# How valid and reliable is the use of lesson observation in supporting judgements on the quality of education?

**Published:** June 2019

**Reference no:** 190029

# Contents

# Introduction

Recent work on the curriculum[1] and an overview of the research for the new education inspection framework (EIF)[2] have been central to scrutinising the validity of future inspection methods. This evidence provides parents and those working in education with assurance that inspectors will be assessing the right things when judging the quality of education. It has also delivered on one of Ofsted's core strategy promises – we remain committed to constantly improving the validity of inspection.

We have since turned our attention to lesson observation, an important inspection method within an inspector's toolkit.

In November 2017, Ofsted held an international seminar on lesson observation. The purpose of this seminar was to learn from experts from around the world on how best to develop a model of lesson observation that would be fit for purpose in the EIF. The findings of the seminar provided several aspects for us to consider for re-designing our current lesson observation instrument.[3]

Following this seminar, we developed an observation instrument specific to our inspection context. We tested this in a range of schools and colleges during the autumn term 2018. This report provides details on the findings from these research visits, particularly around the validity and reliability of the first iteration of the observation method. We were not expecting to design a perfect model first time around. However, we hoped to produce a prototype that would be a strong basis for shaping and refining how we carry out lesson visits in the future.

Initial development focused on deciding an agreed purpose for our observation protocols. The main intent was to produce a valid method that contributes to school-level evaluation. This focus is something that differs from most other lesson observation models, which prioritise teacher performance and feedback. Our unit of interest was instead a subject department or similar. We also set out to develop an observation model that worked effectively alongside the inspection process for assessing the curriculum.[4] We see lesson observation in our context, therefore, as a contributing method among many others towards judging quality of education.

---

[1] 'An investigation into how to assess the quality of education through curriculum intent, implementation and impact: Phase 3 findings of curriculum research', Ofsted, December 2018; www.gov.uk/government/publications/curriculum-research-assessing-intent-implementation-and-impact.

[2] 'Education inspection framework: overview of research', Ofsted, January 2019; www.gov.uk/government/publications/education-inspection-framework-overview-of-research.

[3] 'Six models of lesson observation: an international perspective', Ofsted, May 2018; www.gov.uk/government/publications/ofsted-research-on-lesson-observation-models.

[4] 'An investigation into how to assess the quality of education through curriculum intent, implementation and impact: Phase 3 findings of curriculum research', Ofsted, December 2018; www.gov.uk/government/publications/curriculum-research-assessing-intent-implementation-and-impact.

From these guiding principles, we produced a set of measurable indicators that offered inspectors a clear, structured system to focus lesson visits. We developed 18 indicators across three domains of interest – curriculum, teaching and behaviour – along with detailed guidance to help inspectors with assessing the indicators. The indicators were scored on a high-inference, five-point scale. These indicators were strictly for the purpose of the research study, as they allow us to carry out quantitative analysis. They were not intended for use on inspection.

Along with testing the validity of the domains and indicators, we also wanted to test for inter-rater reliability between observers. This highlights how frequently observers agree on their scoring of the indicators. We applied a paired observation method to the study design. This involved two observers observing the same lesson at the same time. We selected observers from 10 Her Majesty's Inspectors (HMI) and four researchers from our research and evaluation team. Observers carried out lesson visits between 15 and 30 minutes in length.

The study design also asked observers to carry out multiple lesson observations across two subject departments or similar during a one-day visit. This allowed us to test reliability at both the individual lesson and the subject department level, our preferred units of interest.

Both schools and colleges were included in the research so that we could identify whether the indicators and protocols developed worked across different contexts. Inspectors visited 22 schools and 15 colleges. In total, they completed 346 paired observations across 74 departments.

The findings from the research are encouraging. The evidence and analysis suggest that the observation model has good 'face validity'. This is the degree to which it appears effective in terms of its stated aims. For instance, the patterns of distribution of the indicator data across the three domains replicate patterns seen in other observations models. Exploratory factor analysis (a statistical method for uncovering the underlying structure of a relatively large set of variables) also establishes a two-factor model that partially reflects the original design intentions of our observation model.

The observers achieved a reasonable level of reliability in the school sample, given the context for which the instrument was designed. The level of consistency that we have achieved has mostly been obtained through the design of a structured system that guides observers. Reliability was enhanced further when we took other factors into consideration, such as paired observations involving two HMI.

However, reliability was considerably weaker in the college sample. The further education and skills (FES) context is likely to be incompatible with the current model design. We therefore need to develop an alternative observation model that is not associated with the school context.

Discussions with the observers and other analyses suggest several ways that we could improve validity and reliability based on these initial findings. The lack of

standardised training in using the indicators before the visits is one obvious limitation of the study. However, based on the lesson observation research literature available, we would expect a well-designed training programme to improve observer reliability further.

We realised that asking observers to score 18 indicators created a cognitive load on them. We will reduce the scale of the model so that it can be used more effectively.

Overall, this study gives us confidence that our observation model is assessing useful aspects from a lesson that can contribute evidence to a wider quality of education judgement. Further development is, of course, required as we look to turn these findings into recommendations and training models for lesson visits.

# Main findings

## Validity

- The evidence we collected suggests that inspectors were able to make valid assessments using the observation model.
- Observers were able to identify which indicators were most and least useful. This resulted in indicators for reading/numeracy strategies and assessment from the model. As our phase 3 curriculum study suggests, these aspects of curriculum are better evidenced through other inspection methods instead.
- The distribution of all the indicator scores across the three domains suggest our model has face validity. Observers scored behaviour indicators more strongly than those for teaching and curriculum. This replicates the pattern found in most lesson observation models.
- Importantly, observers were not using pupil behaviour as a proxy for deciding whether teaching in the class was effective. This gives us some confidence that the focus of the model and the guidance in the rubric were being applied as intended.
- We used exploratory factor analysis to determine whether the indicators coalesce into coherent domains that can be assessed. This identified a two-factor model – behaviour (factor 1) and a factor that included the indicators of both curriculum and teaching (factor 2). This partially reflects the original design intentions of the model (curriculum and teaching are both part of the quality of education judgement area in EIF), but also suggests observers were scoring similarly across the indicators in the curriculum and teaching domains.
- Additionally, all the indicators have strong factor loadings that are highly correlated. This suggests two things:
  - firstly, that the different indicators are measuring the same two domains and do so to a very similar extent

- secondly, that it is possible that there is a halo effect present, reflecting the overlap of the teaching and curriculum indicators in factor 2.

Analysis of the qualitative data collected from the observations supports this theory. This suggests that the traits from the model's teaching criteria often influenced scoring on the curriculum indicators. This is not entirely unsurprising, considering the unique design of curriculum in our model. Further work is therefore needed to develop how observers approach looking at the curriculum in lessons.

## Reliability

- Observations from the primary schools visited generally attained a reasonable level of reliability. We found the overall curriculum, teaching and behaviour domain statistics to be above 0.6. This meets the generally accepted rule-of-thumb for reaching substantial reliability using the metric we employed.
- The secondary school data shows that we found the overall behaviour score to have substantial reliability. The curriculum and teaching statistics for the secondary school sample achieved a moderate level of reliability.
- 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 the school sample is encouraging given that observers have reached this level of consistency without standardised training. Observers confirmed how useful having a structure in place was for ensuring that they maintained a specific focus while observing.
- Agreement in the college lessons observed was much weaker. The kappa statistics only attained a mild level of reliability or worse across all indicators. This is a disappointing, albeit useful, result. It suggests that the more variant factors that exist between the school and FES sectors reduce reliability and perhaps require a different approach for their specific contexts. We are developing this approach in the pilots for the EIF.
- One very encouraging finding shows 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 nine of the indicators, whereas only moderate reliability was reached when a non-HMI was involved.
- Data suggests that 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. Furthermore, reliability

increased in the departments visited in the afternoons. This suggests that a practise effect may be enhancing the level of consistency achieved.

- Observer feedback suggested there might be an issue with central tendency, which was confirmed through analysis of the indicator scores. The five-point scale used to score the indicators created potential for inspectors to opt for the 'middle ground'. There was less variation in observation scores at either end of the scale.
- Inspectors involved in the study stated that additional training and reducing the cognitive load of the model (because too many indicators were included in this initial iteration) were likely to result in a more consistent approach. These views are in line with what we find from the available research literature.

# Literature review

The following section provides details on some of the research literature on lesson observation that is currently available. We have used this to ground our observation model in a research frame for testing validity and reliability. This is along with the findings from our international seminar in November 2017[5] and the recent curriculum research.[6] We have therefore designed the basis for our observation model with a strong evidence base in mind.

## Validity or reliability?

Validity is core to the work of educational research but must be carefully distinguished from the notion of accuracy. For instance, the level of accuracy, or reliability, required from a research instrument should be determined by considerations of validity and not the other way around. Validity is 'a matter of assessing the right thing, in the right way, to provide accurate and useful assessment results'.[7]

Reliability is still an important aspect and is often seen as a check on validity. If it is impossible to replicate scores or marks reliably then it could be argued that it is also unreasonable to claim that anything at all is being measured, let alone what needs to be measured.[8] However, while more agreement is obviously always better than less agreement, moves towards producing perfect alignment can often undermine validity. Perfect alignment could result in oversimplifying instruments to a point

[5] 'Six models of lesson observation: an international perspective', Ofsted, May 2018; www.gov.uk/government/publications/ofsted-research-on-lesson-observation-models.
[6] 'An investigation into how to assess the quality of education through curriculum intent, implementation and impact: Phase 3 findings of curriculum research', Ofsted, December 2018; www.gov.uk/government/publications/curriculum-research-assessing-intent-implementation-and-impact.
[7] P Newton, 'An approach to understanding validation arguments', Ofqual, October 2017; www.gov.uk/government/publications/an-approach-to-understanding-validation-arguments.
[8] P Newton, Validity and validation overview, internal report for Ofsted, 2018.

where the approach is mechanistic. In turn, this could result in us measuring things that can be easily measured, rather than what we value.[9] Therefore, a delicate trade-off needs to be managed to ensure that we can reach an appropriate level of sufficiency for attaining both a credible level of validity and reliability.

## Purposes of an observation model

In the context of lesson observations, questions of validity focus on how observation systems relate to the inferences drawn from the findings they produce. Classrooms offer complex environments with a rich range of variables available for observation,[10] suggesting that observation models should seek to establish clear justifications for design decisions. For instance, not everything that can be observed from the classroom or in lessons needs including. What to include is often based on the intended use of the model to ensure that the evidence collected suitably validates any inferences being drawn.[11]

Additionally, the experts from the international seminar were clear that it was the explicit 'relationships between constructs, instruments and inferences' that are crucial design features for a valid lesson observation framework.[12] Other literature also emphasises the importance of determining validity through consideration of the purpose that the model was being developed for, the focus of the framework, how it will be put into operation and how findings will be understood.[13] The experts agreed, therefore, that it would be a mistake to pick up an off-the-shelf observation model from elsewhere and apply it wholesale to a different context.[14]

Further complexities arise from the phrase 'lesson observation' itself. This is because it denotes a wide range of activities across schools and educational research that are carried out in different forms and for different purposes.[15] For instance, improving teaching quality appears to be critical to improving schools and education systems.[16]

---

[9] G Biesta, 'Good education in an age of measurement: on the need to reconnect with the question of purpose in education', in 'Educational Assessment, Evaluation and Accountability', 21(1), 2009, pp. 33–46.
[10] E C Wragg, 'An introduction to classroom observation', Routledge, 1999.
[11] A-K Praetorius and C Y Charalambous, 'Classroom observation frameworks for studying instructional quality: looking back and looking forward', in 'ZDM', 50, 2018, pp. 521–534.
[12] C Bell, D H Gitomer, D F McCaffrey, B K Hamre, R C Pianta and Y Qi, 'An independent approach to observation protocol', in 'Educational Assessment', 17(2–3), 2012, pp. 62–87.
[13] A-K Praetorius and C Y Charalambous, 'Classroom observation frameworks for studying instructional quality: looking back and looking forward', in 'ZDM', 50, 2018, pp. 521–534.
[14] 'Six models of lesson observation: an international perspective', Ofsted, May 2018; www.gov.uk/government/publications/ofsted-research-on-lesson-observation-models.
[15] L Page, 'The impact on lesson observations on practice, professionalism and teacher identity', in M O'Leary, 'Reclaiming lesson observation', Routledge, pp. 62–74; M O'Leary and P Wood, 'Performance over professional learning and the complexity puzzle: lesson observation in England's further education sector', in 'Professional Development in Education', 43(30), 2017, pp. 573–591.
[16] H C Hill, C Y Charalambous and M A Kraft, 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', in 'Educational Researcher', 41(2), 2012, pp. 56–64; J P Allen, R C Pinata, A Gregory, A Y Mikami and J Lun, 'An interaction based approach to enhancing secondary school instruction and student achievement', in 'Science',

This has created two main focuses in the field of lesson observations that are often conflated: developing clear systems of teacher evaluation and ensuring that these systems provide feedback that allow for teacher improvement.[17] Lesson observations typically focus on the individual teacher or classroom. However, our intent was to provide indicators of practice at subject level as one aspect of quality of education more generally. This is why we opted to call our approach 'lesson visits' rather than 'lesson observations'.

The Measures of Effective Teaching (MET) project revealed that the level of certainty required from lesson observation depends on the function the observation framework is designed to deliver. For instance, the MET project carried out extensive work that demonstrated that only a modest relationship exists between lesson observations, which are designed for teacher evaluation, and pupil outcomes.[18] This has led to greater consideration being given to lesson observation as just one important tool among a range of evaluation methods for measuring teacher effectiveness. In this way, the validity of observation models is dependent on how well the evidence they collect is triangulated with a range of other evidence to make informed assessments on teaching quality.[19]

## Reliability

The following aspects are often considered to be issues of reliability:

- the number of indicators to include in an observation model
- the timing and number of observations
- the extrapolation of indicator scores to a judgement of teacher quality

333(6045), 2011, pp. 1034–1037; M Strong, J Gargani and O Hacifazlioglu, 'Do we know a successful teacher when we see one? Experiments in identification of effective teachers', in 'Journal of Teacher Education', 64(4), 2011, pp. 1–16; P Black and D Wiliam, 'Inside the black box: Raising standards through classroom assessment', King's College, 1998; D Wiliam, 'Assessment for learning: Why, what and how?', Institute of Education, 2009.

[17] E L Baker, P E Barton, L Darling-Hammond, E Haertel, H F Ladd and R L Linn, 'Problems with the use of student test scores to evaluate teachers', Washington, D.C: Economic Policy Institute, Vol. 278, 2010.

[18] T J Kane and D O Staiger, 'Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains', Bill and Melinda Gates Foundation, 2012.

[19] M Strong, J Gargani and O Hacifazlioglu, 'Do we know a successful teacher when we see one? Experiments in identification of effective teachers', in 'Journal of Teacher Education', 64(4), 2011, pp. 1–16; S Cantrell and T J Kane, 'Ensuring fair and reliable measures of effective teaching: Culminating findings of the MET projects' three-year study', Bill and Melinda Gates Foundation, 2013; http://k12education.gatesfoundation.org/resource/ensuring-fair-and-reliable-measures-of-effective-teaching-culminating-findings-from-the-met-projects-three-year-study/; Measure of Effective Teaching Project (MET), 'Gathering feedback for teaching', Bill and Melinda Gates Foundation, 2012; K Mihaly, D F McCaffrey, D O Staiger and J R Lockwood, 'A composite estimator of effective teaching', Bill and Melinda Gates Foundation, 2013; R Coe, C Aloisi, S Higgins and L E Major, 'What makes great teaching? Review of the underpinning research', Sutton Trust, 2014.

- the intended outcomes of the observation.[20]

Some of the main considerations for designing a reliable observation model follow below.

## Multiple observers

The phrase 'multiple observers' can refer to either multiple observers visiting one classroom, multiple observers visiting a series of classrooms or several observers visiting classrooms individually.[21] Some studies have linked greater reliability of results with multiple observers.[22] Observation models that focus on teacher development – such as Classroom Assessment Scoring System (CLASS), Framework for Teaching (FfT), UTeach Observation Protocol (UTOP), Mathematical Quality of Instruction (MQI) and Protocol for Language Arts Teacher Observations (PLATO) – all advocate the need for multiple observers in the observation process. Evidence from the health sector has also suggested that groups of inspectors produce more reliable assessments than individual inspectors alone.[23]

The International Comparative Analysis of Learning and Teaching (ICALT) observation instrument, which the Dutch inspectorate developed, provides an example of a model where reliability is determined by only one inspector visiting a lesson. This requires a slightly different approach to gathering evidence and achieving levels of certainty. The ICALT framework manages this through incorporating indicators that must be observable in (almost) each lesson.[24]

## Number of lessons

Reliability is also linked to the number of observations required. One-time observations have been critiqued for being open to substantial measurement error, through either a bad moment or a difficult class that may not be indicative of a teachers' typical performance.[25] There is general agreement that more observations

[20] C Bell, D H Gitomer, D F McCaffrey, B K Hamre, R C Pianta and Y Qi, 'An independent approach to observation protocol', in 'Educational Assessment', 17(2–3), 2012, pp. 62–87.
[21] R van de Lans, W van de Grift, K van Veen and M Fokkens-Bruinsma, 'Once is not good-enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations', in 'Studies in Educational Evaluation', 50(1), 2016, pp. 88–95.
[22] H C Hill, C Y Charalambous and M A Kraft, 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', in 'Educational Researcher', 41(2), 2012, pp. 56–64; Measure of Effective Teaching Project (MET), 'Gathering feedback for teaching', Bill and Melinda Gates Foundation, 2012.
[23] A Boyd, R Addicott, R Robertson, S Ross and K Walshe, 'Are inspectors assessments reliable? Ratings of NHS acute hospital trust services in England', in 'Journal of health services research and policy', 22(1), 2017, pp. 28–36.
[24] W van de Grift, 'Quality of teaching in four European countries: A review of the literature and application of an assessment instrument', in 'Educational Research', 49(2), 2007, pp. 127–152.
[25] R van de Lans, W van de Grift, K van Veen and M Fokkens-Bruinsma, 'Once is not good-enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations', in 'Studies in Educational Evaluation', 50(1), 2016, pp. 88–95.

are required when teachers are receiving developmental feedback rather than being evaluated. Research points to between three to 10 observations being enough to achieve modest reliability.[26]

The time of day and point in the year that observations take place can influence observer ratings, as well as the length of time spent in the classroom.[27] The presence of an observer may also disrupt regular routines and practices of the lesson,[28] particularly because teachers may choose to alter their practice to put on a 'fire-work display'[29] to impress an observer.[30] Multiple lesson observations seek to eliminate the impact of these variations to provide a more comprehensive and stable picture of a teachers' practice.

## Inference levels and indicators

Observation systems are typically high-inference practices by design. Yet, even taking into account their high-inference nature, observation systems exist on a spectrum of low to high inference, depending on the number and type of indicators they employ. Low-inference observation models offer greater reliance on quantitative data such as counting the number of times an event occurs within a lesson. These quantitative indicators can help to focus the eye of observers onto the same aspects of lessons. High-inference models that build in a greater collection of qualitative data[31] can allow observers to capture aspects of the lesson beyond those prescribed by indicators. Even in models that focus on collecting quantitative data, the experts at the seminar were clear that the criteria applied in these models nearly always required high-inference judgements to be made by observers.[32]

---

[26] H C Hill, C Y Charalambous and M A Kraft, 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', in 'Educational Researcher', 41(2), 2012, pp. 56–64; T J Kane and D O Staiger, 'Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains', Bill and Melinda Gates Foundation, 2012.
[27] H C Hill, C Y Charalambous and M A Kraft, 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', in 'Educational Researcher', 41(2), 2012, pp. 56–64; A J Mashburn, J P Meyer, J P Allen and R C Pianta, 'The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality', in 'Educational and Psychological Measurement', 74(3), 2014, pp. 400–422.
[28] E C Wragg, 'An introduction to classroom observation', Routledge, 1999; E C Wragg, F J Wikelely, C M Wragg and G S Haynes, 'Teacher appraisal observed', Routledge, 1996; B Jeffrey and P Woods, 'Testing teachers: the effect of school inspections on primary teachers', Falmer, 1998.
[29] G Marriot, 'Observing teachers at work', Heinemann Educational, 2001, p. 8.
[30] S J Ball, 'The teacher's soul and the terrors of performativity', in 'Journal of Education Policy', 18(2), 2003, pp. 215–228; S J Ball, 'Performativities and fabrications in the education economy: Towards the performative society?' in 'Australian Education Researcher', 27(2), 2000, pp. 1–23; B Jeffrey and P Woods, 'Testing teachers: the effect of school inspections on primary teachers', Falmer, 1998, p. 122; J Perryman, 'Panoptic performativity and school inspection regimes: Disciplinary mechanisms and life under special measures', in 'Journal of Education Policy', 21(2), 2006, pp. 147–161.
[31] W van de Grift, 'Quality of teaching in four European countries: A review of the literature and application of an assessment instrument', in 'Educational Research', 49(2), 2007, pp. 127–152.
[32] 'Six models of lesson observation: an international perspective', Ofsted, May 2018; www.gov.uk/government/publications/ofsted-research-on-lesson-observation-models.

The experts at the international seminar made a further point, that a lack of focus in the design of an observation model can lead to observers' personal biases disagreeing or collectively ignoring some aspects of lessons. The range and number of indicators that observers are trained to focus on can, therefore, guide and limit the aspects they attend to in a lesson. This can lead to observers ignoring less essential things, although increasing the number of quantitative or qualitative indicators in a framework can also increase the cognitive load on observers.[33] Assigning weight to any one indicator, while increasing the reliability of that measure, has also shown that observers can ignore or miss other important aspects within the classroom.[34] In-attentional blindness, where the focus on one aspect leads to a failure to notice other phenomena, means that even multiple observers can readily miss aspects of effective teaching.[35]

Ideally, indicators should offer the opportunity for agreement between observers – by helping them to know what to look for – but without creating complex checklists that distract from the lesson itself.[36]

### Subject-specific or generic indicators

Lesson observation frameworks sit within a continuum depending on whether the focus is generic or subject-specific. A lot of interest has focused on developing mathematics-specific frameworks.[37] Other models include UTOP, which is focused on science, technology, engineering and mathematics (STEM) subjects,[38] although a deficiency in subject-related observation instruments for numerous other subject areas remains.

Proponents of subject-specific lesson observations highlight the ways in which generic indicators, such as 'asks higher order questions', mask how teachers of different subjects and phases should adapt questioning for their pupils.[39] However,

---

[33] H C Hill, C Y Charalambous and M A Kraft, 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', in 'Educational Researcher', 41(2), 2012, pp. 56–64.
[34] K Mihaly, D F McCaffrey, D O Staiger and J R Lockwood, 'A composite estimator of effective teaching', Bill and Melinda Gates Foundation, 2013.
[35] M Strong, J Gargani and O Hacifazlioglu, 'Do we know a successful teacher when we see one? Experiments in identification of effective teachers', in 'Journal of Teacher Education', 64(4), 2011, pp. 1–16.
[36] J E Good and T L Brophy, 'Teachers' communication of differential expectations for children's classroom performance: Some behavioural data', in 'Journal of Educational Psychology', 61(5), 1970, pp. 365–374; C Tilstone, 'Observing teaching and learning', David Fulton, 1998.
[37] M D Boston and A G Candela, 'The instructional quality assessment as a tool for reflecting on instructional practice', in 'ZPD', 50, 2018, pp. 427–444; C Y Charalambous and A-K Praetorius, 'Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively', in 'ZPD', 50, 2018, pp. 355–366.
[38] C Walkington and M Marder, 'Using the Uteach observation protocol (UTOP) to understand the quality of mathematics instruction', in 'ZDM', 50(3), 2018, pp. 507–519.
[39] H Hill and P Grossman, 'Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems', in 'Harvard Educational Review', 83(2), 2013, pp. 371–384.

others question how far generic and subject-specific frameworks do differ in indicator content.[40] Furthermore, the MTP-S programme found evidence of teacher–student interactions that were linked to motivation and achievement regardless of the subject being taught.[41]

One argument for subject-specific observations is that they allow for more fine-grained observations to take place, ensuring that more detailed feedback can be provided for feedback to teachers.[42] However, for wide-scale observations taking place across a provider, this focus can increase the potential for disagreement among observers and increases training needs, because subject-experts are often needed for carrying out observations.[43]

## Observer training

Training is considered to be an essential component of a lesson observation system.[44] High-quality initial training is required to introduce observers to protocols. The training should be supported by ongoing calibration to ensure the continued reliability of judgements.[45] Ongoing calibration also allows for observation instruments to be reviewed and fine-tuned. It recognises that attempts to measure agreement levels between observers do not always guarantee agreement or ensure that observers will always produce a comprehensive picture of teaching.[46]

[40] A-K Praetorius and C Y Charalambous, 'Classroom observation frameworks for studying instructional quality: looking back and looking forward', in 'ZDM', 50, 2018, pp. 521–534.
[41] J P Allen, R C Pinata, A Gregory, A Y Mikami and J Lun, 'An interaction based approach to enhancing secondary school instruction and student achievement', in 'Science', 333(6045), 2011, pp. 1034–1037.
[42] H Hill and P Grossman, 'Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems', in 'Harvard Educational Review', 83(2), 2013, pp. 371–384.
[43] C Walkington and M Marder, 'Using the Uteach observation protocol (UTOP) to understand the quality of mathematics instruction', in 'ZDM', 50(3), 2018, pp. 507–519.
[44] H C Hill, C Y Charalambous and M A Kraft, 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', in 'Educational Researcher', 41(2), 2012, pp. 56–64.
[45] R Coe, C Aloisi, S Higgins and L E Major, 'What makes great teaching? Review of the underpinning research', Sutton Trust, 2014; S Cantrell and T J Kane, 'Ensuring fair and reliable measures of effective teaching: Culminating findings of the MET projects' three-year study', Bill and Melinda Gates Foundation, 2013; http://k12education.gatesfoundation.org/resource/ensuring-fair-and-reliable-measures-of-effective-teaching-culminating-findings-from-the-met-projects-three-year-study/.
[46] H C Hill, C Y Charalambous and M A Kraft, 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', in 'Educational Researcher', 41(2), 2012, pp. 56–64.

# Method

The main aim of this study was to design and test a prototype observation model for use in lesson visits in the EIF. The main determining factors for the model's success would be that it is both valid (it assesses the right things) and reliable (inspectors assess the right things consistently).

Importantly, we were looking for the prototype to provide a **reasonable** level of assurance that validity and reliability have been reached. As most of the models in the literature have been developed over a long period of time, we were not expecting to design a perfect model first time around. However, we were hoping to produce a prototype that would provide a strong baseline for shaping and refining the use of lesson visits in the future.

From this aim and the available literature, we developed a series of research questions for the model to test:

- What are the most important dimensions and indicators of the quality of education that can be observed and assessed by inspectors during a lesson visit?
- What structures should guide lesson visits for inspectors to elicit valid and reliable evidence to support judgements on the quality of education?
- Can observation be used effectively in contributing to departmental/subject level evaluation?

## Designing the prototype observation model

The research summary provides details on some useful approaches for developing a new observation instrument. One of the main findings from the literature is the need to develop an agreed purpose for observation protocols from the outset. This pointed towards lesson visits working best when they are tied to a clear intent and operating within strong parameters. To ensure that we built validity into the prototype model, our initial design decisions agreed the following:

- **Evaluation at the school level** – Ofsted's context is one of whole-school evaluation. Lesson visits, therefore, are one evidence base among many that contributes towards determining the quality of education at the school level. The purpose of a lesson visit is neither to form a view on an individual teacher nor to be used for teacher feedback or development purposes.[47] Instead, we chose the subject department (or similar) as our unit of interest under which lesson visits would be carried out.
- **Synthesis and triangulation** – As the model was orientated towards school level outcomes, we decided that replicating aspects of the design

[47] This is one of the rationales for removing the requirement for inspectors to grade individual teachers in September 2014.

approach used in the phase 3 curriculum study[48] would help with grouping observations together in a meaningful way. The subject department (or similar) would be the focus for triangulating evidence across observations to identify subject quality and contribute to a school-level overview. This would also mean that lesson visits would not be randomly carried out across a provider.

- **Structured focus** – The model needed a structure to aid consistency between inspectors. Scaled indicators covering measurable features of a lesson, such as teaching quality and classroom management, were essential to the study design. The inclusion of indicators would ensure that observers focused on the most important elements across lessons and would allow us to carry out quantitative analyses.
- **High inference** – Owing to the implications on validity, we needed to avoid a straightforward tick-box approach to data collection. The high-inference nature of observation means that we still required inspector professional judgement, therefore scoring the indicators would be enhanced through providing inspectors with a detailed set of instructions to improve consistency.
- **Importance of teaching** – There was a consensus among the experts at the international seminar that 'learning' cannot be directly observed and should be determined through other evidence collection methods. Observation should, instead, take on a greater focus on **teaching** because this tends to be observable and measurable. We therefore decided that our model would include teaching as a core domain on this basis.
- **Generic indicators** – Owing to the paucity of observation models that extended beyond mathematics and literacy, the decision taken for this prototype was to focus initially on generic indicators. These are more readily available and can be assessed across most types of lesson to determine quality.

We then turned our attention to developing a set of measurable indicators that offered a clear framework to standardise the activities of observers.

Along with the misconception that learning is something that can be observed, we have tried to avoid building several perceived proxies for learning into the model. This meant that pupils' attitudes, particularly whether they were engaged or motivated, were not included. We made this decision on the basis that even though pupils may appear engaged in a lesson, it does not mean they have learned anything. Therefore, it was unclear what measuring engagement or motivation would tell us about the quality of education on offer.

[48] 'An investigation into how to assess the quality of education through curriculum intent, implementation and impact: Phase 3 findings of curriculum research', Ofsted, December 2018; www.gov.uk/government/publications/curriculum-research-assessing-intent-implementation-and-impact.

Instead, we developed the indicators for the model based on factors that have been found to be valid and reliable in existing lesson observation research. We have assumed that, because these indicators were considered reliable across multiple scenarios, they would work within our inspection context. We also assumed that they would tell us something about the quality of education. The experience of the two HMI leads on the study also contributed to the indicator development. This process generated two specific domains of interest – the **quality of teaching** and **behaviour/classroom management**[49] – and the inclusion of six indicators for the former and four indicators for the latter.

A third domain was also developed based on the indicators used in the phase 3 curriculum research.[50] This was so that we could test whether lesson visits offered valid evidence on the **curriculum** through observation. If this proved to be the case, it would give inspectors an additional method for assessing the effectiveness of a provider's curriculum. Using the phase 3 curriculum findings as a starting point, we developed eight indicators. Wider lesson observation research was not available to inform curriculum indicator development because its focus tends to be on subject-specific aspects of lessons.

[49] We expected that the behaviour/classroom management domain would tell us something about the quality of education (is the classroom managed well to facilitate good teaching?) but would also feed into the behaviour judgement in the EIF.

[50] 'An investigation into how to assess the quality of education through curriculum intent, implementation and impact: Phase 3 findings of curriculum research', Ofsted, December 2018; www.gov.uk/government/publications/curriculum-research-assessing-intent-implementation-and-impact.

**Figure 1: List of indicators used in the lesson observation model**

| No. | Indicator |
|---|---|
| **1** | **Curriculum** |
| 1a | Teachers use subject expertise, knowledge and practical skills to provide learning opportunities |
| 1b | Teachers ensure there is an equality of opportunity for all learners to access every lesson, as building blocks to the wider curriculum |
| 1c | Strategies to support reading/vocabulary understanding/numeracy are in place for pupils who need it/cannot access the curriculum |
| 1d | The content of the lesson is suitably demanding |
| 1e | The lesson content is appropriate to the age group and does not lower expectations |
| 1f | There is a logical sequence to the lesson |
| 1g | Teachers provide opportunities to recall and practise previously learned skills and knowledge |
| 1h | Assessment provides relevant, clear and helpful information about the current skills and knowledge of learners |
| **2** | **Teaching** |
| 2a | Teachers demonstrate good communication skills |
| 2b | Teachers' use of presentation allows pupils to build knowledge and make connections |
| 2c | Teachers use relevant and appropriate resources during presentation to clarify meaning to pupils |
| 2d | Teachers possess good questioning skills |
| 2e | Teachers give explicit, detailed and constructive feedback in class |
| 2f | Teachers effectively check for understanding |
| **3** | **Behaviour** |
| 3a | Teachers create supportive classrooms focused on learning |
| 3b | Teachers create focused classrooms through their high expectations for pupils |
| 3c | Teachers communicate clear and consistent expectations which are understood and followed |
| 3d | Pupils' behaviour contributes to the focus on learning |

Based on the model used in the phase 3 curriculum research, we developed a detailed rubric to guide inspectors in making informed assessments against each indicator on a scale of one to five. Evidence relating to a score of five on an indicator would show that this was a strength of the lesson and embedded into practice, whereas a score of one would indicate that there are major weaknesses in this aspect. The rubric therefore provided a systematic structure around the indicator design to ensure that inspectors focused on these domains during an observation. It would also assist inspectors in making consistent judgements.

The scoring was on a five-point scale for two reasons:

- First, the scale avoids reference with current Ofsted judgements. Observers would need to engage with the guidance set in the detailed rubric to make accurate assessments on the indicators.

- Second, the scale increases the variability in scoring options for testing the reliability of the indicators.

A 'not applicable' option was also included for each indicator. If elements for an indicator were not observed during a lesson, observers were expected to mark the score as N/A instead. Figure 2 provides further details of the categories each number represents on the five-point scale:

**Figure 2: Categories applied in the rubric for scoring the indicators**

| 5 | 4 | 3 | 2 | 1 | N/A |
|---|---|---|---|---|---|
| This aspect is embedded in practice (many examples of exceptional teaching) | This aspect is embedded with minor points for development (leaders taking action to remedy minor shortfalls) | This aspect is sufficient but there are some weaknesses overall in a number of examples (identified by leaders but not yet remedying) | Major weaknesses evident (leaders have not identified or started to remedy weaknesses) | This aspect is absent in practice | Unable to score this indicator as not observed in the time provided |

## Paired observer design

An important aspect of the study was to test for observer consistency during the lesson visits. Based on the available literature, the structured and focused design of our observation instrument would likely enhance both the validity of what we were assessing and aid inspector consistency. However, to be completely sure of the latter, the research visit was also designed so that we could assess the inter-rater reliability of observers' use of the model.

Paired observation was, therefore, the required approach across the lessons visited for the study. This involved two observers observing the same lesson at the same time, with the aim to see how regularly they reached agreement across the indicators in the model. Observers in the study consisted of 10 HMI and four researchers from Ofsted's research and evaluation team (non-HMI). The involvement of non-HMI in the study design was partially down to scheduling. HMI were not always available on the same days to carry out a paired observation. In these circumstances non-HMI were used instead. One benefit of non-HMI involvement in the study was that scoring consistency between the two participating groups could be analysed.

We gave observers proformas and asked them to score each of the 18 indicators separately, along with recording qualitative evidence to support their scoring. This method would allow us to test statistically the inter-rater reliability between paired observers. We also asked observers to give an overall domain score for each of the three domains. This was applied through synthesis of the indicators under each domain to provide a best fit of the evidence against the rubric categories.

## Independent scoring protocols

We designed the paired observation process to be carried out independently by the observers. We minimised their contact time during the observation so that independent assessments could be made on the indicators by each observer. The sharing of views between observers could, in this context, bias the level of agreement achieved. Therefore, we applied two methods to minimise this risk. First, a protocol was developed setting out the expectations for observers. This ensured that they understood the rationale for this aspect of the study and could implement it effectively. Second, we allocated neutral observers to some of the provider visits to ensure that the approach was being carried out as intended and with integrity.[51]

## Synthesising across multiple observations

One aim of the study, beyond looking at general reliability, was to determine whether the process of synthesising lesson data could give a valid and consistent method for assessing the quality of a subject department or similar.[52] The rationale for this was that a departmental overview would move the focus away from individual lessons (and, linked to that, teacher evaluation) towards an appropriate evaluation of the quality of education being provided across a department. We also felt that minimising the variability in approach, by making 'scatter-gun' observations across multiple departments less prevalent, would improve reliability.

The study design therefore required observers to carry out multiple lesson observations across two subject departments or similar during a one-day visit. Typically, observers would view the lessons for one department in the morning of the visit, with the afternoon visits focusing on the second department. This was to make the department the unit of interest, not the individual lessons. We asked observers to give an overall department score for each of the indicators alongside each of the individual lesson scores collected within a department. The department score would be derived from a synthesis of the evidence collected at the individual lesson level. This would allow us to analyse differences in inter-rater reliability between the individual lesson and department level.

A suitable sample of lessons were required on each visit for observers to assess the quality of each indicator at the departmental level. It was agreed that between four to eight lessons across a department would be suitable for synthesis. In larger departments (above six teachers), the protocol was to see teachers only once. In smaller departments (below six teachers), necessity meant teachers could be seen more than once.

[51] Neutral observers were selected from researchers in Ofsted's research and evaluation team. This was applied in six of the visits involving two HMI observers.

[52] In general, 'departments' refer to subject departments. However, it was not always possible to view a subject owing to specific school contexts. In these cases, a year group, phase of education or theme/aspect was identified as the unit of interest to focus observation on instead.

Observers selected the subject departments to investigate in agreement with school leaders. This was often based on what was being taught on the day (particularly for the primary schools in the sample), although considerations for reducing sample attrition were also factored in. That is, if school leaders were keen for us to look at a specific department then that was typically built into the visit programme.[53] One condition we did specify was that only one core subject be selected as part of the selection process for the departmental focus.

### Other design decisions

Observations ranged between 15 and 30 minutes in length. This provided observers with the flexibility to remain in a lesson if they felt this was required to score the indicators appropriately. Additionally, the domains and indicators in our model were designed so that they could be observed at any point in a lesson, providing further flexibility to the observation length required.

Inspectors were also asked to carry out some additional activities before and after the observation. The main reason for this was to help inspectors with accurately assessing the curriculum domain of the model. It was felt that this domain may be particularly tricky to score accurately without some degree of context beforehand. For instance, it was felt that some understanding of the curriculum aims in the subject would allow inspectors to better evaluate delivery across a department. Therefore, inspectors spent a short amount of time with subject leads at the beginning of the visit to discuss their pupils' progression through the curriculum. They also spoke to the teachers and a small number of pupils they had observed about where this lesson fitted into their curriculum progression. Inspectors were advised to carry out these discussions before or after an observation and not during, as this could otherwise distract inspectors from directly observing teaching and behaviour.

## Sample

As this study was looking to inform inspection methodology in the new EIF it was important to understand how the lesson observation model worked across different phases of education. We decided, therefore, to involve both schools and colleges in the fieldwork to see if the indicators and protocols work similarly across different contexts.[54]

Our sampling approach was to construct a convenience sample for the fieldwork. This was for the following reasons:

- there was no theoretical basis to suggest that findings from lesson observation would drastically differ across regions

[53] It is worth noting that this will not be the process on inspection, but was a method applied during the research to provide schools with flexibility to warrant their participation in the research.
[54] Observation in the college sample was carried out by further education and skills HMI. Observation in the primary and secondary school sample was carried out by schools HMI.

- we were not attempting to evaluate whether lessons were better delivered in one context over another, making sampling on a representative basis unnecessary
- because multiple observations per visit were possible, this removed the need for many providers to be included in the sample to test out the model
- a more complicated statistical sampling approach would have made it difficult to allocate two observers per visit to meet the reliability component of the study design.

We drew the convenience sample from the available pool of providers by provider type and previous inspection judgement (outstanding, good and requires improvement (RI) schools only). This was so that some variability and balance existed in the sample.

We visited 22 schools and 15 colleges during the fieldwork. The school sample consisted of 10 primary, 10 secondary, one special school and one pupil referral unit (PRU). The college sample consisted of 11 further education colleges, three sixth form colleges and one independent specialist college. Independent learning providers were out of scope for this study.

Fourteen of the provider visits involved an HMI/non-HMI pairing. The other 23 visits were carried out by a pairing of HMI.

## HMI focus groups

It was important to gather the views of the HMI on the processes and experiences of carrying out these observations, as well as to get their feedback on the usefulness of the indicators and rubric. The HMI were split into two focus groups for these discussions: one of further education HMI; the other of schools. The data from the observations had not been analysed at the time of the focus groups, so the feedback was directly based on the HMI experience of using the observation model. Researchers took notes during the focus groups to compare the views recorded with the statistical analysis of the indicator scores.

## Training

Inspectors attended a one-day training session that focused on the observation instruments and rubric before carrying out visits. This developed their understanding and applied use of the observation model to increase consistency. However, the training involved did not intend to replicate the processes of other lesson observation frameworks, which tend to include extensive standardised training in the use of their indicators and observation tools.

One purpose of this study was to identify an initial baseline on the effectiveness of the model, centred solely on how inspectors applied the instrument and methodology. The levels of reliability attained would, therefore, help with specifying

the degree of standardised training required to enhance consistency in future iterations.

## Limitations

- The number of indicators included in the model was intended to offer broad coverage across these three domains. However, this was just our starting point. One purpose of the study was to identify the most useful indicators and refine the model for lesson visits in the new inspection framework. We realised that 18 indicators may also add to observer cognitive load in applying the indicators and guidance correctly.
- The limited amount of training time available may have reduced the level of reliability achieved. However, we needed to understand how well the design of the indicators and rubric worked in isolation of training. Establishing a baseline for the model would allow us to understand which aspects of training would really benefit inspector consistency in the future.
- This study looks at the reliability between HMI but does not feature reliability between Ofsted Inspectors or compared with HMI. Differences between these groups may exist. This is something we will evaluate as the model is further developed.

While designed to test for reliability in a conventional research sense, the study design does not necessarily replicate the processes of routine inspection. In general, multiple inspectors on a routine inspection of a large provider would be able to discuss outcomes from observation between them to help inform their judgements. It is possible, therefore, that some of the inter-rater reliability scores that follow may under-represent the level of reliability achieved due to a lack of inspector collaboration.

## Evaluation

After completing the fieldwork, we collated the scores from the evidence forms into a data-set. In total, 346 paired observations were completed across 74 departments. This data was cleaned before being analysed in the statistical product R. Cleaning mostly involved changing blank entries provided by inspectors to an N/A score, as we assumed this was something that inspectors were unable to observe during the lesson.[55]

We used this data-set to carry out several analyses to establish the validity and reliability of the indicators. Model validity was determined through an exploratory factor analysis. Reliability was examined using Cohen's kappa as the statistic to

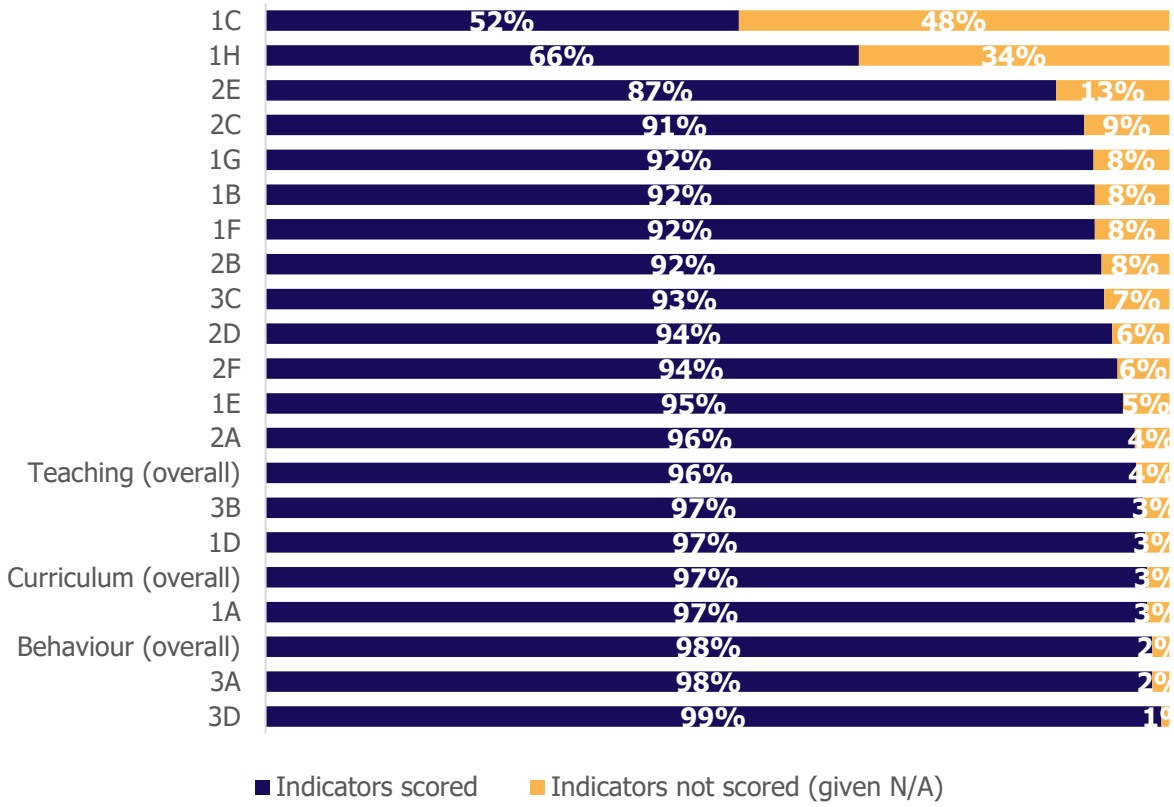
[55] Owing to missing data, the quantitative analyses carried out do not always correspond to the number of individual lesson visits and departments that contributed to the study.

compare observers' inter-rater reliability across lesson and department levels. We also looked at other variables to see what effect this had on inspector reliability.

## How valid is the lesson observation model?

One way of understanding the validity of our observation model is to look at descriptive statistics of the indicator scores to see what this can tell us about it.

**Figure 3: Proportion of scored and non-scored indicators across 346 paired lesson observations**

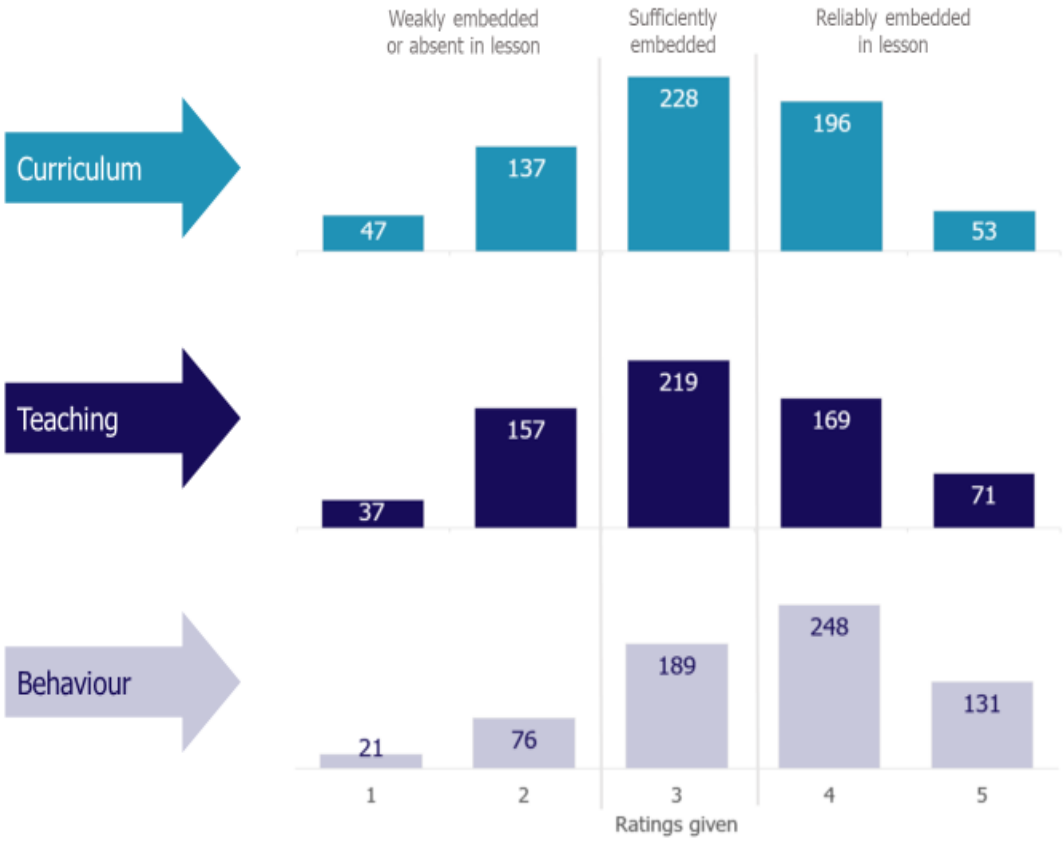



First, we investigated whether inspectors had difficulties with assessing any of the indicators. This helps with determining whether the indicators measured what they were intended to measure. Figure 3 shows that inspectors clearly struggled to provide scores on two of the indicators in the curriculum domain. Strategies to support reading (indicator 1c) and determining whether assessment provides relevant information about the current skills and knowledge of learners (indicator 1h) commonly received a 'not applicable' response from inspectors.

The HMI involved confirmed that both indicators proved difficult to score, because these were aspects that did not regularly feature during observations. Additionally, some HMI indicated that, although they did provide a score for these indicators, they felt the evidence secured from the observation process was too flimsy to match against the rubric with any real accuracy. In fact, it was suggested that lesson observation is not the ideal way for assessing either of these indicators and that the evidence required, as our phase 3 curriculum study proved, probably comes from

sources outside of lessons. As such, these two indicators fall out of our lesson observation model.

**Figure 4: Observation scores for all indicators grouped by curriculum, teaching and behaviour domains**




A further way of assessing the validity of our model is to compare the data collected with that of other lesson observation models. Figure 4 shows the distribution of all the indicator scores across the three domains.

Observers typically scored behaviour indicators more favourably, as shown by the skew in the distribution towards a score of 4 and 5 on the rubric. The teaching and curriculum indicators, by comparison, feature a distribution centred around a score of 3. This shows a similar pattern of scoring to that found in most other observation models, in which classroom management tends to be scored more highly than across other teaching-related domains. Importantly, this also shows that observers were not typically using pupil behaviour as a proxy for whether teaching in the class was effective. This gives us some confidence that the observers were applying the focus of the model and the guidance in the rubric as intended.

We also carried out exploratory factor analysis on the indicator data to determine whether they coalesce into coherent domains that inspectors can assess. Based on the data from Figure 4, we assumed that the differing response patterns of inspectors' scoring on behaviour, compared with the teaching and curriculum

domains, would at the very least identify two important factors in the model.[56] Figure 5 provides the factor loadings for each indicator.

**Figure 5: Results of factor analysis carried out on 18 indicators and three overall domain scores**

| Indicator | Factor loading | |
|---|---|---|
| | **Factor1** | **Factor2** |
| Curriculum overall | 1.00 | |
| Teaching overall | 0.95 | |
| Behaviour overall | | 1.03 |
| 1a | 1.00 | |
| 1b | 0.92 | |
| 1c | 0.84 | |
| 1d | 0.91 | |
| 1e | 0.86 | |
| 1f | 0.89 | |
| 1g | 0.92 | |
| 1h | 0.84 | |
| 2a | 0.85 | |
| 2b | 0.93 | |
| 2c | 0.87 | |
| 2d | 0.85 | |
| 2e | 0.89 | |
| 2f | 0.90 | |
| 3a | | 0.91 |
| 3b | | 0.79 |
| 3c | | 0.93 |
| 3d | | 0.96 |

This data does indeed identify two factors – behaviour (factor 1) and a factor that included the indicators of both curriculum and teaching (factor 2). This partially reflects the original design intentions of the model, but suggests that observers were scoring similarly across the indicators in the curriculum and teaching domains.

All the indicators have strong factor loadings that are very highly correlated (the correlation matrix can be found in Annex A). This suggests two things. Firstly, the different indicators measure the same two domains and do so to a very similar extent. Secondly, it is possible that there is a halo effect present, reflecting the overlap of the teaching and curriculum indicators in factor 2.

[56] When multiple (observed) variables have similar response patterns because they are all associated with something else (which is not directly measured), the 'something else' can be thought of as a factor.

Analysis of the qualitative data collected from the observations supports this theory, suggesting that the traits from the teaching criteria in the model often influenced scoring on the curriculum indicators. HMI views on the practical use of the instrument also corroborate this view. 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 we had originally envisaged in our design. This is not entirely unsurprising, considering the unique design of curriculum in our model. Further work is required to develop how observers approach looking at the curriculum in lessons, although our phase 3 curriculum research[57] highlights that lesson visits will not be the only way we will assess curriculum under EIF.

Finally, inspector feedback from the focus groups helps to explain why this distinction between the behaviour and teaching/curriculum domains exists. It was felt that focusing separately on the behaviour, teaching and curriculum constructs told them very different things about the overall quality of a lesson. For instance, factors such as the age of learners, school culture and school intake appeared to contribute to behaviour in lessons, so that behaviour was clearly not the product of the lesson. In some cases, strong scores in the behaviour domain could simply reflect pupils' compliance with a schools' behaviour policy. However, the quality of pupils' behaviour told inspectors very little about the progress being made, pupil engagement and motivation or the level of challenge in the lesson. So, while the behaviour indicators were generally advocated as the easiest to observe and score, they were also seen by inspectors as a poor proxy for assessing the quality of teaching. It was not uncommon to view lessons in which pupils' behaviour was exemplary, yet the quality of teaching observed was particularly poor.

The evidence collected suggests that inspectors were able to make valid judgements in these areas. We have identified the indicators that do not work in practice and should be removed from the model. We have also replicated patterns seen in other observation models, which is encouraging. The model also distinguishes between the indicators to confirm that pupil behaviour and quality of teaching are not mutually exclusive. This provides some confidence that the instrument we have developed is measuring important aspects of the domains we intend to measure.

## How reliably did inspectors score the indicators?

The design decisions of our model, such as providing a structure around what is important to observe in a high inference design, will have some implications on the level of reliability achieved, as does the purpose of our model. Lesson observation, in the EIF, will be part of a suite of methods available for the deep-dive process to aid inspectors in evaluating the quality of education. Valid outcomes will be generated from synthesising the evidence that comes from these independent activities. This

[57] 'An investigation into how to assess the quality of education through curriculum intent, implementation and impact: Phase 3 findings of curriculum research', Ofsted, December 2018; www.gov.uk/government/publications/curriculum-research-assessing-intent-implementation-and-impact.

means that lesson observation alone is not the main factor of a judgement in the same way it is for observation models that are more concerned with teacher performance management. The purpose of our lesson observation model suggests, therefore, that failing to achieve almost perfect reliability is not necessarily a problem.

We used Cohen's kappa as the statistic to measure inter-rater agreement between paired observers using our research instrument. The kappa coefficient is applicable for categorical or ordinal data. It is generally seen as a stronger measure than a simple percentage agreement. This is because its calculation takes into account whether the agreement reached has occurred by chance. The values of the coefficient range from 1, where there is exact agreement, to 0 where there is no agreement. A negative kappa suggests that the inter-rater reliability is worse than when the ratings were produced by chance. Early research in the use of kappa coefficients to measure inter-rater agreement[58] 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). For the purposes of our analyses, Figure 6 shows how we have interpreted the level of agreement reached for the kappa coefficients calculated.[59]

**Figure 6: Cohen's kappa interpretation**

| Kappa statistic | Agreement |
|---|---|
| $0<x\leq0.2$ | Slight |
| $0.2<x\leq0.4$ | Fair |
| $0.4<x\leq0.6$ | Moderate |
| $0.6<x\leq0.8$ | Substantial |
| $0.8<x\leq1$ | Almost perfect |

The structure and focus of the research model, along with the use of lesson observation for school evaluation purposes, suggested that, as a rule of thumb, a substantial level of agreement (kappa statistic >.6) would be a good level of reliability for our observers to achieve. We would expect to achieve a higher level of reliability with standardised training in place.

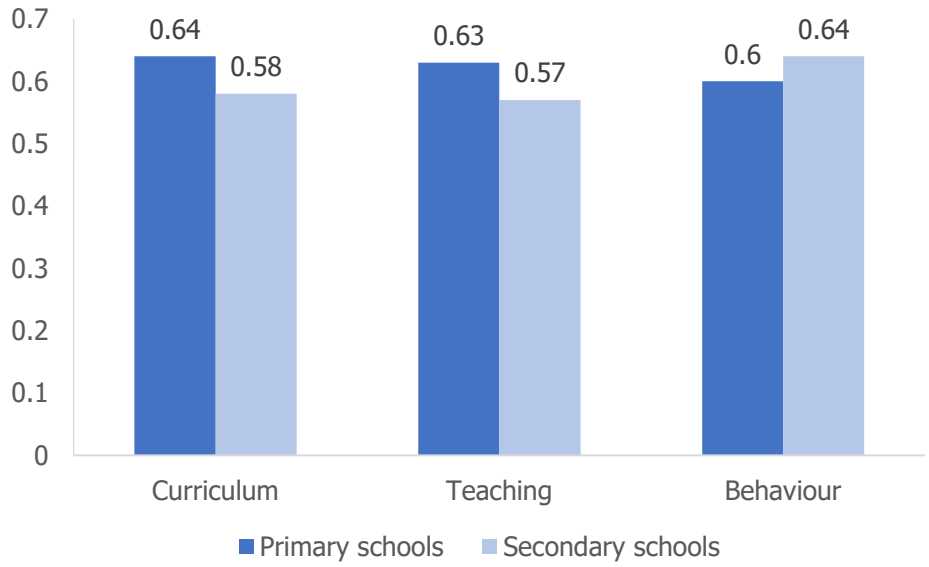
The reliability analyses that follow are based on a weighted linear kappa, which excludes the N/A scores from the fieldwork.

[58] J Landis and G Koch, 'The measurement of observer agreement for categorical data', Biometrics, 33(1), 1977.
[59] M McHugh, 'Interrater reliability: The kappa statistic' in 'Biochemia Medica', 22(3), 2012, pp. 276–282.

## Reliability in the school sample

Figure 7 shows that inter-rater reliability between observers was good in both the primary and secondary schools' sample.

**Figure 7: Inter-rater reliability of overall domain lesson observation scores, by primary and secondary schools**




Secondary school data includes observations from the special school and PRU.
Primary school data is based on 90 observations. Secondary school data is based on 107 observations.

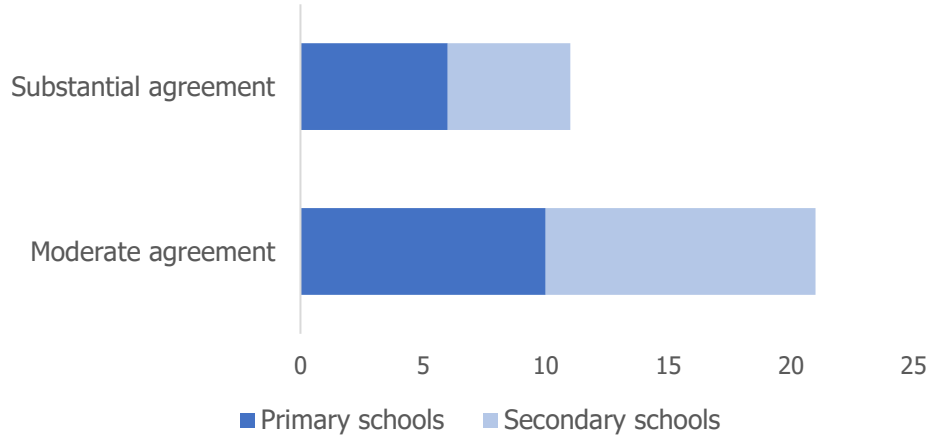
The data indicates that a substantial level of agreement was reached across each of the overall domain scores provided by observers in the primary school sample.

Observation outcomes from the secondary schools were found to be slightly lower, although the overall domain score for behaviour also reached a substantial level of agreement at 0.64. This suggests that observers may find it slightly easier to agree on the quality of pupils' behaviour and teachers' behaviour management strategies in secondary schools than in primary schools.

The statistics for the 16 indicators feeding into the overall domain scores also appear to have a high level of reliability.[60] Figure 8 shows that six indicators from the primary observations and five from the secondary sample achieved a substantial level of reliability. The distribution of the kappa statistics for the 16 indicators in the primary school observations ranged from 0.46 to 0.69. For the secondary school observations, the distribution was similar and ranged from 0.48 to 0.62.

[60] Indicator 1c and 1h have been excluded from the reliability analysis as they have already been determined as invalid indicators.

**Figure 8: Number of indicators reaching differing levels of agreement between observers, by phase at the lesson level**




Secondary school data includes observations from the special school and PRU.
Primary school data is based on 90 observations. Secondary school data is based on 107 observations.

Of the six indicators reaching a substantial level of agreement from the primary school observations, three were in the curriculum domain, two were in the teaching domain and one was in the behaviour domain. These indicators are listed below:

- Teachers use their subject expertise to provide effective learning opportunities (indicator 1a).
- The content of the lesson is suitably demanding (indicator 1d).
- The lesson content is appropriate to the age group and does not lower expectations (indicator 1e).
- Teachers demonstrate good communication skills (indicator 2a).
- Teachers give explicit, detailed and constructive feedback in class (indicator 2e).
- Teachers communicate clear and consistent expectations which are understood and followed by pupils (indicator 3c).

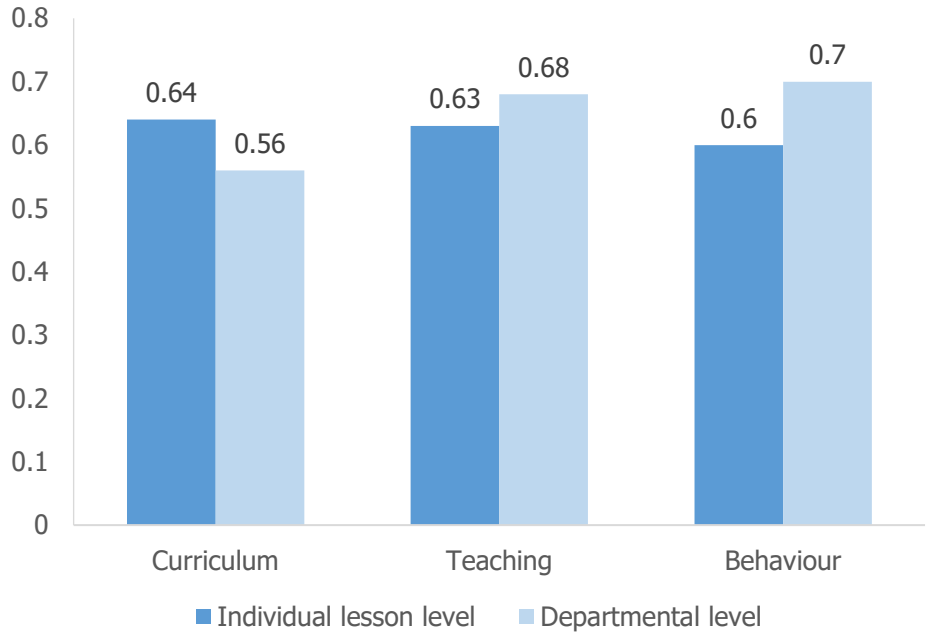
Teachers' subject knowledge had the highest level of inter-rater agreement (0.69) in the above indicators. Indicators 1e, 2a and 3c also attained substantial reliability in the secondary school observations. The following two indicators from the curriculum and behaviour domains also achieved a substantial level of reliability in the secondary school observations:

- Teachers ensure there is an equality of opportunity for all learners to access every lesson (indicator 1b).
- Teachers create supportive classrooms focused on learning (indicator 3a).

Interestingly, agreement on teachers' subject knowledge appeared more difficult for observers to deduce in the secondary school sample, as this only achieved a kappa score of 0.54. This suggests that as specialisation in a subject area increases, it

perhaps becomes more difficult for non-specialists in the subject to agree on what sufficient subject knowledge may look like in a lesson visit.

**Figure 9: Comparison of individual lesson and department level inter-rater reliability domain scores in the primary school sample**




Analysis of the data at the department level – where observers synthesised evidence from multiple lessons across a subject department, or similar, to form indicator scores – also shows some consistency. Figure 9 compares the individual lesson data from the primary school observations with the synthesised lessons scores of 19 departments. This shows that while the overall curriculum score is lower, the overall teaching and behaviour domains are higher at 0.68 and 0.7, respectively. This suggests that observers were generally more comfortable with corroborating evidence across multiple lessons in the two domains they are more used to observing.

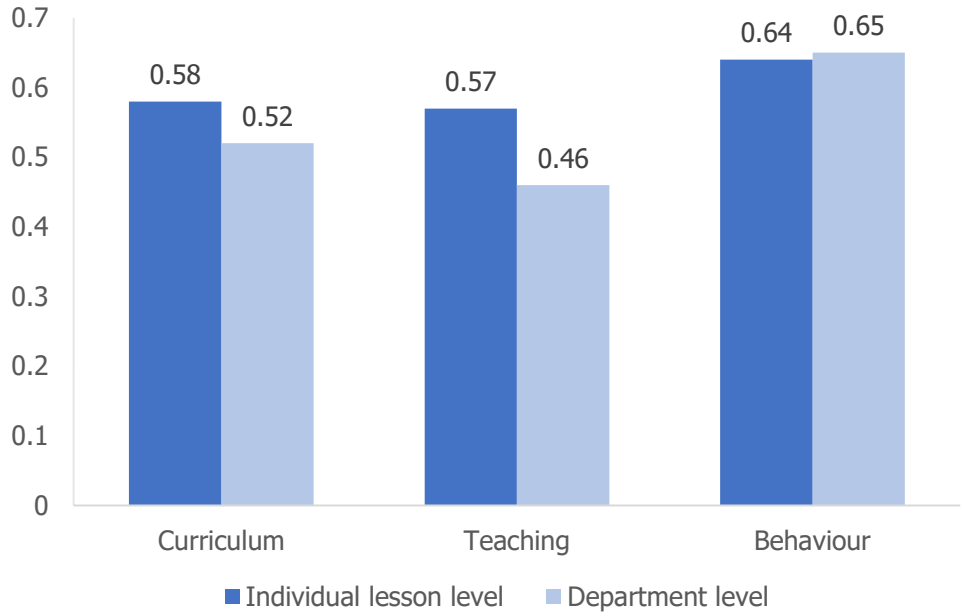
There were six indicators reaching a substantial level of agreement from the primary school departmental synthesis. This included the same three curriculum indicators and one teaching indicator from the individual lesson analysis:

- Teachers use their subject expertise to provide effective learning opportunities (indicator 1a).
- The content of the lesson is suitably demanding (indicator 1d).
- The lesson content is appropriate to the age group and does not lower expectations (indicator 1e).
- Teachers give explicit, detailed and constructive feedback in class (indicator 2e).
- Teachers effectively check for understanding (indicator 2f).

- Pupils' behaviour contributes to the focus on learning (indicator 3d).

This pattern was not replicated by observers when synthesising the data across 20 secondary school departments. Figure 10 shows that inter-rater agreement was weaker in the curriculum and teaching domains. This is particularly when compared with the individual lesson level scores, although the behaviour domain was equally strong.

**Figure 10: Comparison of individual lesson and department level inter-rater reliability domain scores in the secondary school sample**

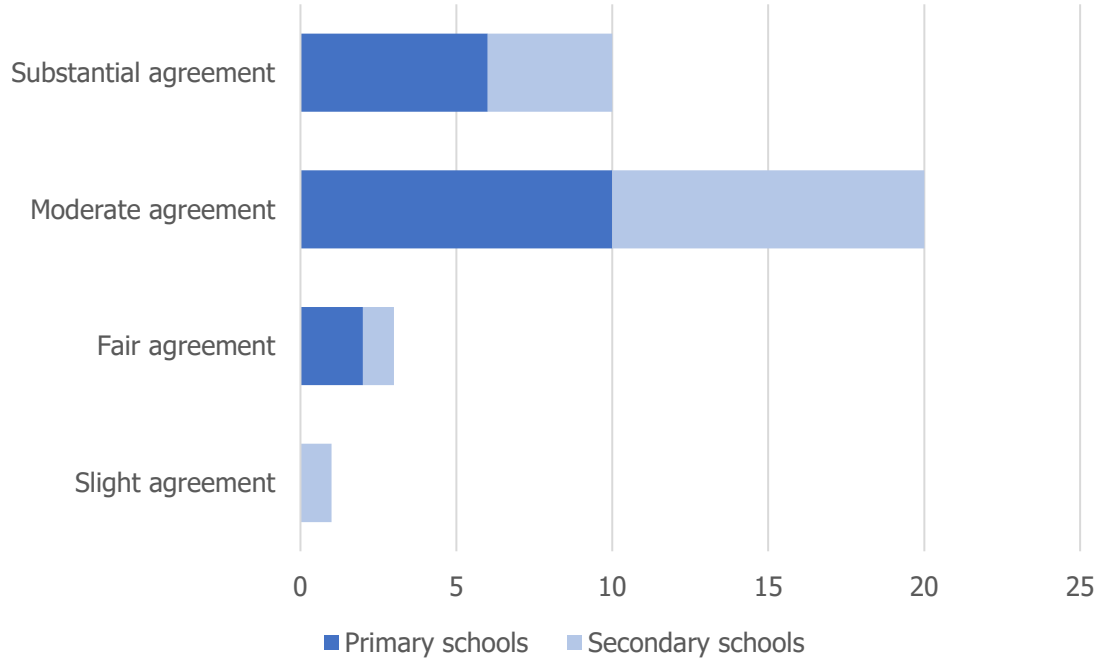



Secondary school data includes observations from the special school and PRU.

Although the overall reliability scores are lower at the department level for the secondary sample, this was a new unit of interest for observers to assess. It required the mastery of a different process that observers would not have previously used. The HMI in the focus groups spoke about the usefulness of having the department as the unit of interest and felt that this was beneficial in allowing them to develop an overview of the provision. However, they tended to disagree on whether the departmental scores should be arrived at through a process of aggregation (essentially averaging the indicator scores from individual lessons to derive a departmental score) or through a more holistic process of synthesis. This suggests a conflict in approach that may have affected the reliability of departmental scoring in the secondary phase.

Figure 11 shows the inter-rater reliability distribution of the 16 indicators at the departmental level. This highlights the fact that, despite the lower reliability across the three overall domain scores, variation exists across the indicators in the secondary sample. Fourteen indicators, for instance, still attained a moderate or higher level of reliability.

**Figure 11: Number of indicators reaching differing levels of agreement between observers, by phase at the department level**




Secondary school data includes observations from the special school and PRU.

One further reason for the lower level of reliability at the department level could, therefore, involve observers placing greater on less valid indicators when generating an overview across a series of lesson observations. For instance, Figure 12 shows that, despite the decrease in the overall curriculum score between individual and departmental level evidence in the primary school sample, the same three curriculum indicators attained a substantial level of agreement. One interpretation may be that observers placed more weight on the less reliable curriculum indicators when synthesising evidence for the overall curriculum score, although other latent variables may also be responsible.

**Figure 12: Comparison of curriculum indicators at the individual lesson and department levels in the primary school sample**

| Indicator | Lesson | Department |
|---|---|---|
| Curriculum overall | 0.64 | 0.56 |
| Subject expertise (1a) | 0.69 | 0.61 |
| Equality of opportunity (1b) | 0.49 | 0.48 |
| Demanding lesson content (1d) | 0.62 | 0.65 |
| Appropriate content for age (1e) | 0.61 | 0.66 |
| Logical sequencing (1f) | 0.53 | 0.57 |
| Recall and practise (1g) | 0.51 | 0.49 |

Indicators 1c and 1h not included as earlier analysis shows them to be less valid.
Indicators in green show those with a substantial level of inter-rater reliability.

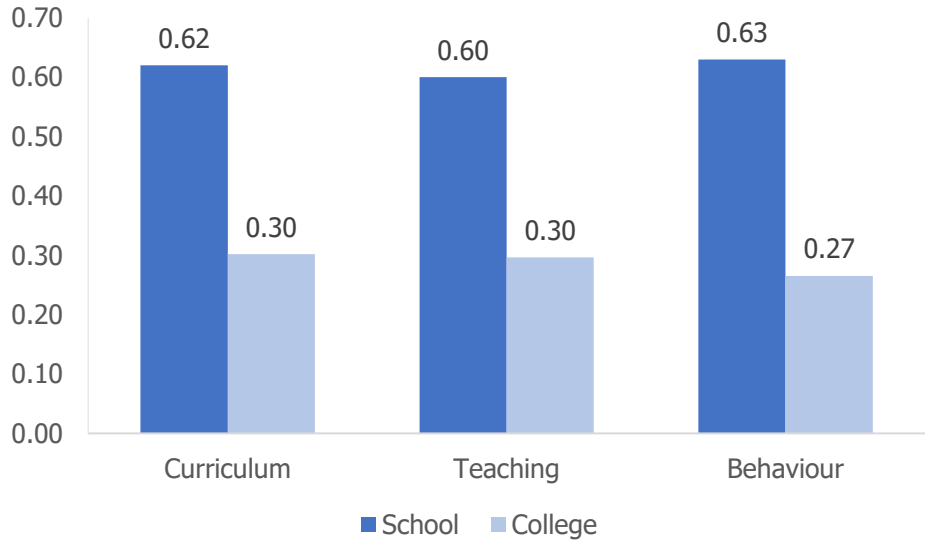
Narrowing the indicators down to those that are the most valid should help with consistency around evidence synthesis in the future. For instance, it is likely that

fewer indicators will reduce cognitive load on observers – from both the act of observing and synthesising evidence to the unit of interest.

## Reliability in the college sample

Figure 13 shows that differences in inter-rater reliability between observers in the schools and college samples was particularly stark. Agreement between observers was generally much less reliable than in the school's sample. The kappa statistics show only a fair level of agreement achieved between observers on the overall domain scores. This was also a common feature across the 16 indicators contributing to the overall domain scores.

**Figure 13: Inter-rater reliability of overall domain lesson observation scores, by phase of education**



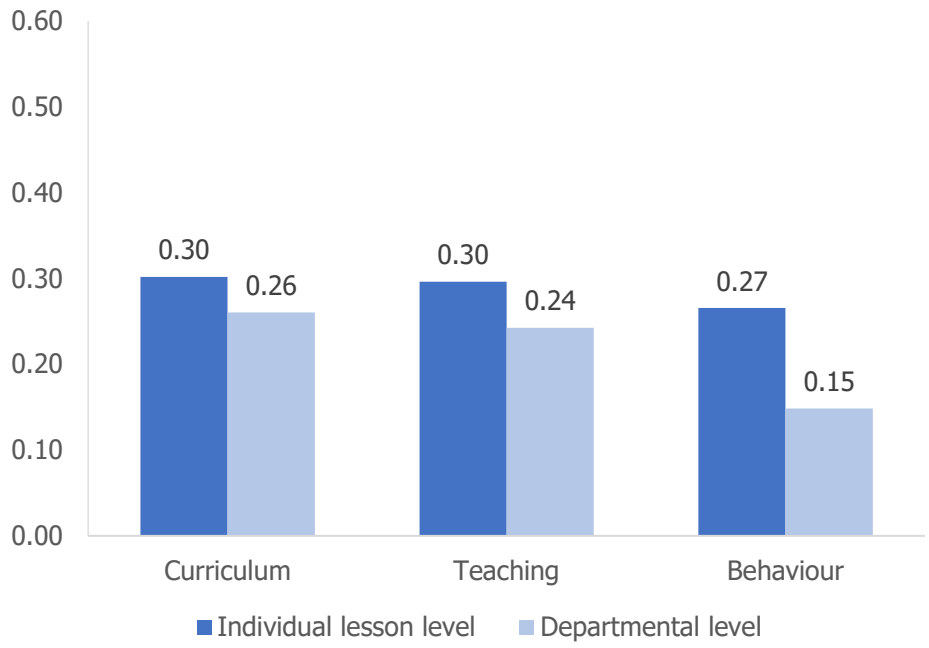

School data merges observations from all the school phases in the study. This includes 197 observations in total.
College data is based on 119 observations.

Department-level reliability provides similar findings to that identified at the individual lesson level. Inter-rater reliability was much lower in each domain, albeit the behaviour domain declines to a slight agreement being identified between observers (Figure 14). Unsurprisingly, the majority of the 16 indicators also decline to a slight level of agreement at the department level. For two of the behaviour judgements, inter-rater reliability was worse than had the ratings been produced by chance.

The low level of reliability in the behaviour domain is particularly interesting. Our initial assumptions were that this would be the strongest of the domains across each of the phases investigated. This was on the basis that across other observation models, behaviour tends to be the easiest and most reliable component of a lesson to assess. That it has not been the case here says something about the nature of behaviour expectations between school and college providers in England – they are perhaps fundamentally different. That is, expectations of behaviour vary considerably

across the FES sector on account of different definitions, for example in adult technical classes, apprentices on-the-job and 16 year olds taking level 1 or 2 qualifications compared with less variance across schools.

**Figure 14: Comparison of individual lesson and department level inter-rater reliability domain scores in the college sample**




Theoretically, the relative stability in classroom conditions and expectations at primary and secondary schools perhaps allow for greater consensus across subjects on what behaviour might look like. Additionally, the lower age of pupils means the teacher takes a larger role in determining how they should act.

Colleges' wider range of contexts, ages, experiences and health and safety implications makes for a much more complex environment in which to establish general guidance on behaviour. For example, being ready to learn in a practical lesson on perming hair in a classroom that seeks to resemble a professional hair salon is necessarily different from a classroom teaching A-level French. Different views on what constitutes acceptable behaviour in this context (including among observers) makes it less straightforward to assess using the indicators designed for our model.

It is also likely that the wide range of very specialised subjects that observers looked at in the college sample, compared with a narrower range of subjects often at a lower level of specialisation in the school sample, may also compensate for the low level of reliability encountered. For instance, as more variant factors come in to play, the lower the level of inter-rater reliability achieved. This is certainly something that the greater reliability between observers in the primary schools' hints at, with increased specialisation in subjects at secondary schools then leading to slightly less reliability among observers. By the time you get to the curriculum in college

provision, which is heavily differentiated in terms of pedagogy and cultural behaviours, some of the indicators become an issue.

Overall, the findings from the college observations suggest that our prototype model is not a good fit for lessons in a FES context, as it is likely to be looking at the wrong things. This requires more research.

## What other factors help to explain the levels of inter-rater agreement identified?

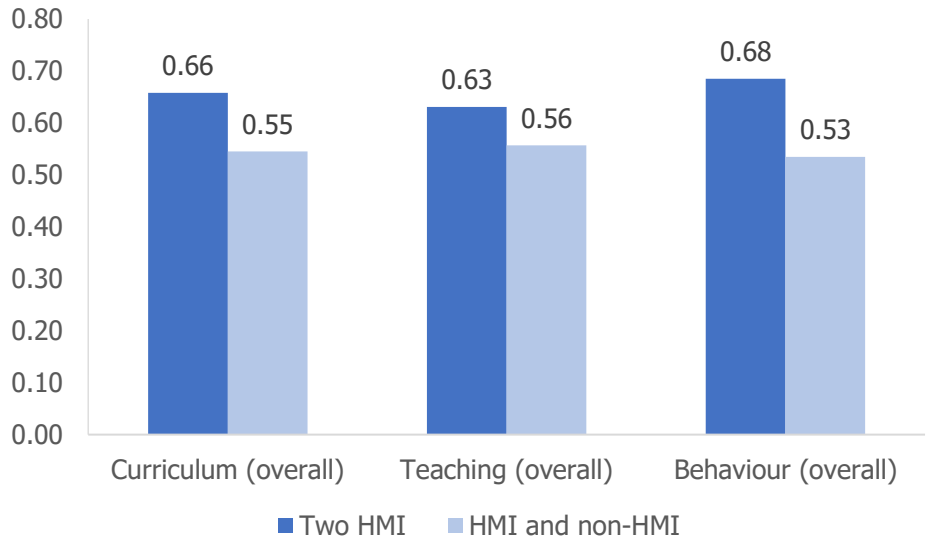
The method of data collection along with the views of participating HMI in the fieldwork allowed us to look at other variables that may be contributing to the levels of inter-rater reliability achieved. The following section looks at these factors in greater detail. Owing to the issues with reliability identified in the college sample, the analyses in this section are based on the school sample only.

### Experienced observers

Figure 15 shows that observation carried out by inspectors generally leads to greater reliability. Two school inspectors paired together for observation scored the indicators more consistently than when an inspector was paired with a non-inspector. This suggests that non-inspectors and their lack of experience in observing may have had a slight negative effect on the overall reliability scores.

Inspector pairings achieved some of the highest kappa scores across the study. This was also the case for the 16 indicators informing the overall domain scores. Of these, nine achieved a substantial level of reliability (>0.6) and a further five achieved a score of 0.58 or 0.59. By comparison, none of the indicators achieved a substantial level of reliability in the HMI and non-HMI pairings. Overall, the pattern of the data here is particularly encouraging regarding inspector reliability. It suggests that even without standardised training, experienced school inspectors are generally in-tune with each other when it comes to observing specified components of a lesson consistently.

**Figure 15: Inter-rater agreement between inspector pairings and inspectors observing with non-inspectors**

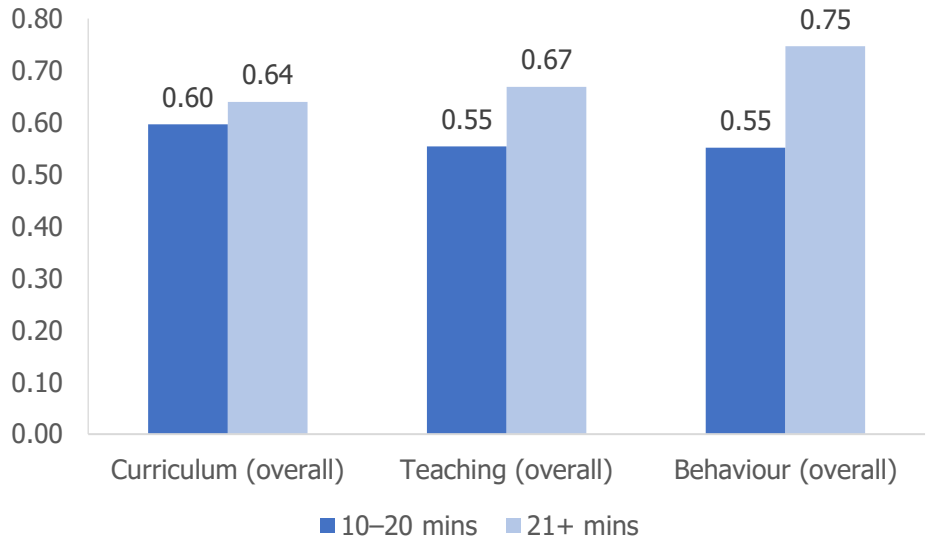



Based on observations from primary schools, secondary schools, the special school and PRU. Eighty-three observations were carried out by a pairing of two HMI, with 129 carried out by a HMI and non-HMI pairing.

## Lesson length

Reliability appears to increase the longer the lesson is observed. Figure 16 shows this is particularly the case for the overall teaching and behaviour domains. Furthermore, 12 of the 16 indicators also achieved a substantial level of reliability when the observation lasted longer than 20 minutes, compared with three indicators when the observation lasted between 10 and 20 minutes in length.

**Figure 16: Inter-rater agreement by length of observation, across the school sample**

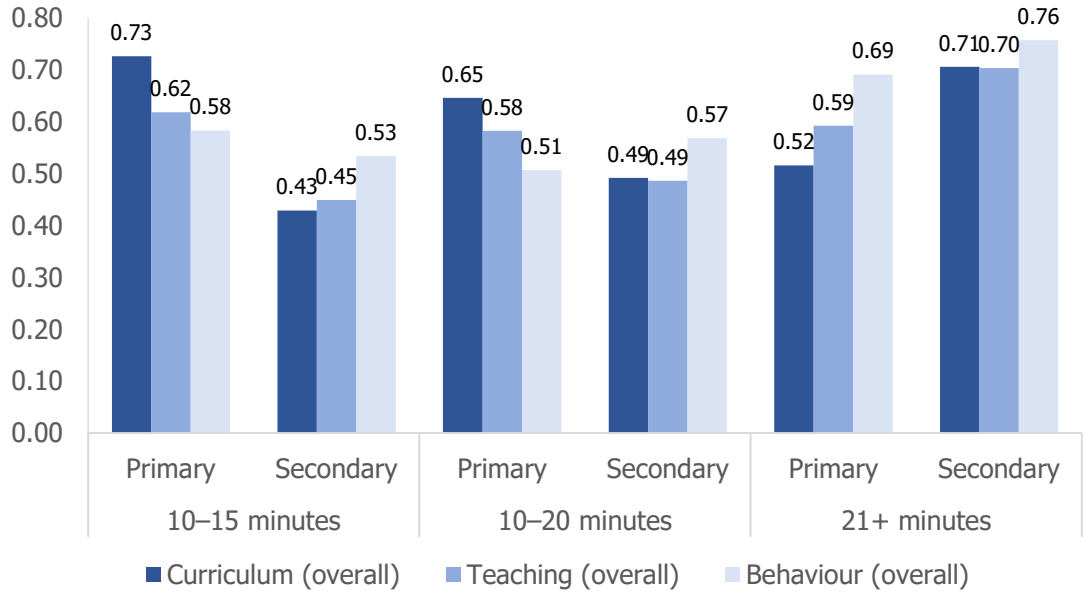



Based on observations from primary schools, secondary schools, the special school and PRU.

However, variation between phases exists. Figure 17 shows a mixed picture, although every kappa score provides at least a moderate level of inter-rater reliability. The data indicates that observation in secondary schools was generally more reliable the longer the length of the observation. In the primary schools, curriculum was more reliable to score in shorter observations. Across both phases, it appears that the longer the lesson, the more reliable the assessment on behaviour between both observers.

In the focus groups, HMI told us that the flexibility of the model was helpful. This is because sometimes 15 minutes was not quite enough to determine the score for an indicator. For example, one HMI talked about the usefulness of being able to extend observation length to allow for more time to observe learners participating in activities. A fixed time protocol would have prevented the inspector from seeing an activity that went on to contribute to how the indicator scores were applied. Similarly, some inspectors highlighted that sometimes 15 minutes was more than enough time to observe a lesson, particularly when the lesson was not going so well. They felt it was appropriate to move on to the next observation quickly when this was the case, so as not to place unnecessary burden on staff and pupils.

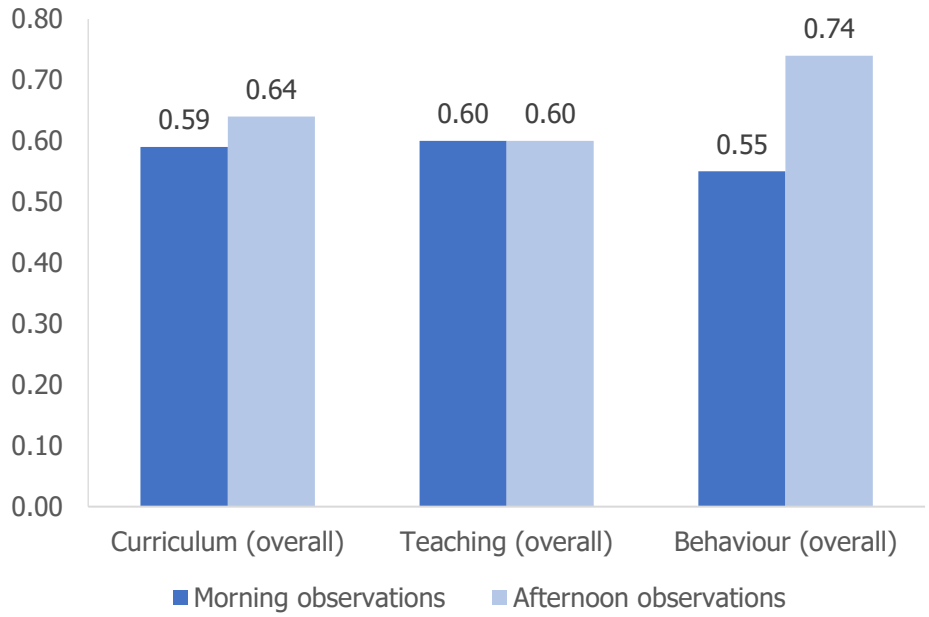
**Figure 17: Inter-rater agreement by length of observation, by phase**




## Time of day

The focus on two departments across a provider visit allowed us to analyse whether the time of day had any influence on reliability. We structured the visits so that observers carried out lesson observations in the morning for one department and the afternoon for the second department. We hypothesised that as the day wore on and tiredness from observing set-in, reliability would decrease. In fact, the data suggests the inverse happened, as shown in Figure 18.

**Figure 18: Inter-rater agreement by observation carried out in the morning or afternoon sessions**




Based on observations from primary and secondary schools only.

The behaviour domain shows a particularly strong level of reliability in the afternoon compared with those lessons observed in the morning. Furthermore, of the 16 indicators in the model, eight achieved a substantial level of agreement in the afternoon sessions, whereas only one indicator was found to have substantial agreement from the morning observations. The distribution of the kappa statistics for these 16 indicator scores from the morning observations ranged from 0.45 to 0.62. For the afternoon observations, the distribution was much higher and ranged from 0.55 to 0.7.

This suggests a practice effect may be influencing the level of consistency instead (although there may be other explanations). That is, as inspectors got used to a new method of observing, their application of the model became more consistent. This is supported by the focus group evidence. Several of the inspectors commented that the indicators and rubric became easier to use as they became more familiar with the instrument, but that, initially, it had proved to be a challenge. This was partly through the number of indicators and the amount of information within the rubric they were expected to know and apply during the observations. Additionally, some inspectors had a week or two between fieldwork visits, meaning that they needed to re-learn aspects of the instrument before applying it effectively.

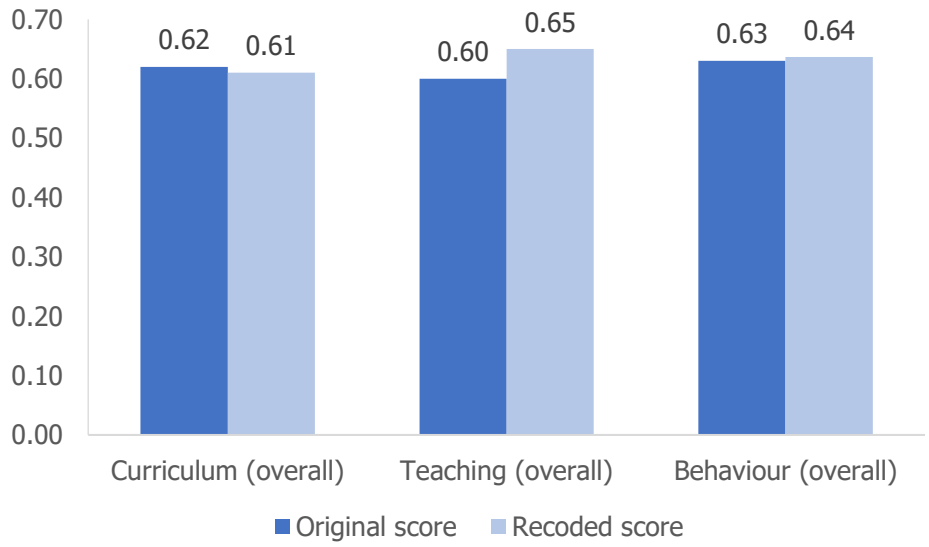
## Central tendency

The HMI in the focus groups felt that the five-point scale used for scoring the indicators was not without problems. They often used the score of 3 in the rubric when they were uncertain. They found the criteria at either end of the scale easier to

identify and score, but the five-point scale provided them with the means to 'sit on the fence'.

The inference here was that central tendency could be causing some degree of inconsistent practice in the way observers scored. To test out this assumption, we recoded the data from the five-point scale into a three point-scale[61] and calculated inter-rater reliability using this amended data-set. Figure 19 compares the original kappa statistics for the school's data with that of the recoded data-set. The data shows that the assumption for central tendency is supported, particularly at the individual lesson level. The difference in the kappa statistics remains small, despite the level of agreement increasing slightly. If central tendency was not a factor, we would expect larger increases in the recoded data. However, inspector ratings usually differed across scores of 2, 3 and 4 rather than the boundaries at either end of the five-point scale.

**Figure 19: Inter-rater reliability of two observers at the individual lesson level using the original five-point scale and recoded three-point scale**




Based on observations from primary schools, secondary schools, the special school and PRU.

However, a move to a three-point scale would be problematic. For instance, it is likely to undermine validity as it will reduce the amount of variation that inspectors can attribute from their observations. So, while reliability might increase, validity might be reduced to a point where meaningful measurement is compromised.

## Additional training

The level of inter-rater reliability reached in the school lesson observations provides a useful baseline for taking the prototype model forward. In some areas, a substantial level of agreement has been identified based on the strength of the indicator,

[61] Scores of 1 and 2 in the rubric were combined, as were scores of 4 and 5.

strength of the rubric design and observer knowledge alone. We expect, however, that further rigorous training would enhance reliability.

As the international seminar on lesson observation identified, many existing frameworks strengthen reliability through extensive programmes of training – including accredited training. Inspectors from the focus groups felt that they would have benefited from further training to ensure consistency in interpretation of the rubric and around the processes used to arrive at overall scores. This was particularly the case in relation to department level synthesis of multiple observations. One common view expressed was that:

> 'My concern was whether we have a consistent understanding of the rubric. We needed a bit more time to make sure we all understood and applied it correctly.'

A single training day introducing inspectors to the rubric and indicators was not nearly long enough to generate a common understanding. The time-of-day analysis also bears this view out. In addition, several inspectors felt that the absence of standardised training, particularly a lack of examples to interpret practice against the rubric, hindered their application of the observation model.

The focus group involving FES inspectors suggested that future training requirements may need to be phase-specific. They particularly highlighted complexities in scoring aspects of curriculum and behaviour. Expectations around behaviour in colleges can depend on health and safety requirements, the age of students and the number of students in classes, all of which tend to be less standardised than in schools. A focus on subject specialism in FES providers, therefore, may mean there is less consensus on aspects like behaviour and curriculum, which would need to be the focus of future training.

### Reducing cognitive load

Part of the rationale for including 18 indicators in the model was so that we could determine the most useful indicators. This was particularly the case for the curriculum indicators, which is a new construct. However, we expected the number of indicators within the model to have a negative impact on reliability, as pointed out in the existing literature.[62]

This was corroborated through HMI feedback in the focus groups. They were clear that the structuring of observations by the three domains had a positive impact. As one HMI noted, this allowed observers to be more focused and structured in what they were looking for:

[62] H C Hill, C Y Charalambous and M A Kraft, 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', in 'Educational Researcher', 41(2), 2012, pp. 56–64; K Mihaly, D F McCaffrey, D O Staiger and J R Lockwood, 'A composite estimator of effective teaching', Bill and Melinda Gates Foundation, 2013.

'[it] empowered me to look at lessons more effectively and to know what I am supposed to be doing there.'

However, they also felt that the cognitive load of the model was substantial. Several identified that, despite the positives of the structure, they needed a significant amount of time at the start of the study to assimilate the rubric. Some inspectors also reported overlap between a few of the indicators that made it more complicated to directly relate observations to the rubric. Indicators 1c and 1h were generally identified as less useful and were recommended as indicators to remove from the model.

The findings from the study suggest that the domains and nearly all the indicators selected are valid aspects of the model design. As such, we can now refine these to a core selection of indicators, which we would expect to reduce cognitive load on observers. We also assume that this would improve departmental level reliability.

## Which are the most useful indicators?

As the validity analysis for our model suggests that different combinations of the indicators may yield similar results in practice, we have instead used the inter-rater reliability data to specify an optimum model for further testing. This considered differences in the levels of agreement reached across phase and between the lesson and department level to identify the indicators where observers performed most consistently. The kappa statistics from the college analysis, however, have not been included in this process.

The validity analysis had already identified that indicators 1c and 1h were typically aspects of curriculum that inspectors did not see enough evidence during an observation to provide a score. This suggests that other methods are perhaps better suited for identifying quality on these indicators. Additionally, three other indicators each featured inter-rater reliability often below 0.5 across several of the analyses carried out. This suggests that they are likely to be aspects of a lesson that are harder for observers to assess with a secure level of consistency. On this basis, we have removed the following three indicators from the model:

- Teachers ensure there is equality of opportunity for all learners to access every lesson (indicator 1b).
- Teachers effectively check for understanding (indicator 2f).
- Teachers create focused classrooms through high expectations for pupils (indicator 3b).

Based on the inter-rater reliability of the remaining 13 indicators, we identified the following eight as commonly featuring substantial or a high degree of moderate reliability across the various analyses carried out. These will be the indicators that we will prioritise in developing further on pilot visits for the EIF:

Curriculum domain:

- Teachers use their subject expertise to provide effective learning opportunities (indicator 1a).
- The lesson content is appropriate to the age group and does not lower expectations (indicator 1e).
- There is a logical sequence to the lesson (indicator 1f).

Teaching domain:

- Teachers demonstrate good communication skills (indicator 2a).
- Teachers possess good questioning skills (indicator 2d).
- Teachers give explicit, detailed and constructive feedback in class (indicator 2e).

Behaviour domain:

- Teachers create supportive classrooms focused on learning (indicator 3a).
- Pupils' behaviour contributes to the focus on learning (indicator 3d).

Furthermore, a few of the inspectors noted during the focus groups that the teaching and behaviour indicators might be indicators that are already within their comfort zone. The uniqueness of the curriculum indicators, however, means that further investigation is perhaps warranted before decisions are made on their validity. We will take this into consideration as we refine the model for the EIF.

# Annex A: Inter-rater reliability data tables

Data tables showing the full inter-rater reliability Kappa scores for each indicator in the research model are available on the Ofsted website.

# Annex B: List of the 37 schools and colleges that participated in the research visits

| School name | Local authority | Type of provider | Ofsted phase |
|---|---|---|---|
| Arthur Mellows Village College | Peterborough | Academy Converter | Secondary |
| Bath College | Bath and North East Somerset | GFE College | College |
| Beehive Lane Community Primary School | Essex | Community School | Primary |
| Boroughbridge High School | North Yorkshire | Community School | Secondary |
| Boxford Church of England Voluntary Controlled Primary School | Suffolk | Voluntary Controlled School | Primary |
| Boxted St Peter's Church of England School | Essex | Voluntary Controlled School | Primary |
| Brockenhurst College | Hampshire | GFE College | College |
| Burston Community Primary School | Norfolk | Academy Converter | Primary |
| Campion School | Warwickshire | Academy Converter | Secondary |
| Chirbury CofE VC Primary School | Shropshire | Voluntary Controlled School | Primary |
| Coundon Court | Coventry | Academy Converter | Secondary |
| Exeter College | Devon | GFE College | College |
| Fareham College | Hampshire | GFE College | College |
| Good Shepherd Catholic School | Coventry | Academy Converter | Primary |
| Ledbury Primary School | Herefordshire | Community School | Primary |
| Melbourn Village College | Cambridgeshire | Academy Converter | Secondary |
| Much Marcle CofE Primary School | Herefordshire | Voluntary Aided School | Primary |
| North East Essex Co-operative Academy | Essex | Academy Alternative Provision Converter | PRU |
| North East Surrey College of Technology | Surrey | GFE College | College |
| Richard Huish College | Somerset | Sixth Form College | College |
| Royal National College for the Blind | Herefordshire | Independent Specialist College | College |
| Sandwell College | Sandwell | GFE College | College |
| Shorefields School | Essex | Community Special School | Special |
| Sidegate Primary School | Suffolk | Academy Sponsor-led | Primary |
| South Gloucestershire and Stroud College | South Gloucestershire | GFE College | College |
| Spalding High School | Lincolnshire | Community School | Secondary |
| St Brendan's Sixth Form College | Bristol | Sixth Form College | College |
| Stanmore College | Harrow | GFE College | College |
| The De Montfort School | Worcestershire | Community School | Secondary |
| The Westwood Academy | Coventry | Academy Converter | Secondary |
| Thirsk School & Sixth Form College | North Yorkshire | Community School | Secondary |
| Weston College | North Somerset | GFE College | College |
| Weymouth College | Dorset | GFE College | College |

| | | | |
|---|---|---|---|
| Whitchurch CofE Primary School | Herefordshire | Voluntary Aided School | Primary |
| Worcester Sixth Form College | Worcestershire | Sixth Form College | College |
| Wymondham High Academy | Norfolk | Academy Converter | Secondary |
| Yeovil College | Somerset | GFE College | College |

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. 190029