Workbook scrutiny

Ensuring validity and reliability in inspections

Her Majesty's Inspectors (HMI) can assess the quality of education by using workbook scrutiny indicators and they do so reliably. The report outlines the findings and the next phase of research.

Published: June 2019

Reference no: 190028

Contents

Introduction

The focus of inspection under the new framework has shifted to the quality of education more broadly. Some of the evidence inspectors will gather during an inspection under the new education inspection framework (EIF) will feed in to the overarching judgement on the quality of education. For example, conversations with leaders shed light on how curriculum is conceptualised, while lesson observations and workbook scrutiny provide a window into the quality of curriculum implementation.

In order to ensure a standard and consistent approach to inspections, we developed and piloted a number of indicators (or assessment criteria). These indicators unpack essential aspects of education in relation to curriculums, teaching and learning. We selected a few of those indicators and further tailored these to workbook scrutiny.

This report sets out a recent pilot of indicators and rating scales for workbook scrutiny. We needed to investigate their validity and fitness for purpose, so our first question was:

1. Does the piloted approach to workbook scrutiny allow meaningful assessment of the quality of education?

The study design also included an initial and small-scale exploration of reliability, so we also asked:

2. Can inspectors rate reliably using the piloted workbook scrutiny indicators?

The first research question was answered through the findings arising from questionnaire and focus group feedback of the participating HMI. The second one was answered through a statistical analysis of the level of agreement between HMI judgements.

Methodology

This was a mixed methods study, with the convergent parallel design.[1] We collected both quantitative and qualitative data to allow a more rounded validation of the piloted indicators and rating scales. This is the first phase of a multi-phase research project.

Participants

Nine HMI participated in the pilot study. Most of them (n=7) have substantial experience of two to three years or more in the role. Two HMI have one to two years of experience or less. Their subject expertise were in English, mathematics, science,

[1] JW Creswell and VL Plano Clark, 'Designing and conducting mixed methods research', SAGE Publications, 2011.

history and geography. They scrutinised workbooks within and outside of their areas of expertise (see Table 1).

Table 1: Areas of expertise

	Areas of expertise	**Areas outside of specialism**
HMI 1	English	Science
HMI 2	English	History/geography
HMI 3	Mathematics	English
HMI 4	Mathematics	English
HMI 5	Science	English
HMI 6	Science	Mathematics
HMI 7	History	French, science
HMI 8	French	History/geography
HMI 9	French	Mathematics

Materials

We obtained workbooks from primary and secondary schools to ensure that key stages 2 and 3 were represented (see Table 2). The subjects matched the participating HMIs' areas of expertise.

Table 2: The range of workbooks

Subject areas in workbooks	**Primary**			**Secondary**	
	Year 3	Year 4	Year 5	Year 8	Year 9
Mathematics	15	15	15	2	2
English	15	15	15	8	12
History and geography	29	15	15	14	
Science	15	15	15	10	15
French	15	15	15		

Workbooks in each subject were scrutinised by at least two HMI specialising in the subject. The exceptions with only one subject specialist were history workbooks and primary workbooks for science.

The same workbooks were also scrutinised by two or three non-specialist HMI. The exceptions were French workbooks and primary school English workbooks, which were examined by only one non-specialist HMI.

The study design of at least two HMI per workbook and subject allowed an initial examination of reliability.

Indicators

In order to develop the indicators for the EIF, we consulted several HMI and looked at the available research literature. We selected four indicators for workbook scrutiny from a wider range of the indicators designed for the whole inspection process (see Table 3). We drew the workbook scrutiny indicators from the 'implementation indicators' and tailored them further with the following in mind:

- the aspects of the quality of education described in the indicators should be observable in workbook scrutiny
- the indicators should cover different aspects of the quality of education, for example:
  - what is taught and learned (the breadth and depth of subject-matter content)
  - how subject matter is taught and learned (from the perspective of how learning is structured to allow for efficient and meaningful acquisition of new knowledge)
  - whether and how pupils consolidate knowledge so that it remains in their long-term memory.

Table 3: Book scrutiny indicators selected for the pilot

Building on previous learning	Depth and breadth of coverage	Pupils' progress	Practice
Pupils' knowledge is consistently, coherently and logically sequenced so that it can develop incrementally over time. There is a progression from the simpler and/or more concrete concepts to the more complex and/or abstract ones. Pupils' work shows that they have developed their knowledge and skills over time.	The content of the tasks and pupils' work show that pupils learn a suitably broad range of topics within a subject. Tasks also allow pupils to deepen their knowledge of the subject by requiring thought on their part, understanding of subject-specific concepts and making connections to prior knowledge.	Pupils make strong progress from their starting points. They acquire knowledge and understanding appropriate to their starting points.	Pupils are regularly given opportunities to revisit and practice what they know to deepen and solidify their understanding in a discipline. They can recall information effectively, which shows that learning is durable. Any misconceptions are addressed and there is evidence to show that pupils have overcome these in future work.

Each indicator has a five-point rating scale, ranging from 1 (minimum) to 5 (maximum). Each of the five bands in each indicator was accompanied by a descriptor – a text which describes the quality of education at a particular level.

Data collection process

We obtained workbooks from three schools. Nine HMI took turns scrutinising them without discussing their judgements during the exercise. This took place in one of our offices and in a single day.

Before starting workbook scrutiny, HMI were given time to familiarise themselves with the four indicators. They then applied the indicators to the workbooks, recording their judgements by year and key stage within the allocated subject areas and providing a rationale for their judgements. Following that, they completed a questionnaire about the indicators and the piloted workbook scrutiny process. Finally, they participated in a focus group interview.

This process is different from live inspection. In live inspection, workbook scrutiny is intended to complement conversations with leaders and pupils, as well as lesson observations. The aim in live inspection will be to establish whether the quality of pupils' workbooks matches leaders' curriculum intent of the curriculum. We could not achieve this in this pilot because the workbook scrutiny took place in isolation, due to practical constraints.

Data analysis

We collected both qualitative and quantitative data.

We obtained the qualitative data through open-ended questions in questionnaires and through a focus group interview with HMI. We then identified the main and recurrent themes.

We obtained some quantitative data through fixed-choice questions in the feedback questionnaire. Judgements awarded for each subject and year group also constitute quantitative data: they were marked on a 1- to 5-point scale.

In order to assess reliability, we used Cohen's kappa as the statistic to measure agreement between each two raters (HMI) who rated the same books using our indicators. The kappa coefficient is applicable for categorical or ordinal data. It is generally seen as a stronger measure than a simple percentage agreement calculation. This is because it takes into account whether the agreement reached has occurred by chance.

Research findings

Research question 1: Does the piloted approach to book scrutiny allow meaningful assessment of the quality of education?

The general finding derived from HMI feedback is that the piloted indicators are a step in the right direction. They helped HMI focus on the essential aspects of the quality of education, while minimising the effect of irrelevant factors such as

neatness or handwriting. The HMI all agreed that using the indicators 'allowed them to delve under the surface'. Some of the illustrative comments are provided below:

> 'It forced me to look at curriculum subjects in a new, deeper way. For example, I noticed in the history books I scrutinised that lower ability pupils focus more on literacy (reading comprehension), but not so much on grappling with the historical concepts or deepening history knowledge.'

> 'The indicators and descriptors eliminate questions about marking, handwriting, neatness, etc. They focus HMI more and can eliminate variation in what they focus on. This helps you think about what pupils are actually learning.'

The indicators require HMI to focus on knowledge sequencing as well as depth and breadth of content coverage (see Table 3). Therefore, we investigated how confident HMI were in their judgements and how easy they found it to use the indicators, both within and outside of their areas of specialism.

All HMI (9/9) were confident in 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, most HMI (6/9) felt confident in the bands they awarded. One HMI explained:

> 'to be fully confident out of your subject area, you need to have a secure understanding of the curriculum content in order to be able to judge progress etc.'.

Subject expertise did not affect the reported ease with which HMI were applying the rating scale. Using the indicators and descriptors, most HMI (6/9) found it easy to arrive at a judgement for the subject in their own area of expertise, while five out of nine reported the same when making judgements outside of their area of expertise.

It should be noted that using the indicators for workbook scrutiny was a novel experience for all participating HMI. Training and workbook exemplars should help increase inspectors' confidence in making judgements outside of their individual specialism, as well as the ease with which they can apply the indicators both within or outside subject specialism.

The difficulties that HMI experienced for this study 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. As one HMI explained:

> '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'.

Another HMI pointed out that:

> 'the exercise of work scrutiny needs to be complemented and triangulated with other evidence for the descriptors to have more validity'.

Differentiation across levels of the quality of education

Another factor that could have affected the ease with which some HMI applied indicators was the ability to distinguish between different bands.

HMI were asked whether they found it difficult to distinguish between different bands (1 to 5) that represent different levels of the quality of education. The main finding here is that there is not a sufficiently clear distinction between some bands. The bands that HMI found the most difficult to distinguish were the following:

- Bands 1 and 2 (6/9 HMI).
- Bands 4 and 5 (4/9 HMI).

HMI 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':

> 'Clarity of interpretation of language used such as *some*, *sufficient*, *considerable* – if this was being used there would need to be very clear definition of what some of this language means when applying it to judgements.'
>
> 'The use of terms like *adequate* need to be aligned between inspectors when talking about progress – as what one person considers adequate another may not. Might need some more "pulling out".'
>
> 'Establishing consistency in use of language and expectations – all inspectors need to be able to know what makes it *sufficient* or *adequate* for example. Important that there are benchmarks for all to be able to measure against and be accurate in doing so.'

Some asked for exemplification, 'particularly in terms of the tension between coverage and depth'.

HMI also asked for fewer bands because band 3 may 'end up as a dummy bit', or to otherwise increase differentiation between some bands.

The above suggests the following:

- The piloted five-point rating scale may benefit from shortening, combining bands 1 and 2, and 4 and 5, to form a three-point scale. We explore this further in the following section.
- Quantifiers would need to be exemplified to ensure that they are interpreted in a standard and consistent manner. This could be resolved through training and guidance materials with exemplars.

Research question 2: Can inspectors rate reliably using the piloted book scrutiny indicators?

The reliability of HMI judgements was investigated through Cohen's kappa co-efficient (see Methodology/Data analysis section above).

The values of the coefficient range from 1, where there is exact agreement, to 0 where there is no agreement (see Table 4). A negative kappa suggests that the inter-rater reliability is worse than it would have been had the ratings been produced by chance.

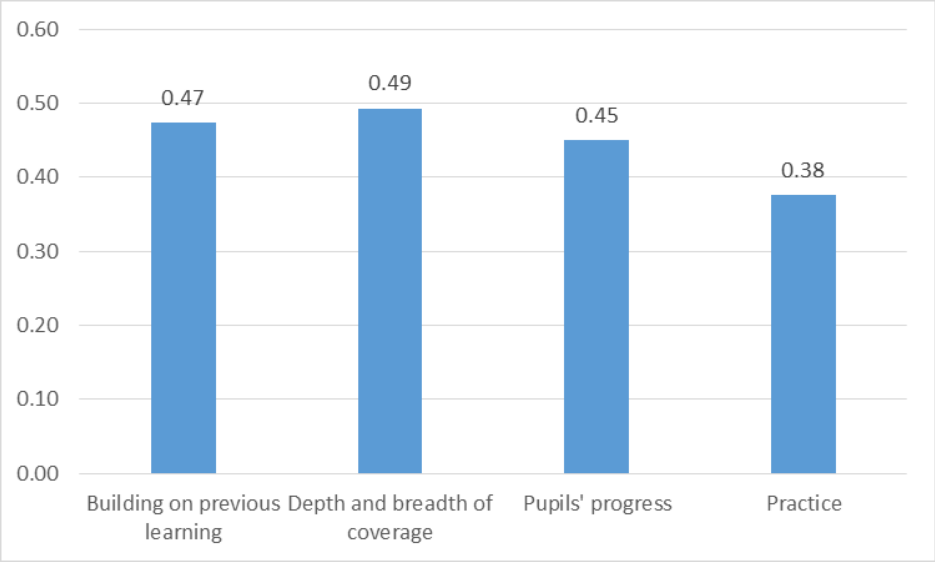
Table 4: Levels of agreement

Kappa statistic	Agreement
0<x≤0.2	Slight
0.2<x≤0.4	Fair
0.4<x≤0.6	Moderate
0.6<x≤0.8	Substantial
0.8<x≤1	Almost perfect

Reliability levels: overall

Figure 1 shows moderate levels of agreement between the marks awarded by nine HMI on three indicators. The agreement on the fourth indicator, 'Practice', is fair, but only marginally below the 0.41 cut-off for moderate agreement. This suggests that HMI rated reliably, using the workbook scrutiny indicators and rating scales.

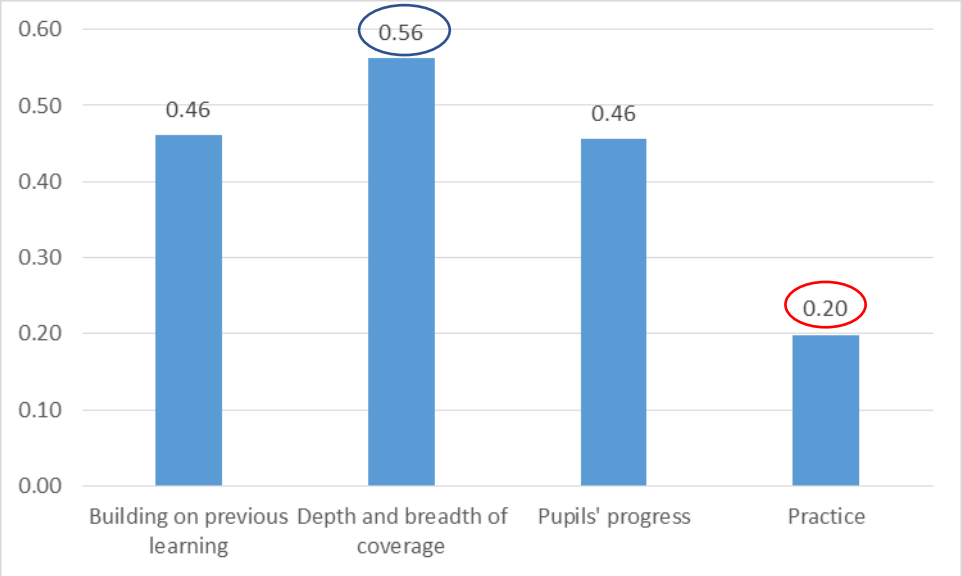
Figure 1: Inter-rater reliability by indicator (overall) – five-point rating scale



Given that HMI reported difficulties distinguishing between some bands (1 and 2, and 4 and 5), we tested whether merging the awarded bands of 1 and 2 into a single band, and 4 and 5 into another band would increase reliability. According to Figure 2, 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'.

Figure 2: Inter-rater reliability by indicator (overall) – three-point rating scale

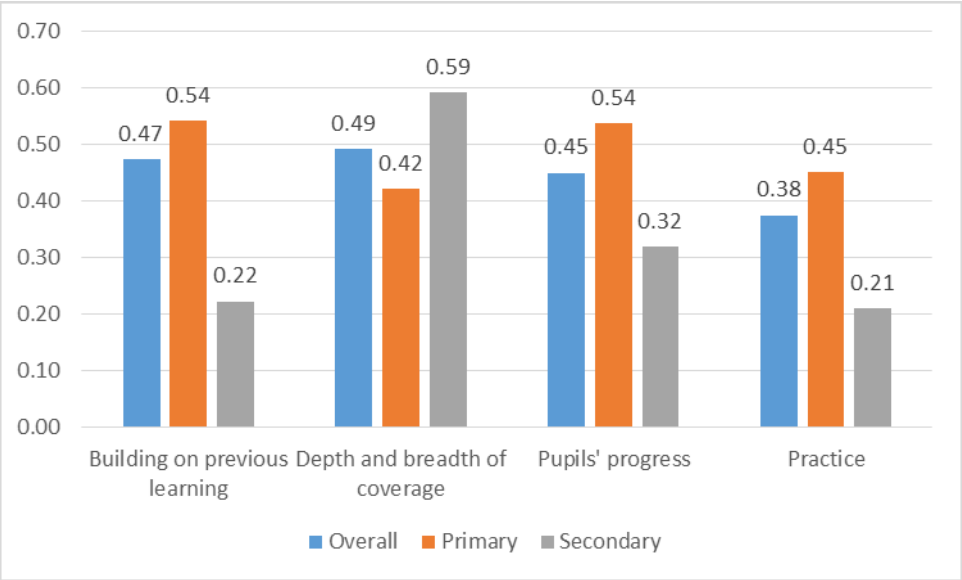


Reliability levels: education phase

Inter-rater reliability levels by education phase are displayed in Figure 3. The main finding is that inter-rater reliability is higher across all indicators at a primary school level:

- At a primary school level, reliability was moderate for three criteria ('Building on previous learning', 'Depth and breadth of coverage' and 'Pupils' progress') and fair but close to moderate for 'Practice'.
- At a secondary school level, reliability was moderate only for 'Depth and breadth of coverage', while being fair for the rest of the indicators.

Figure 3: Inter-rater reliability by education phase



It could be that subject-matter knowledge is more important in workbook scrutiny in secondary school inspections. Inter-rater agreement figures were based on the ratings of both subject specialists and non-specialists. The lack of subject matter expertise may mean that non-specialists could struggle to achieve agreement with specialists at a secondary school level, where subject matter is more complex.

At the same time, it is important to bear in mind that the sample size is small, particularly at the secondary school level. There were fewer paired comparisons in the secondary education phase because:

- workbooks covered two year groups (Years 8 and 9) and only one year group in history. This is in contrast to three year groups in the primary phase across subjects.
- Fewer subjects were represented in the secondary school workbooks: there were no French books at that level.

It is possible that shortening the scale in conjunction with other actions could increase reliability. One of the next steps is the revision of some descriptors, particularly those for 'Practice', to help distinguish across bands more clearly. Eliminating the indicator 'Pupils' progress' could be another one, because HMI found that it overlapped considerably with 'Building on previous learning'.

Conclusions and next steps

Research question 1: Does the piloted approach to workbook scrutiny allow meaningful assessment of the quality of education?

This study showed that the HMI could assess the quality of education by using the workbook scrutiny indicators. Having a clear focus on what to look at in pupils' work helped HMI concentrate on what was relevant. So, the indicators informed

inspectors' judgements on whether subject-specific learning was taking place. However, the HMI found it difficult to distinguish between the bands at either end of the rating scale. They raised the issue of potentially different interpretation of quantifiers such as 'some' and 'considerable'.

Research question 2: Can inspectors rate reliably using the piloted workbook scrutiny indicators?

We saw moderate levels of agreement for three indicators ('Building on previous learning', 'Depth and breadth of coverage' and 'Pupils' progress') and fair but close to moderate agreement for one of them ('Practice'). This suggests that HMI rated workbooks reliably, with the exception of 'Practice'.

The findings are indicative only, but they show that reliability is higher at a primary school than at a secondary school level. This is probably due to the fact that subject knowledge required of secondary school pupils is deeper and more specific than it is at primary school level. Subject matter expertise is therefore likely to be beneficial for workbook scrutiny in secondary schools, which is why we are producing detailed subject guidance and training for inspectors.

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.

Training, guidance materials and illustrative examples for HMI are crucial to ensuring validity and reliability of workbook scrutiny, and especially given that this is a novel approach to workbook scrutiny. We would also refresh the training (every year or every two years) to ensure that HMI make judgements in a standard and consistent manner over time.

Based on our research and discussions with HMI, we have concluded that the following factors are important for book scrutiny:

- **Structure** – We found that using indicators and rubric provided focus on scrutinising what matters, making the approach more meaningful and standardised. They help minimise the effect of potential biasing factors (for example neatness of handwriting or text length) and eliminate variability in terms of what inspectors should focus on during book scrutiny. It is the structure itself that matters rather than specific indicators.
- **Departmental and year group focus** – Focusing workbook scrutiny (as well as lesson observations) across a single subject/department/year group is helpful in securing greater validity and reliability. This is because some subjects (for example history in primary schools) may not be taught every day or may not be taught to different year groups on the same day. Hence, in some cases, it would not be possible to observe lessons within a subject area across year groups, but just within a single year group. Workbook scrutiny should go hand in hand with lesson observations and conversations

with leaders and teachers in order to allow triangulation of findings (see 'Triangulation' bullet point below).

- **Context** – Carrying out workbook scrutiny without context is likely to limit validity. Conversations with subject leaders or teachers on the purpose of tasks in the workbooks and how they contribute to learning progression can help provide that context.
- **Triangulation** – 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.

There are also certain issues to be aware of:

- Workbook scrutiny may not be possible to implement in special schools. Those schools may not use workbooks as pupils' work and progress may be captured in a different way (such as through post-it notes or videos).
- Workbook scrutiny may also not be applicable to further education and skills (FES) settings. Students in this 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.
- Pupils' work may look different in schools that use alternative methodologies in teaching and learning (for example Montessori schools) and may not necessarily be captured in workbooks.
- Modern foreign languages may not lend themselves as easily as other subjects to workbook scrutiny because a lot of classroom activity could be spoken rather than written. This points to the importance of triangulating inspection activities (such as combining book scrutiny with conversations with the subject lead and teachers, and with lesson observations).
- The amount of work in workbooks at the beginning of an academic year (for example in September, October and possibly November) may not be sufficient for inspectors to make a valid and reliable judgement about curriculum and learning progression. However, this would not be an issue if workbooks from the last few months of the previous academic year were also available for the pupils.

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. 190028