Introduction and background

Ofsted’s new Education Inspection Framework (EIF) has an increased focus on the quality of education in a broader sense. Ofsted is keen to ensure that the methods which its inspection methods are reliable. To this end, it is conducting a multi-phase research project of which these are the initial reports. The first report examines the results of a recent pilot of workbook scrutiny indicators. The second focuses on a lesson observation instrument for inspections which was piloted in the autumn term of 2018.

Key findings

Lesson observation: research design

- In November 2017, Ofsted held an international seminar on lesson observation; its purpose was to learn from experts around the world how to develop a model of lesson observation which would be aligned to the principles of the EIF. Ofsted also undertook a literature review looking at a range of current research on lesson observation.
- On the basis of the seminar and the literature review, an observation instrument was developed. It was made up of 18 indicators across 3 domains of interest, namely curriculum, teaching and behaviour. Each of the indicators was scored on a 5-point scale.
- In order to test reliability, 2 inspectors observed the same lesson at the same time. Inspectors also carried out multiple lesson observations across at least 2 subject areas. In total, inspectors visited 22 schools and 15 colleges, completing 346 paired observations across 74 departments.

The reliability of the observation model

- Ofsted examined whether inspectors had had difficulties assessing lessons against any of the indicators. Inspectors identified 2 indicators which were particularly difficult to score as they did not always feature in lessons: strategies to support reading (1c) and determining whether assessment provides relevant information about the current skills and knowledge of learners (1h). It was felt that these indicators are best measured in other ways and they were therefore removed from the lesson observation instrument.
- The pattern of scoring which emerged was similar to that produced by other lesson observation models. As in other models, observers typically scored behaviour indicators more favourably, as shown by the skew in the distribution towards a score of 4 and 5. The teaching and curriculum indicators, by comparison, feature a distribution centred around a score of 3. This shows, encouragingly, that observers were not using behaviour as a proxy for good teaching. It was not uncommon to view lessons in which pupils’ behaviour was exemplary, yet the quality of teaching observed was particularly poor.
- Data from the observations suggests that elements in the teaching criteria often influenced scoring on the curriculum criteria. Inspectors explained that they often found it difficult to distinguish between the teaching and curriculum indicators during observations, hinting that these are perhaps focused on relatively similar things and are more closely related than had originally been envisaged. Further work is required to develop how observers approach looking at the curriculum in lessons.

The validity of the observation model

- Under the EIF, lesson observation alone will not be the main factor of a judgement as it is for models which focus on teacher performance management. Therefore, failing to achieve perfect reliability in the model is not necessarily a problem.
- The level of agreement between observers was measured according to Cohen’s kappa. The values of the kappa coefficient range from 1 (exact agreement) to 0 where there is no agreement. Agreement of between 0 and 0.2 is ‘slight’; between 0.2 and 0.4 is ‘fair’; between 0.4 and 0.6 is ‘moderate’; between 0.6 and 0.8 is ‘substantial’ and between 0.8 and 1.0 is ‘almost perfect’.
- Observations from the primary schools generally attained a reasonable level of reliability (curriculum 0.64, teaching 0.63 and behaviour 0.6).
- The secondary school data shows that the overall behaviour score had substantial reliability (0.64). The curriculum and teaching statistics for the secondary school sample achieved a moderate level of reliability (0.58 and 0.57 respectively).
- Synthesis of lesson observations at the subject department level enhanced reliability across several indicators. Six indicators in the primary school sample and four from the secondary school sample attained a substantial level of reliability when observers combined observation evidence from across a subject department.
- The inter-rater reliability in schools is encouraging given that observers have reached this level of consistency without standardised training. Observers confirmed how useful it was to have a structure in place.
- Agreement in the college lessons observed was much weaker. The kappa statistics only attained a mild level of reliability or worse across all indicators; 0.3 for curriculum,
0.3 for teaching and 0.27 for behaviour. This would suggest that the more variant factors that exist between the school and further education and skills sectors reduce reliability and perhaps require a different approach for their specific contexts.

- It is encouraging that school HMI observing together generally reached a more substantial level of reliability than when a non-HMI observer was involved. This was the case across the overall domain scores and 9 of the indicators, whereas only moderate reliability was reached when a non-HMI was involved.
- Longer observations tended to lead to greater reliability. This was particularly the case for behaviour indicators. Observers commented that flexibility for them to determine the length of time to remain in lessons was a positive aspect of the model.
- Inspectors involved in the study stated that additional training and reducing the cognitive load of the model by cutting the number of indicators would probably result in a more consistent approach. This view is echoed by the research literature.

Workbook scrutiny: research design

- In order to develop the indicators for workbook scrutiny within the EIF, Ofsted consulted several inspectors and looked at the available research literature. Four indicators for workbook scrutiny were selected from the ‘implementation’ indicators. These were: building on previous learning; depth and breadth of coverage; pupils’ progress; and practice (i.e. opportunities to revisit and practice what they know). As with lesson visits, each indicator was scored on a 5-point scale.
- Ofsted obtained workbooks from three schools. Nine inspectors took turns scrutinising them without discussing their judgements during the exercise.
- Qualitative data from the exercise was obtained through open-ended questions in questionnaires and through a focus group interview. Data was assessed for reliability using Cohen’s kappa (see above).

Workbook scrutiny: qualitative feedback

- Feedback from inspectors was generally positive. They felt that the indicators gave them a good focus and prevented them from concentrating on irrelevant factors such as neatness or handwriting.
- All inspectors were confident about the bands they awarded when using the indicators for the subjects in their area of expertise. When scrutinising books for the subjects outside of their expertise, 6 out of 9 felt confident in the bands they awarded.
- The difficulties experienced by inspectors in applying the indicators may have been partly due to the lack of other evidence that they would usually gather as part of live inspection. One inspector explained that; ‘depth and breadth of coverage really also depends on what the school’s own curriculum is, e.g. in year 9 they may still be doing key stage 3 work’.
- Inspectors felt that there was not always a sufficiently clear distinction between bands. They found it particularly difficult to differentiate between bands 1 and 2 and between bands 4 and 5. They emphasised the need to make the language of certain descriptors more precise. For instance, they needed more precision on the meaning of quantifiers such as ‘some’ and ‘considerable’. Some inspectors suggested moving form s 5-point to a 3-point scale.

Work scrutiny: reliability

- Overall, the levels of agreement between inspectors were fair. The kappa score for building on previous learning was 0.47; for depth and breadth of coverage it was 0.49; for pupils’ progress it was 0.45; and for practice it was 0.38.
- Since inspectors had reported difficulties distinguishing between some bands (1 and 2, and 4 and 5), Ofsted tested whether merging the awarded bands of 1 and 2 into a single band, and 4 and 5 into another band would increase reliability. When this was done, the reliability stayed nearly the same for ‘Building on previous learning’ and ‘Pupils’ progress’. It increased for ‘Depth and breadth of coverage’, whereas it decreased for ‘Practice’.
- There were some significant differences between the primary and secondary phase. For ‘Building on previous learning’ the primary agreement score was 0.54, whereas in secondary it was just 0.22. For ‘Depth and breadth of coverage’ it was 0.42 for primary and 0.59 for secondary. For ‘Pupils’ progress’ it was 0.54 for primary and 0.32 for secondary. For ‘Practice’ it was 0.45 for primary and just 0.21 for secondary. There are a number of factors which may account for these differences such as: small sample size; the fact that fewer subjects were covered in the secondary sample; and lack of inspectors’ subject matter expertise.
- Following the workbook scrutiny exercise, one of the next steps is the revision of some descriptors, particularly those for ‘Practice’, so that it is easier to distinguish across bands more clearly. Eliminating the indicator ‘Pupils’ progress’ will also be considered, because inspectors found that it overlapped considerably with ‘Building on previous learning’.

Workbook scrutiny: conclusions

- Results suggest that using indicators and a rating scale requires a further trial. However, a clear focus and consistency supported by inspector training are important to maximise validity and reliability of work scrutiny. Using indicators and rubric provide a focus on scrutinising what matters,
- Focusing workbook scrutiny across a single subject/department/year group is helpful in securing greater validity and reliability.
- Carrying out workbook scrutiny without context is likely to limit validity. Conversations with subject leaders or teachers on the purpose of tasks in the worksbooks and how they contribute to learning progression can help provide that context.
- Including workbook scrutiny, alongside lesson observation and discussion with the subject lead and the teachers and pupils observed, provides greater confidence that the overall assessment of the subject area would be valid and reliable.
- It is important that workbook scrutiny may not be as relevant in all settings. It may not, for example, be possible to implement in special schools where pupils’ progress is captured in different ways. It may not be as relevant in Modern foreign languages, since a lot of classroom activity might be oral rather than written.
- Students in the further education and skills sector may not typically be required to bring in their work to classes (for example sixth form pupils), and the main written activity during lessons may be note-taking.