a short inspection to a full inspection. This could happen at any point during the day, not just at the end of the inspection. In two of the pilots, this had led to just a single shared outcome being recorded by the inspectors (as opposed to separate decisions). In this scenario, one inspector had identified the need to convert, had consulted with their colleague on the evidence they had collected and subsequently reached agreement to convert. This obviously had an impact on whether the second inspector had reached this decision to convert independently.

Developing the approach

39. The findings from the pilot inspections led to a number of changes being implemented to the study design. First, the selection criteria for the sample was amended to exclude small primary schools. School leaders were also given the opportunity to opt out of the study if their school was selected. Importantly, more detailed guidance and focused training was provided for participating inspectors to ensure that they fully understood the purpose of the study and the practical requirements for these inspections. This included specific protocols that:

- clarified the roles of the two inspectors
- minimised the burden on schools
- ensured that the inspection could be converted without compromising the independent decision of either inspector.

40. The key feature added to the design was the designation of a 'lead' and 'methodology' inspector on each inspection. These roles would differ slightly to encourage further independence in the approach of each inspector. The lead inspector would essentially conduct a routine inspection of the school. Likewise, the methodology inspector was expected to carry out their inspection in parallel, but with the exception that they would concede authority to the lead inspector's inspection trails where any overlap in processes occurred. In practice, this would mean that if both inspectors needed to observe the same lesson or speak to the same member of staff, the lead inspector would have priority. The methodology inspector in these instances would be expected to follow other inspection trails they had established in order to avoid either undermining or influencing the integrity of the live inspection. However, as the discussion with the headteacher at the start of the inspection is considered an important factor for developing lines of enquiry, it was important that both inspectors had this opportunity to meet with the headteacher. To avoid additional burden for the school and rehearsal bias, it was decided that both inspectors would jointly attend this meeting, although dialogue would be with the lead inspector only. The purpose of this meeting would ensure that:

    - the logistics of the inspection were agreed, including the timing of meetings and final feedback
    - the headteacher and senior leadership team had an opportunity to explain their self-evaluation to both inspectors
    - key lines of enquiry were agreed (although the methodology inspector would generate their own separate lines of enquiry).

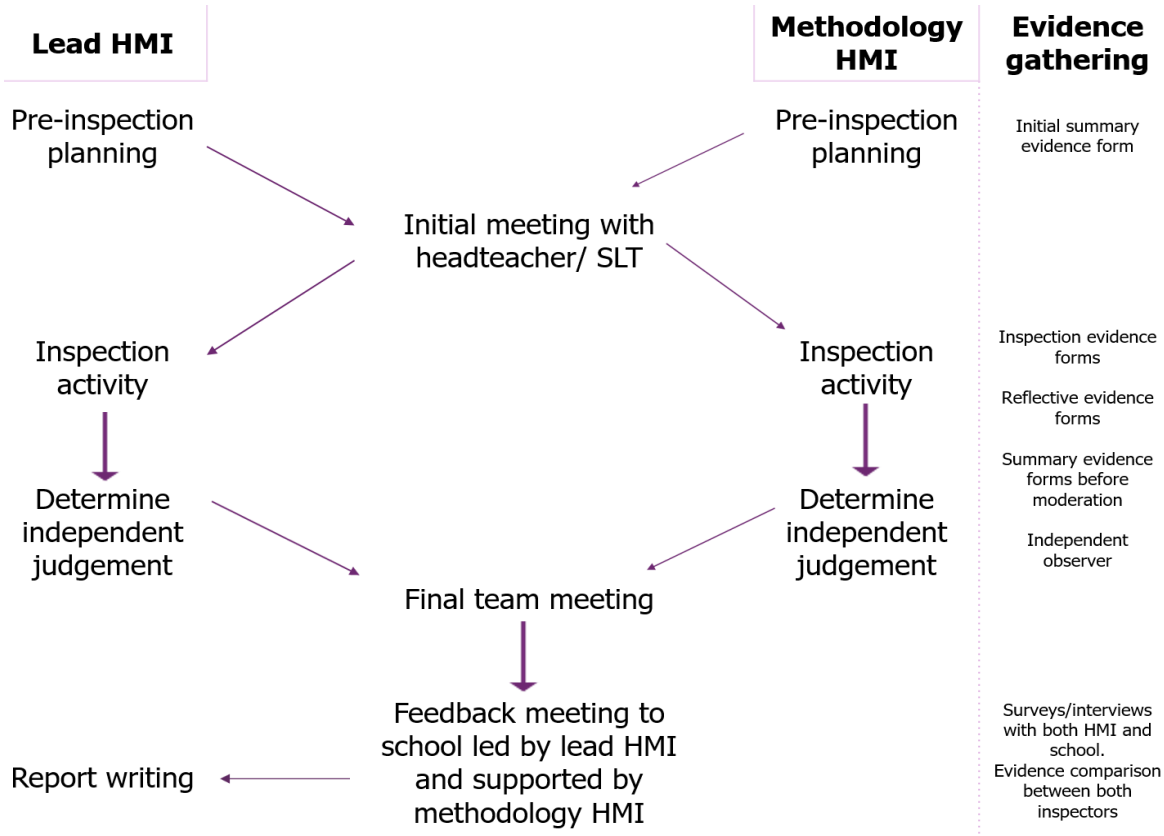
41. It is worth noting that this part of the process was still likely to introduce an element of confirmation bias. These discussions were based on the lines of enquiry developed by the lead inspector from the pre-inspection evidence available and not necessarily the methodology inspector's line of questioning at this point, which may have differed. This could, therefore, have influenced the methodology inspector's thinking and how they tackled the remainder of the inspection. On balance, it was considered that this limitation was an accepted element of the study design as the integrity of the live inspection was of greater concern.

42. With this in mind, at all times other than the initial meeting with the headteacher, the lead and methodology inspectors were instructed not to inspect together. They were expected to work entirely separately. This included how they accessed and interpreted pre-inspection information. It was also deemed important that the inspectors did not conduct interviews or

observations together. Although this may lead to some members of staff or pupils being interviewed or observed more than once, it was felt that too much prescription on the process would compromise each inspector's ability to investigate their individual lines of enquiry fully and successfully (which may also have an impact on the reliability of their decisions). It was accepted that this may lead to additional burden on the school, but not to the same extent as the initial leadership meeting. For this reason, selected schools were permitted to opt out of participating if they felt they were unable to accommodate the methodology design.

43. As the evidence collected by both inspectors was to be used for live inspection purposes as well as for the methodology tests, it was also important to identify where the reliability testing ends and the process of determining the official corporate outcome begins. It is worth reiterating that, as the inspections in the study were also 'live', the headteacher would need to receive clear, unequivocal feedback on the effectiveness of the school at the end of the process. Receiving feedback from two potentially different lines of enquiry would, therefore, be unhelpful, particularly if the inspectors independently disagreed on the inspection outcomes. A protocol was therefore established to determine the point at which the inspectors could meet to discuss their decisions and feedback for the school. The planned process involved each inspector sending their summary evaluation form (including strengths, weaknesses, areas of improvement and, most importantly, their final outcome) by email to the relevant regional management team overseeing the inspection. Once both summary documents had been received, the regional teams would contact the inspectors to confirm receipt. At this point, the reliability test component of the inspection would end and the inspectors were free to discuss their evidence to agree a final outcome.

Figure 1: Reliability inspection approach



44. As with all short inspections, conversion to a full inspection was one of three possible outcomes. In these methodology tests, either inspector may have had good reason for concluding that conversion was necessary. The decision to convert would, in most cases, be an agreed corporate outcome made by the inspectors in the team meeting following the submission of summary evaluation forms. There were, however, two possible exceptions:

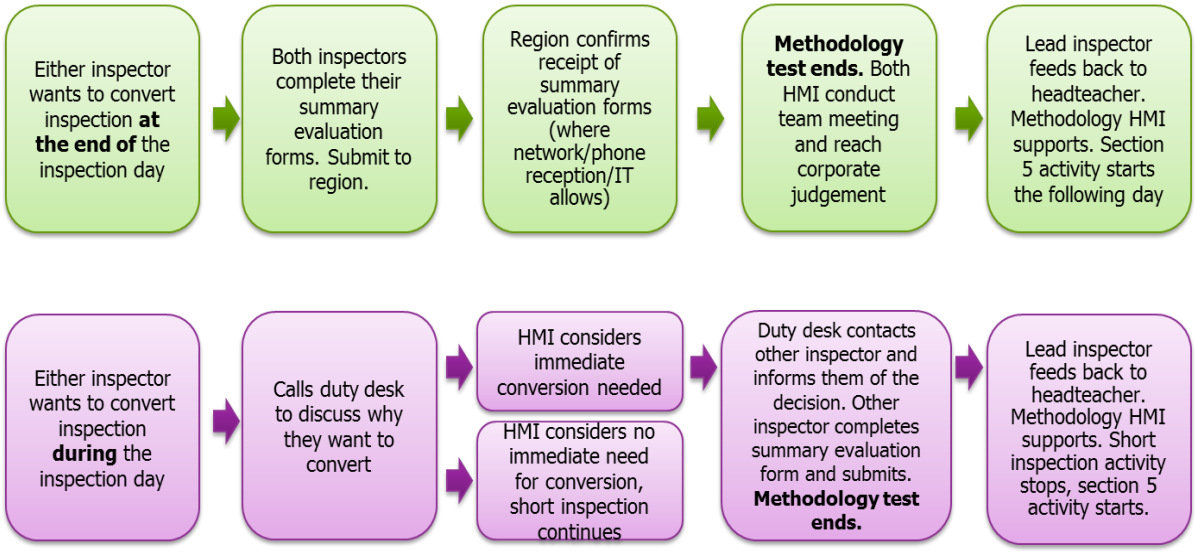
- where one of the inspectors considered it necessary to convert the short inspection before the end of the day
- where one inspector identified a potential safeguarding issue.

45. In either of these situations, the process for the methodology element of the inspection would involve the inspector contacting regional teams to discuss the possibility of conversion. If the inspector considered that an immediate conversion was needed (and conversion could not wait until the end of the day), the regional team would contact the other inspector and ask them to complete and submit their summary evaluation form. Once the summary evaluation forms had been submitted by both inspectors, they were free to meet. Following this, the reliability test would be deemed over and the inspection would continue as a converted, full inspection.[35] This process was designed to ensure that any differences between the inspectors' views would

[35] If the decision was made to convert for a safeguarding reason, the reliability test would stop as discussed, but the inspection would not be used as part of this studies evidence base.

have at least been captured. However, if the conversation with the regional team helped the inspector to conclude that immediate conversion was not justified at this point, the short inspection would continue under the arrangements of the study design.

Figure 2: Examples of the conversion process to ensure independent outcomes can be captured



Ensuring independence

46. We established a small panel of external experts to provide us with advice and guidance on this aspect of the research design. There were two specific ways in which the study attempted to provide evidence that inspector decisions were being made independently. As part of the process, both inspectors were asked to complete a 'reflective evidence form' at three specific points during the inspection (typically mid-morning, midday and mid-afternoon). This was a short note based on five minutes' reflection of the evidence the inspector had gathered on the school so far and utilised the same three-point scale to capture their internal 'barometer' of how the inspection was progressing. It is important to note that these three data-points in the day did not result in specific judgements. Instead, they were a mechanism through which to understand the differing ways inspectors formed their decisions throughout the day. The expectation was that these forms would help to show greater variation than the final outcomes each inspector reached at the end of the inspection and, therefore, the differing ways inspectors tackled the lines of enquiry to reach their conclusions.

47. While the reflective evidence forms could also help to identify where independent decisions had not been made, there remained a concern that we would only be able to infer that independence and separation were maintained in the inspections conducted. Essentially, there was no way to prove that the two inspectors had acted completely independently to arrive at their decision to convert or not. For instance, even insubstantial contact during the inspection, such as a hand gesture across a corridor, could lead to influencing the views of the other inspector. As such it was agreed that the study design would include provision for an independent observer on the inspection.[36]

48. Independent observers were therefore planned to be involved in some of the inspections carried out near the end of the autumn term 2015 and during the summer term 2016. A protocol was developed to guide the observer through the planned process and an evidence form designed to collect relevant information from the inspection. In general, no training was required, as the expectation of the independent observer was simply to record whether the two participating inspectors had acted independently throughout the inspection. As the study was about reliability rather than the validity of these outcomes, the observers were not asked to pass comment on the school, the decisions made or the conduct of the inspection. This would require further training in the inspection process for observers to comment effectively. Invitation to the role of an independent observer was offered to the small number of academics from the expert panel.

[36] This was added to the design after most of the autumn term inspections for the study had been completed.

Sample

49. Ordinarily, when investigating inter-observer agreement, the inclusion of a large number of events (data-points) is preferable. This is so that any subsequent analysis can account for factors that could coincide with and affect an inspection. Testing under live inspection conditions, however, limits the available sample size. The initial intention was to carry out the two inspector approach in approximately 80 schools. This sample size was calculated using an estimate of inter-rater agreement between inspectors of 80%, which was informed, by the assessment and teacher rating literature available, as to what might be expected as a high degree of inter-rater reliability for inspection.[37] In order to minimise undue influence on inspectors from other factors, the sample was designed to include schools that were similar to each other in certain ways. Short inspections already offered some commonality. The schools in the sample were, therefore, selected based on the following criteria:

    - school must have been good at its previous full inspection (fixed by the use of short inspections)
    - primary school only
    - between 250 and 500 pupils on roll.

50. The second and third criteria here helped to reduce the burden on the school because they were the planning conditions that meant ordinarily just one inspector would visit the school to carry out a short inspection. A large primary school or any size secondary school would warrant two inspectors for a standard short inspection, which would then have required a further two 'methodology' inspectors on site. Additionally, at this stage, only HMI participated in the study. This was because, at the time of the study, Ofsted inspectors had not been trained to carry out short inspections. It is also worth noting that all of the schools included in the sampling pool were scheduled for a routine inspection in the same way as any other previously good school.

51. The inspections and inspectors selected for the study were scheduled by the eight Ofsted regions using existing scheduling protocols. In total, 54 inspections were arranged for the study, that is, the date of the inspection had been established as had the two independent inspectors allocated for each to carry it out. In eight of the inspections, however, the headteacher opted out of participating in the study. In these instances, the short inspection continued in the usual way, without the methodology inspector. In a further 17 schools, the methodology component of the inspection was cancelled by the Ofsted region concerned. This was often due to having insufficient resource to supply a methodology inspector and, in a few instances, inspector illness. Finally, a

[37] The suggested sample size of 80 was a compromise aimed at achieving a margin of error for our reliability estimate with a confidence interval of plus/minus 10% or better.

further three test inspections did not go ahead as part of the study because the schools were subsequently found to have more than 500 pupils on roll.

52. This left 26 schools in the sample that regions could progress to the fieldwork stage.[38] Eighteen of these inspections were carried out in the autumn term 2015. The other eight were conducted in the summer term 2016. The smaller number of inspections carried out in the summer term was due to more headteachers opting out of the study (seven in total) in that term. These headteachers often indicated that the time of year was the main barrier preventing them from participating, often citing the clash with end of key stage national testing as their reason.

53. In total, four of these inspections included independent observers to check that the inspectors involved were collecting evidence and forming conclusions on the school's quality independently of each other. The two independent observers involved in the autumn term 2015 inspections were not fully independent, as they were carried out by former HMI. They were involved at this stage to further test the design methodology and develop the protocols of the role for subsequent use in the summer term 2016 inspections. The invitation to the expert panel to participate in the study as an independent observer was accepted by three of the members. One of these observers was scheduled to participate in a summer term reliability inspection on three occasions, but each time the methodology component of the inspection was cancelled. This meant that only two of the three volunteers from the panel were able to participate in the study.

Post-inspection interviews

54. Following the completion of these inspections, participating inspectors were interviewed about the process. Their views, along with the inspection evidence, were collected to provide further knowledge on how well independence was established on these inspections and to identify any additional barriers with the method that may undermine the findings. Inspectors were selected for discussion due to their availability, although attempts were made to speak to a balance of lead and methodology inspectors. In total, 23 of the participating inspectors, who were involved in 19 of the test inspections, were interviewed about their experience by telephone. This consisted of 12 lead inspectors and 11 methodology inspectors. Due to problems with the way some evidence was recorded, most of the post inspection interviews were conducted several months after the inspection took place. In addition, six headteachers from the autumn term 2015 inspections were interviewed by telephone.

[38] The evidence bases for 24 of these 26 are reviewed as part of this study. See 'Results', below, for an explanation.

Limitations

55. It is worth reiterating that the context in which the study was conducted was highly specific – only good primary schools of a certain size involving HMI carrying out the inspection were considered. Alongside the small number of inspections that were eventually carried out, this means that the results of the study should not be generalised more broadly, particularly to reflect the reliability of all types of Ofsted school inspection. While the study provides some insight on how likely it is that an inspection can be considered reliable within this specific group of schools, further research is needed to establish the effect that Ofsted inspectors,[39] different types of provision or different school contexts will have on reliability. The analysis that follows should be considered in the light of these limitations.

Results

56. Figures 3 and 4 provide details of the findings from 24 of the inspections carried out under the independent inspector methodology. Two inspections from the sample are not included in these findings. One of these inspections was removed from the results as it was identified that inspectors did not carry out the inspection independently of each other.[40] The other inspection converted early on the short inspection day owing to the identification of a potential safeguarding issue, which led to the methodology element of the inspection being discontinued. The tables show each inspector's views from the reflective evidence forms (where these were completed), their final independent decision at the end of the inspection, whether or not the inspection converted and the agreed corporate judgement following the completed short or full inspections.

[39] We directly contract Ofsted inspectors to carry out our inspections.
[40] The lead inspector carried out a pre-inspection data meeting with the methodology inspector which would likely have biased the decisions of both.

Figure 3: Outcomes from the methodology inspections carried out during the autumn term 2015

	Region	Inspector type	Reflection 1	Reflection 2	Reflection 3	Independent judgement	Converted	Final judgement
School A	South West	Lead inspector	Convert - more evidence	Still Good	Still Good	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School B	East Midlands	Lead inspector	Still Good	Still Good	None provided	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School C	South East	Lead inspector	Convert - more evidence	Convert - more evidence	Convert - more evidence	Convert - more evidence	Yes	Requires improvement
		Methodology inspector	Convert - more evidence	Convert - more evidence	Convert - more evidence	Convert - more evidence		
School D	North East	Lead inspector	Still Good	Still Good	Convert - more evidence	Convert - more evidence	Yes	Good
		Methodology inspector	Convert - more evidence	Convert - more evidence	Convert - more evidence	Convert - more evidence		
School E	East Midlands	Lead inspector	Still Good	Still Good	Still Good	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School F	South East	Lead inspector	Convert - outstanding	Still Good	Still Good	Still Good	No	Good
		Methodology inspector	Convert - outstanding	Still Good	Still Good	Still Good		
School G	North East	Lead inspector	Still Good	Still Good	Still Good	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School H	London	Lead inspector	None provided	None provided	N/A	Convert - more evidence	Yes	Inadequate
		Methodology inspector	Convert - more evidence	Convert - more evidence	N/A	Convert - more evidence		
School I*	East	Lead inspector	Convert - more evidence	Convert - more evidence	Convert - more evidence	Convert - more evidence	Yes	Inadequate
		Methodology inspector	Convert - more evidence	Convert - more evidence	Convert - more evidence	Convert - more evidence		
School J	East	Lead inspector	Convert - Outstanding	Convert - Outstanding	N/A	Convert - outstanding	Yes	Good
		Methodology inspector	Convert - more evidence	Convert - more evidence	N/A	Convert - more evidence		
School K	North West	Lead inspector	Still Good	Still Good	Still Good	Still Good	No	Good
		Methodology inspector	Convert - Outstanding	Still Good	Still Good	Still Good		
School L*	West Midlands	Lead inspector	None provided	None provided	None provided	Still Good	No	Good
		Methodology inspector	None provided	None provided	None provided	Still Good		

School M	North East	Lead inspector	None provided	None provided	None provided	Convert - more evidence	Yes	Requires improvement
		Methodology inspector	Convert - more evidence	Convert - more evidence	N/A	Convert - more evidence		
School N	London	Lead inspector	Convert - more evidence	Convert - more evidence	N/A	Convert - more evidence	Yes	Requires improvement
		Methodology inspector	Convert - more evidence	Convert - more evidence	N/A	Convert - more evidence		
School O	West Midlands	Lead inspector	Still Good	Still Good	Still Good	Still Good	Yes	Good
		Methodology inspector	Convert - outstanding	Convert - Outstanding	Convert - outstanding	Convert - outstanding		
School P	London	Lead inspector	Convert - outstanding	Convert - Outstanding	N/A	Convert - outstanding	Yes	Outstanding
		Methodology inspector	Convert - outstanding	Convert - Outstanding	N/A	Convert - outstanding		

Figure 4: Outcomes from the methodology inspections carried out during the summer term 2016

	Region	Inspector type	Reflection 1	Reflection 2	Reflection 3	Independent judgement	Converted	Final judgement
School Q	South East	Lead inspector	Convert - more evidence	Convert - more evidence	Convert - more evidence	Convert - more evidence	Yes	Inadequate
		Methodology inspector	Convert - more evidence	Convert - more evidence	Convert - more evidence	Convert - more evidence		
School R*	London	Lead inspector	Convert - more evidence	N/A	N/A	Convert - more evidence	Yes	Good
		Methodology inspector	Convert - more evidence	N/A	N/A	Convert - more evidence		
School S*	South West	Lead inspector	Still Good	Still Good	Still Good	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School T	East	Lead inspector	None provided	None provided	None provided	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School U	London	Lead inspector	Still Good	None provided	Still Good	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School V	South East	Lead inspector	Still Good	None provided	Still Good	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School W	East Midlands	Lead inspector	Still Good	Still Good	Still Good	Still Good	No	Good
		Methodology inspector	Still Good	Still Good	Still Good	Still Good		
School X	West Midlands	Lead inspector	Convert - more evidence	Still Good	Still Good	Still Good	No	Good
		Methodology inspector	Convert - more evidence	Still Good	Still Good	Still Good		

* = independent observer participated in inspection
N/A = the inspection had converted to a full inspection at this point, so no further reflective evidence forms were completed
Convert (outstanding) = convert on basis of potential outstanding overall effectiveness judgement
Convert (more evidence) = convert as there is insufficient evidence to confirm the school remains good
None provided = in some cases, reflective evidence forms were not completed by participating inspectors or only two reflection points were completed instead of three. In school T, the lead inspector was given the wrong advice from the regional team, who instructed the reflection forms only affected the methodology inspector and therefore did complete the documents.

Are inspection outcomes reliable for short inspections?

57. The data shows that in 22 of the 24 inspections the lead and methodology inspectors independently agreed on the outcome of the short inspection. Within the limitations of this small-scale, exploratory study, this suggests a high rate of inter-observer agreement between inspectors for these inspections. There was also common agreement across the reflective evidence forms completed by the inspectors, although more variation was observed at the first reflection point. By the end of the inspection, any initial differences tended to converge into a common view of the school that was arrived at separately by both inspectors.

58. For two of the schools, however, the inspectors involved did not reach agreement on the outcome of the short inspection. In school O, the lead inspector was confident that they had secured enough evidence to judge that the school remained good. The methodology inspector, however, required more evidence to determine if the school was potentially outstanding. There was a more pronounced difference between the lead and methodology inspector in school J though. Both inspectors determined that the inspection needed to convert, but for different reasons. The lead inspector perceived that more evidence was required to support a potential outstanding judgement, whereas the methodology inspector required more evidence to establish whether the school remained good. In both cases, the extra inspection activity allowed by converting to a full section 5 inspection led to a confirmation that the schools remained good.

59. It is important to note that, within the sample, some variation in inspectors' views would be expected. This is due to the relatively subjective nature of inspection. While Ofsted's framework and inspection guidance aim to provide consistency by ensuring that inspectors use the same assessment criteria and by suggesting types of evidence to consider, we cannot necessarily ensure that all inspectors responding to and interpreting evidence in the same way. Nor would we expect them to follow identical pathways through the inspection process. These expectations are consistent with the literature on inter-rater reliability discussed earlier; the key is for any potential inconsistency to be reduced as much as possible by the protocols available.

60. With inter-rater percentage agreement at 92% and a Cohen's kappa coefficient of 0.795, this suggests the precision of the 80% and 0.7 estimates for inspectors agreeing on their outcomes was reasonably good, but perhaps reflects a slightly pessimistic initial stance regarding inspector reliability.[41] However, while a relatively high level of inter-rater agreement has been

[41] This statistic is a weighted kappa calculation, where each result was coded for the lead inspector and methodology inspector as follows: convert (more evidence) = 0, still good = 1, convert (outstanding) = 2.

achieved, the small sample size means there is limited external validity to these findings. It is someway short of the calculated sample required for statistical validity. Therefore, we cannot be sure that a different sample of primary schools meeting the study criteria would produce similar results with regards to inspector agreement, so some caution with the interpretation of these findings is required.

61. In total, 11 of the inspections converted to a full inspection. This was generally to secure more evidence that the school remained good, as opposed to conversion because the school was potentially improving or in decline. For two inspections, the additional time was used to gather evidence to help establish whether the school had improved to outstanding. Overall, one school improved to outstanding after converting, four remained good, three declined to requires improvement and three were judged inadequate for overall effectiveness.

Was independence maintained between inspectors?

62. At face value, the outcomes from the sample suggest that the inter-observer agreement between two independent inspectors was relatively strong, but the validity of these results is dependent on the extent to which the inspections were carried out in the way intended, that is, without one inspector unintentionally influencing the view of the other and compromising their independent observations.

63. That two of the inspections provided outcomes where the lead and methodology inspectors disagreed provides some security that the study design was appropriately applied. This is also confirmed by the reports completed by the independent observers of schools I, L and S. In their view, the independent observers strongly agreed that the inspectors involved applied the methodological approach correctly throughout. While the observer of school S did record that the lead and methodology inspector had an initial discussion at the beginning of the day, this was only to agree the logistics of the inspection. It is clear from the observer's notes that the lines of enquiry each inspector intended to prioritise and pursue were not discussed at this point or at any other point during the inspection. In addition, five of the six headteachers spoken to indicated that they were confident that the inspectors did not confer during the inspection and confirmed that the inspectors conducted inspection activities separately, other than at the initial meeting, in line with the agreed intended test methodology.

64. The reflective evidence forms also provide some assurance that these documents were completed with minimal overlap between inspectors. A review of the handwritten 'snapshot' evidence forms shows variation in form and structure across the sample, not just between the two corresponding inspectors. A few were rather detailed and aligned to the lines of enquiry

established by individual inspectors. Most, however, featured a spontaneity that is likely associated with completing these documents quickly and without rehearsal. Importantly, providing the inspectors with a blank sheet of paper with which to form their reflective views means that these documents differed largely from the more formalised structure of the standardised evidence forms that were also completed during these inspections. The reflective evidence forms therefore help to illustrate the different ways inspectors approached this task of the study.

65. It is worth recognising, however, that the reflective evidence forms are clearly not as effective a mechanism for measuring inspector independence as the involvement of an independent observer in the process. Even where the writing of the reflection points was spontaneous, we cannot be totally sure whether the view given had been formed from an earlier incidence of unintentional contact or discussion between inspectors. Ideally, we would have preferred to have more independent observers involved in the study to account for every inspector interaction to rule this out, but geographical location, scheduling issues and observer availability meant this was difficult to arrange for this sample.

66. The findings of the independent observer from the inspection of school R are important here. Compared with the other independent observers who participated, the independent observer for this inspection noted a very different experience where independence was not as clear cut. In this case, the school was not sufficiently prepared for the methodology to be conducted as two distinct inspection processes. This led to the inspectors involved having no option but to carry out the same activities together. This clearly had implications for the study design. As the independent observer indicated, it became very clear how difficult it would be in practice for two inspectors in this situation not to talk to each other. The observer of school R also noted that, before deciding to convert the inspection at midday, both inspectors talked to each other about this decision, meaning the process established for this eventuality was not followed. In this case, it could be argued that the inspectors were reacting to circumstance and attempting to strike a balance between minimising the burden of the study while maintaining the integrity of the inspection, rather than looking to intentionally influence the other inspector. However, we cannot be sure the decision to convert was reached by the lead and methodology inspectors of school R independently of each other.

67. Indeed, discussions with the inspectors also reveal that mild infractions on the study design did occur for similar reasons in other inspections in the sample. For instance, in school F, the lead inspector indicated that both inspectors carried out the interview with governors, as logistically they felt it was not possible or helpful to conduct two conversations with this group of volunteers at separate points during the day. They were also concerned about the

potential for rehearsal bias with the second inspector to talk to the governors. This arrangement was agreed at the start of the inspection with the headteacher and conformed to the same protocol as the headteacher discussion. It was led by the lead inspector.

68. Inspectors in schools F and I also highlighted technical difficulties with the conversion process. In the former, issues with the telephone signal at the school made it difficult to discuss the decision to convert with their regional team, leading to the inspectors speaking about using email as an alternative to manage this part of the process. In the latter, inspectors highlighted that the conversion mechanism for the study was not well understood by the regional teams involved. This led to some discussion between the inspectors about converting. In each case, the inspectors spoken to were clear that at this point of the inspection they had already independently decided on conversion as a course of action. One of the headteachers interviewed also pointed out that the inspectors at their school had a brief conversation, during the morning, about a minor omission in the school's safeguarding single central record.

69. Overall, we are reasonably confident that these inspections were conducted with minimal overlap between inspectors. Where unintentional overlap is evident, we are also reasonably confident that this had limited effect on each inspector's final decision. The inclusion of more independent observers would likely have provided these findings with greater security.

Where differences were found, why was this?

70. In general, there were a large number of similarities identified in the evidence bases between inspectors. The lines of enquiry at the beginning of the inspection and the areas for improvement each inspector generated at the end generally matched between inspectors, although they were written differently (in order, reflecting the different inspection trails taken and in style). Similarities were also seen in the outcomes of the reflective evidence forms, particularly on the latter parts of the inspection where a general consensus emerged between the inspectors.

71. Inspectors noted that, depending on the circumstances of the inspection, it could also be relatively straightforward for them to reach common decisions independently. This was particularly the case in a few of the schools that converted. For instance, in school C, the headteacher had determined (through self-evaluation) that the school required improvement ahead of the initial senior leadership meeting with both inspectors. Each inspector found no immediate evidence to convince them otherwise and both had reached the decision early on that the school inspection should convert. The pre-inspection evidence of school N also led to early agreement. On this occasion, the annulment of the previous year's key stage data had led to concerns around the previous

leadership of the school and the validity of the school's historical performance in key stage tests. Both inspectors immediately picked up on this issue and independently determined early on in the short inspection that a full inspection would be required.

72. Differences that were found in the decisions between inspectors tended to occur for two reasons. Inspectors either subjectively interpreted the same evidence in different ways or alternative inspection pathways taken led them to differing conclusions. The former was more closely associated with the initial meeting with the headteacher where both inspectors received identical evidence at the same time. This did not necessarily lead to similar views being formed between the inspectors, as school A indicates:

> School A – The initial meeting was considered 'strange' by the lead inspector as it was mostly led by the deputy head, rather than the headteacher. This suggested some potential weaknesses to them. The inspector's view on leadership capacity changed after the learning walk with the headteacher though, as this indicated the headteacher really knew the school. Later discussions with the local authority identified that the headteacher was a very traditional headteacher, with teaching and learning as their speciality. The deputy head's role was more data and systems, hence why they led the opening exchanges of the inspection. The methodology inspector on the other hand did not identify any discrepancy with the deputy head leading the conversation at the initial meeting nor did they consider that the information the deputy head had provided suggested that leadership may be weak.

73. Subjectivity was indeed a main reason for why the inspectors of school J did not reach agreement on the final outcome of this short inspection. In this case, both inspectors interpreted the school's self-evaluation and the initial discussion with the senior leadership team in different ways. The lead inspector perceived the quality, rigour and robustness of the leadership to be particularly strong and advocated that the evidence suggested that the school may be outstanding. The methodology inspector was less convinced of this and was unable to determine whether the school remained good before converting to a full inspection. Both inspectors were in the same room and listened to the same evidence from the school's leaders and each formed different perceptions of the quality of school leadership.

74. In a few instances, the evidence forms indicated that the different inspection trails each inspector followed could also lead to subtle variations. For instance, the lead inspector conducted an in-depth scrutiny of pupils' work in mathematics books, whereas the methodology inspector did not, and the methodology inspector met with the early years foundation stage leader,

whereas the lead inspector did not. Schools D and F provide specific examples of this:

> School D – Owing to the differing inspection trails followed, the methodology inspector identified early on in their inspection an apparent issue with safeguarding at the school. This arose following a meeting with the chair of governors, who provided unconvincing responses about the school's safeguarding procedures. The inspector noted they would have converted the inspection at 10:15am on the basis of this evidence if this had been a routine inspection they were leading. The lead inspector did not pick up on the same concerns until later, as they were conducting a learning walk at the time the methodology inspector was interviewing the chair of governors. However, by the end of the learning walk, some concerns from pupils about bullying and the headteacher's views on the curriculum for preventing bullying suggested some concerns with personal development, behaviour and welfare. Further investigation before the end of the inspection led to the lead inspector making this concern a priority, which led to a decision to convert. Until the third and final reflection point, the lead inspector and methodology inspector differed in their view of the school.

> School F – While both inspectors agreed at the first reflection point that the school may be outstanding, they both had differing reasons for revising this initial view at the second reflection point. For the lead inspector, further scrutiny of the school's self-evaluation form and the discussion with the head of English confirmed that the school was not outstanding. Aspects of the learning walk that the methodology inspector subsequently conducted identified the school had some clear strengths but these were inconsistent across classes and pupil groups. Hence, the methodology inspector arrived at a similar conclusion to the lead inspector that the school remained good.

75. In general, however, even where inspectors followed different pathways through the inspection, they nearly always arrived at the same final outcome on the school's quality. This suggests that the pathways taken and the 'organic' nature of the inspection process, particularly in re-configuring evidence trails as new information comes to light, rarely prevented inspectors from arriving at similar decisions.

Were the inspections equitable for the methodology inspector?

76. The other inspection to show differences between inspectors' conclusions featured in school O. In this case, neither subjectivity nor the inspection approach was the main reason for the differing viewpoints formed, although

the former is certainly present. Instead, it appears that this may have come about due to complications in the design of the reliability methodology.

77. Discussions with the lead inspector of school O revealed that, in their view, the process for the methodology inspector was an artificial experience on this inspection. There was a slight lack of organisation on the school's part in preparing for the methodology inspection, which led to the inspectors engaging in slightly different processes. While both carried out the same activities, they were not conducted with the same people. The lead inspector, for instance, had their learning walk with the headteacher and spoke to the chair of governors. The methodology inspector, however, did not have the opportunity to carry out similar activities with these same individuals, although they did conduct a learning walk with another less senior member of staff and spoke to a different member of the governing board. As such, the quality of the evidence from these different individuals appears to have had an impact on the overall views of the inspectors. Whereas the lead inspector was confident that they had secured enough evidence to maintain that the school was still good, the methodology inspector had determined that they required more evidence to establish whether the school was outstanding.

78. As such, there are some implications from this scenario: chiefly, whether equity in the process was secured to ensure that both inspectors were equally well-informed about the school's strengths and weaknesses. The research design appears to suffer when attempting to minimise the burden of inspection, which in turn may affect the validity of the results received. In the case of school O, it seems the inspection was not truly a parallel exercise for the methodology inspector and, on this occasion, appears to account for the different conclusions reached. Indeed, the question to ask is this: if the methodology inspector had had the same level of access to the headteacher and the chair of governors, would they have arrived at a different outcomes to the one they did on the day?

79. Concerns around the equitability of the research design were also raised by other inspectors, commonly from those inspections where conversion was required. The burden of inspection was often referred to as a reason to explain why these inspections appeared less equitable. A few inspectors specifically referenced the governors' meeting as being particularly tricky to set up separately for each inspector. The integrity of the inspection was paramount in these inspections and this meant that the methodology inspector would tend to receive less priority than they might otherwise have liked. This also suggests that differences in inspection pathways between inspectors are linked to the implementation of the methodology design. Two of the methodology inspectors indicated that the process was frustrating for them, particularly around the ownership of the inspection. Indeed, they tended to feel they were unable to carry out the inspection in the way they normally would on a routine inspection.

80. Not all methodology inspectors suggested the process was artificial though. Where the school was clearly good from the outset, there also appeared to be much greater clarity of a shared process. In these situations, inspectors appeared better able to inspect separately and in line with the methodology process. A couple of inspectors also likened the process to receiving high-quality continuing professional development; the process was structured in such a way that it was much better than a shadowing experience, as the need to carry out tasks and complete evidence forms replicated the activities expected of a routine inspection.

81. This evidence shows the complexity in the design applied and the difficulties in securing a fully parallel approach between inspectors. It would seem clear that, for certain types of inspection, the approach is not always able to sustain a high degree of equity between inspectors in the evidence they are able to secure. This is constrained, as would be expected, by the need to secure balance between the integrity of the inspection and the methodology approach in a live environment – the integrity of the inspection holds greater weight. As such, this may affect the reliability of some decisions made by the methodology inspector that they may not otherwise make on a routine short inspection.

82. Despite these implications, the methodology inspectors were still able to secure final outcomes that agreed with their lead inspector counterparts in the vast majority of the schools inspected. This in turn suggests that the methodology is sufficiently robust to ensure that both inspectors have a similar enough experience of the school's quality to measure reliability.

Inspector views of the reliability study

83. Inspector views from the post-inspection interviews provide further hints that the design was implemented as intended. Some inspectors suggested that they enjoyed the opportunity to participate and that they found it a professionally rewarding process that two HMI could carry out inspections separately and arrive at the same conclusions. A couple of the inspectors likened the study to effective professional development that was reaffirming of their inspection practice, particularly in one instance where the inspector had been inspecting for many years.

84. Others indicated that they did not really enjoy the process. In some instances, this was due to the study design being difficult to manage; avoiding each other while trying to conduct a routine inspection under the parallel method was commonly identified as challenging. For others, the focus of the study being on inspector reliability made the process professionally uncomfortable. In these circumstances, rather than affirming, there was a sense of relief when both inspectors reached the same decisions at the conclusion of the inspection. For instance, one inspector indicated a feeling of nervousness when discussing with

their fellow inspector who was going to reveal their independent outcome first. It is unlikely similar views would be expressed if there had not been something important riding on the eventual independent outcomes of both inspectors.

85. As such, this evidence suggests that inspectors generally understood the rationale and purpose of the study and committed to completing it in a professional manner (all of the inspectors spoken to did tend to say as much). Again, no deliberate intention to undermine independence was identified, although we cannot be completely sure about unintentional contact between inspectors.

Next steps

86. While there is reasonable security that the study design is effective in measuring consistency between two inspectors and that the results suggest that, generally, inspectors reached the same conclusions, there remain two key limitations. First, the sample size is too small to provide the results with external validity. Second, the lack of sufficient involvement of independent observers across the sample means we cannot be completely sure that all of the inspectors in the test inspections arrived at their decisions independently. As such, continuation of the study across a larger sample of primary schools and with a greater extent of independent observers may be warranted to provide further assurance that short inspections are reliable.

87. Furthermore, to ensure that inspection is focused on the right mechanisms to maintain reliability, there is a need to establish which individual factors drive reliability. Based on the findings from the methodology inspections, some broad assumptions can be made about what some of the factors might be, although further testing is required to fully substantiate any association with reliability.

88. These main factors include the method of evidence retrieval during inspection (discussion with the senior leadership team and triangulation of evidence) and the underlying design of inspection (the inspection framework). The study provides some initial evidence in relation to these. It is also possible that two additional aspects of Ofsted practice may act to support reliability: quality assurance procedures and inspector training. This study, however, does not provide any evidence in relation to these factors and these would have to be subject to investigation.

Triangulation of evidence

89. The method of engaging directly with the headteacher and senior leadership team at the start of the short inspection stemmed from an assumption that assessing their views on the school's strength and weaknesses was both desirable and important, at least for collaboration purposes. These discussions

were particularly useful for the leaders of school C, who had already determined that the school required improvement, and for leaders at the other schools found to be inadequate at their subsequent full inspection after converting. The meeting gave these leaders the opportunity to own this decision alongside the inspectors and they often cited their own self-evaluation to conclude that the school suffered from particular weaknesses and would need to convert. In this way, collaboration at the starting point of the inspection could often lead to reliable inspection outcomes.

90. However, while the initial discussion with leaders appears to be an important forum for establishing robust inspection trails, the potential for some leaders to offer an unreliable narrative of their school's standards (as was the case with schools F, J, K and O) means that this mechanism alone does not lead to reliability. Instead, it is the interaction of corroborating these views against other evidence collected from the inspection that appears to be a more influential driver of reliability. The school inspection handbook and inspector training indicate the methods of evidence-gathering that inspectors are required to carry out.[42] These include:

- scrutiny of pupils' work
- talking to pupils about their work
- listening to pupils read
- discussions with school leaders, staff and governance
- discussions with parents
- scrutiny of documentary evidence about the quality of teaching
- scrutiny of the schools' other records and documentation
- short visits to lessons
- short observations of small group teaching
- extended visits to lessons, during which inspectors may observe activities, talk with pupils about their work and scrutinise pupils' work
- joining a class or specific group of pupils as they go from lesson to lesson, to assess their experience of a school day or part of a school day
- learning walks with senior staff
- joint observations of lessons carried out with the headteacher and/or senior staff.

[42] School inspection handbook, Ofsted, 2015; www.gov.uk/government/publications/school-inspection-handbook-from-september-2015.

91. Based on these methods, there are two main strands that can be derived from the triangulation process on the short inspections in the sample.

92. First, triangulation of further evidence did lead inspectors to amend their initial views of school performance following the initial leadership meeting. This is indicated by the reflective evidence forms completed where inspectors' views changed as leaders' interpretation of school performance were challenged and corroborated through wider, first-hand inspection evidence. On this basis, inspectors appear to have kept an open mind about school quality, rather than arriving at conclusions on the basis of data and other pre-inspection information alone. Neither did they frequently conclude that the initial discussion with the leadership team was sufficient enough to determine school quality.

93. Second, inspectors in the sample schools followed different inspection trails but tended to arrive at the same outcomes, suggesting that differences tended to be in areas where the weight of evidence was less influential.

94. The evidence collected suggests, therefore, that convergence is a consequence of aggregation, which helps to establish a possible link between triangulation and reliability. As long as inspectors looked at a broad range of evidence, focused on the targeted lines of enquiry developed in the initial leadership discussion and kept an open mind as to what this evidence might tell them about a school, there is the likelihood that a reliable outcome would be reached. Differences in inspection trails appear to have limited impact on reliability. Additionally, where subjectivity did lead to unreliability, other mechanisms of the short inspection, such as the fail-safe mechanism of the framework design, ensured that an accurate judgement could be eventually reached.

95. However, no work has been undertaken to identify which of these proposed methods of evidence collection is most (and least) closely linked to greater reliability of inspection outcomes. As such, further investigation is required to identify the activities that inspectors should focus their time on when triangulating evidence.

Framework, judgements and grade descriptors

96. From the perspective of the methodology inspection approach, there is some evidence to suggest that the short inspection framework does act as a safety net against inconsistency. In both schools J and O, the conversion to a full inspection to gather more evidence helped to ensure that a more accurate outcome was reached by the end of the process. Without the mechanism to convert, it is likely that two of the inspectors would have provided an unreliable judgement that these schools were outstanding, when the evidence from the other inspectors and consequently the full inspection suggested otherwise. This is an obvious improvement on the previous light-touch framework (reduced-

tariff inspections) that Ofsted employed between 2006 and 2009, which did not feature such a fail-safe mechanism.[43]

97. The security of the full inspection also ensured that secure, accurate judgements were made in nine other schools where both inspectors indicated they needed more time to determine whether the school was securely good or not. Of these, two remained good, six declined to either requires improvement or inadequate and for the other enough evidence was secured from the full inspection to show that the school had improved to outstanding. Again, if the conversion process did not feature as part of the short inspection, inspectors would more likely be making decisions on school quality using an incomplete evidence base, which would potentially decrease reliability. One of the strengths of the short inspection process, therefore, is the greater assurance it provides that the correct judgement can eventually be reached.

98. It is worth reiterating that the outcomes from the methodology inspections also indicate that a degree of subjectivity is removed from the short inspection process. Similar to full inspections, inspectors use the detailed set of grade descriptors set out for all judgements in the school inspection handbook, although these are only drawn from selectively for short inspections. Nevertheless, the training and experience all inspectors receive in using grade descriptors allows them to moderate their own judgements individually, suggesting that inspectors can, more generally, make regular consistent decisions when using a broadly objective yardstick to evaluate the quality of provision.

99. There are two assumptions here, however, that the methodology tests have not addressed. First is that the section 5 framework is itself reliable. This has not been subject to a methodology test in the manner of the work conducted in this study. Second is that a short inspection leads to outcomes that are similarly reliable to those produced by section 5 inspection. Again, this may benefit from further investigation. So, while it is initially apparent that the short inspection framework is particularly robust in helping inspectors to arrive at consistent outcomes, particularly where the design of the fail-safe mechanism is concerned, further work is required to see how reliable it is in comparison to other inspection types.

[43] Reduced tariff inspections (RTIs) were section 5 inspections that were allocated fewer inspector days for the size of the school than a standard section 5 inspection. RTIs normally consisted of one inspector in the school for one day. Schools were identified for RTI on the basis that data and other information available prior to the inspection indicated it was likely that the school would be judged good or outstanding for overall effectiveness. The full range of judgements were available on RTIs: schools could be judged outstanding, good, satisfactory or inadequate for overall effectiveness by the end of the inspection.

Quality assurance processes

100. This study focused on the activities undertaken by inspectors when they were on site at a school conducting a short inspection. No analysis was done of the impact of Ofsted's quality assurance work on the reliability of inspection outcomes. While it may be reasonable to infer that Ofsted's other procedures around inspection contribute to reliability, this study has not produced any evidence to substantiate this.

101. Ofsted's quality assurance and complaints procedures complement the short inspection framework design, providing further mechanisms to ensure that the final judgement a school receives is as secure and accurate as possible. In the event that a short inspection failed to provide a reliable outcome, the quality assurance procedures in place would ensure that the right judgement would be made about the school. In the first year of the short inspection methodology, a total of 74 inspections (including those that converted to a full inspection) were complained about by school leaders.[44] Of these, 26 were partially upheld, mainly on aspects of inspector conduct, administration or the management of the inspection. In only three cases was a concern about the judgement upheld. Two of these cases led to text changes in the inspection report but no change to the judgements awarded. This provides some assurance that Ofsted's complaints procedures and internal quality assurance mechanisms provide a further layer of security against inconsistency, but further study is required before any firm conclusions can be reached as part of this or another study.

Inspector training

102. This study did not assess or control for the training and experience of the individual inspectors who conducted the inspections. It may be reasonable to expect that inspectors' ability to inspect reliably by applying the framework, judgements and grade descriptors accurately and appropriately is determined in part by the quality of the training they receive (in addition to their experience and innate skills). Nevertheless, the design of this study means that it provided no evidence in relation to these factors.

103. Ofsted invests considerable effort and focus on the induction and subsequent professional development of each new HMI and Ofsted inspector. For example, induction takes place over the course of a term, with a mixture of face-to-face and online training and considerable opportunities to shadow inspections and learn from more experienced HMI. Ofsted's training packages and materials are built around helping inspectors to understand and apply the grade descriptors that make up the inspection judgements and the other aspects of the

[44] Figures given relate to the period from 1 September 2015 to 31 December 2016.

inspection handbook so that their judgements can be accurate. It may be, therefore, that further study of this area is important in drawing wider conclusions about which elements of Ofsted's short inspection methodology drive reliability.

Validity

104. The aspects above indicate further research is required in order to establish the reliability of inspection more broadly and suggests a few ways in which this could be achieved. However, investigating the validity of inspection in tandem is perhaps a more useful direction for future evaluation work than a narrow focus on reliability alone. For instance, the positive findings from this current study will be largely irrelevant if the components of current inspection processes are found to have little association in determining school quality. Furthermore, the absence of strong evidence in the research literature on the validity of inspection suggests this will remain a priority for the sector going forwards.

105. As we begin to determine Ofsted's evaluation priorities for the future, therefore, we will look to understand what activities could advance our understanding of the validity of different components of our practice. As part of that process, we will continue to engage with those in the sector, academics and other experts to help shape the approach we take.

The Office for Standards in Education, Children's Services and Skills (Ofsted) regulates and inspects to achieve excellence in the care of children and young people, and in education and skills for learners of all ages. It regulates and inspects childcare and children's social care, and inspects the Children and Family Court Advisory and Support Service (Cafcass), schools, colleges, initial teacher training, further education and skills, adult and community learning, and education and training in prisons and other secure establishments. It assesses council children's services, and inspects services for children looked after, safeguarding and child protection.

If you would like a copy of this document in a different format, such as large print or Braille, please telephone 0300 123 1231, or email enquiries@ofsted.gov.uk.

This publication is available at www.gov.uk/government/organisations/ofsted.

Interested in our work? You can subscribe to our monthly newsletter for more information and updates: http://eepurl.com/iTrDn.

Piccadilly Gate
Store Street
Manchester
M1 2WD

T: 0300 123 1231
Textphone: 0161 618 8524
E: enquiries@ofsted.gov.uk
W: www.gov.uk/ofsted

No. 170004