Six models of lesson observation: an international perspective

In November 2017, Ofsted hosted an international seminar on lesson observation. This paper reports on the observation models presented at the seminar and discusses how they may help Ofsted with future inspection framework development.

Published: May 2018

Reference no: 180022

Contents

Foreword from Her Majesty's Chief Inspector

At the end of last year, I was delighted to host Ofsted's first international research seminar, the focus of which was lesson observation. Observation is an important part of an inspector's toolkit, particularly for making judgements on the quality of teaching. As part of our strategy, we are committed to constantly improving the validity of inspection so that our judgements of schools are the best reflection of the quality of education they can be. Scrutinising the reliability and validity of inspection methods is an important part of improving validity overall. The seminar was set up with this purpose in mind.

We invited 14 experts from around the world to share their knowledge on lesson observation. Over the course of the two days at the seminar, there were many enlightening and sometimes challenging discussions. In no area was there more debate than on the key question of: 'what changes should Ofsted consider in developing lesson observation for its 2019 inspection framework?'

Those attending the seminar recognised the validity of Ofsted's current approach to lesson observation, which is of great encouragement to me. Of course, inspectors never rely on lesson observation alone. They go through a process of triangulation, where different evidence sources are weighed against each other to reach a judgement. The weight that inspectors place on considering a range of evidence points is particularly important for validity.

There was also a consensus that the purpose that observation serves in inspection is very different to the purpose of observation for most of the models presented. This was particularly true for those models that focused on individual teacher observation/accountability. We no longer grade individual lessons, nor do we judge the quality of teaching of individual teachers, because one-off observations of a single teacher are likely to be unreliable for evaluating that teacher. The experts agreed that, in our inspection context, where observation is used to inform judgements of school quality, it would be a mistake to pick up an off-the-shelf model from elsewhere and apply it wholesale.

This does not mean that the lesson observation models presented did not have anything new to offer. The more systematic approaches seen across the models provide a number of areas for Ofsted to investigate as we develop our new framework and refine how we evaluate quality of teaching. In particular, we are interested in the fact that all the models put a large amount of structure around expert judgements. There were also stimulating insights about the subject-specific dimension of lesson observation, which we intend to explore further.

Overall, the seminar has certainly given us plenty to think about. In the coming months, research and policy colleagues at Ofsted will be looking at how best to apply this new knowledge. This will include testing new models to see whether they improve inspection practice and give a richer measure of the quality of education, in order to benefit parents, pupils and schools themselves.

I would also like to take this opportunity to say thank you to the international experts who attended the seminar. We are very grateful for the long journeys some undertook to contribute to an excellent couple of days of knowledge sharing. Additionally, I would also like to thank Professor Robert Coe from Durham University who co-produced this event with research colleagues at Ofsted.

I am looking forward to keeping you informed about how our future models of lesson observation develop and am confident that this work will build on existing progress to improve the validity and reliability of our inspections further.

Introduction

In March 2017, the publication of a small-scale reliability study signalled a change in emphasis in Ofsted's use of research.[1] Her Majesty's Chief Inspector (HMCI) announced that this was the first step towards a continuing programme of research into inspection, particularly around its reliability and validity.[2]

The Ofsted strategy provides further focus on this intention. Ofsted is committed to improving the validity of its inspection practice. In order to act as a force for improvement, our inspection work needs to be evidence-led and the evaluation tools and frameworks we use should be as precise, valid and reliable as possible.[3] This is particularly the case as we work towards developing the 2019 education inspection framework.

The first part of inspection that we decided to critically evaluate for the new framework is inspectors' use of lesson observation. Ofsted has used lesson observation as part of the inspection process since its foundation in 1992. This method is still routinely used in school inspections, as well as in inspections of other educational settings such as further education colleges. At one time, it was expected that 60% of inspection time in schools should be used to observe lessons.[4]

The evidence collected from lesson observation remains an important element of the 'teaching, learning and assessment' judgement, as well as for making judgements about the effectiveness of leadership and management. It is therefore a fundamental part of inspection that deserves focused attention.

Furthermore, because school inspection is carried out in a sector where change can be frequent, inspection needs to keep pace with current educational policy and developments. Research and practice in the use of lesson observation have also seen

[1] 'Do two inspectors inspecting the same school make consistent decisions? A study of the reliability of Ofsted's new short inspections', Ofsted, 2017; www.gov.uk/government/publications/do-two-inspectors-inspecting-the-same-school-make-consistent-decisions.
[2] 'HMCI's commentary: new research into short school inspections', Ofsted, 2017; www.gov.uk/government/speeches/hmcis-monthly-commentary-march-2017.
[3] 'Ofsted strategy: 2017 to 2022', Ofsted, 2017; www.gov.uk/government/publications/ofsted-strategy-2017-to-2022.
[4] 'Inspecting schools: handbook for inspecting secondary schools', Ref: HMI 1360, Ofsted, 2003.

some very significant developments in recent years, particularly in international practice. As part of our commitment to using tools that are valid and reliable, it is right that we continually seek to improve our inspection practice in areas where new knowledge is emerging.

Recognising these developments, we held an international seminar on lesson observation in Westminster on 6 and 7 November 2017, in conjunction with Professor Robert Coe from Durham University. This seminar brought together academics and experts from around the world working in the evaluation of teaching. Its purpose was to help Ofsted explore future framework design, specifically looking at the use of lesson observation in schools. The overarching questions addressed at the seminar were:

- What can Ofsted learn from international best practice in the use of lesson observation for the evaluation and improvement of teaching quality?
- What changes should Ofsted consider in developing its inspection framework for 2019?

This paper describes the six observation models presented by attendees over the two days of the seminar, particularly the similarities and differences between them. Three of the models are commonly used in the US, two are linked with European inspection systems, and the other has been developed for use across borders. All of them have a research base that supports claims to their validity and reliability. The models are:

- Classroom Assessment Scoring System (CLASS)
- Framework for Teaching (FfT)
- International Comparative Analysis of Learning and Teaching (ICALT)
- International System for Teacher Observation and Feedback (ISTOF)
- Mathematical Quality of Instruction (MQI)
- Generic Dimensions of Teacher Quality.

The paper also considers the views and reflections provided by the attendees at the seminar. It does not, however, conclude with Ofsted's future model for observing lessons. This needs further conversation, testing, consultation and decision-making. Instead, it sets out some of our initial thinking on how aspects of the six lesson observation tools presented at the seminar might improve our current observation method. We will use this knowledge to identify the implications for Ofsted when we develop valid observation protocols for the 2019 education inspection framework.

We would like to thank the following for attending the seminar and are grateful for their thoughtful contributions:

- Courtney Bell (Educational Testing Service, Princeton)
- Robert Coe (Durham University)

- Charlotte Danielson (Framework For Teaching)
- Marjoleine Dobbelaer (University of Twente)
- Bridget Hamre (University of Virginia)
- Heather Hill (Harvard Graduate School of Education)
- Kirsti Klette (University of Oslo)
- Eckhard Klieme (DIPF, German Institute for International Educational Research)
- James Ko (Education University of Hong Kong)
- Marcus Pietsch (Lüneburg/Hamburg Inspectorate)
- Pamela Sammons (University of Oxford)
- Sandy Taut (State Institute for School Quality and Educational Research, Munich)
- Wim van de Grift (University of Groningen)
- Adrie Visscher (University of Twente).

Main findings

- The six models presented at the seminar had been designed in a systematic way. All had devised observation criteria that are structured and categorised around the judgement of experts and linked to a core purpose of assessing the quality of teaching. This provides a clear focus for those observing lessons.
- The models all produce predominantly quantitative data. This allows them to provide more detailed feedback to teachers that can go down to the level of individual items and indicators of teacher quality.
- The structure of the six models were typically informed by the research literature on the quality of teaching. This explains some of the overlap noted between the models, particularly around their focus on classroom management, instruction and student behaviour and attitudes. This suggests a degree of concurrent validity between the models.
- Most of the models had also evolved over years of iterative design and implementation. The data collected from years of previous lesson observation studies, therefore, provides further reassurance of each model's credibility.
- Interestingly, none of the models explicitly attempted to measure learning. It was generally agreed among the experts that learning is not something that can be directly observed, while the quality of teaching can.
- Most models are similar in terms of the relationship between observation ratings and pupil attainment measures, which have typically modest correlations. This suggests that it is important to look at observation as just one of a set of measures rather than as the sole measure used to judge against an agreed purpose.

- Despite the emphasis on quantitative measures, most of the models were still considered to be high inference.[5] The rating scales used tended to be informed by subjective judgements that were based largely on observed classroom behaviours. Therefore, the experts at the seminar indicated that observers generally needed a high standard of training in using an observation instrument.
- The experts generally agreed that, in addition to standard training, regular refresher training for observers was also important to maintaining a high level of consistency. The experts all accepted that their models could never be 100% reliable.
- There was some debate over the use of video recording instead of live observations. On the one hand, it was argued that video could improve accuracy because a recording could be watched multiple times. On the other, it was argued that video could lose vital contextual information about the classroom when filmed from a single or even multiple perspective(s).
- No agreement was reached on the ideal observation length of time or number of observations required. It was largely agreed that this is dependent on the context and the intended purpose of the observation.
- Attendees agreed that, although most of the models discussed were capturing generic aspects of the quality of teaching, subject-specific factors were equally important. However, few subject-specific models currently exist beyond mathematics and language and literacy. This can be seen as particularly problematic when curriculum is focused on sharply.
- It was agreed that the educational system around the observation model was as important as the model itself. This suggests that the cultural specificity around the models means that they are unlikely to work exactly as intended in different contexts.
- A key difference between the models presented at the seminar and Ofsted's purpose lies in the focus on the individual teacher. Teacher-level accountability is the main purpose some US models are used for. Most of the other models also focus on collecting teacher-level data. This is in contrast to Ofsted where, owing to concerns around the reliability of single observations, we have moved away from grading lessons of individual teachers to focus observation at the school level.

Ofsted's lesson observation model

Lesson observation has been an important feature of the inspection process since Ofsted was founded in 1992. However, inspection has not stood still over the past 25

[5] A high inference observation model is one that requires the observer to make subjective inferences beyond the behaviours observed. Conversley, a low inference model captures observable facts or events, with minimal interpretation or subjectivity.

years. Framework change has had considerable consequences for the place of lesson observation within Ofsted's inspection practice.

Pre-2005 frameworks

Prior to the September 2005 framework, lesson observation was generally expected to make up typically 60% of the time on inspection. This was on the basis that all inspection findings must be rooted in evidence and that 'the most valuable and informative evidence is that obtained first hand, from observations, the analysis of the processes of the school and examining pupils and their work'.[6] Owing to the scale of school inspection during this period – an inspection team could consist of between 10 and 15 inspectors inspecting a school for a week – it was feasible to carry out a relatively high number of subject-specific observations. This provided detailed subject evidence in the published inspection report for each school.

The instrument for collecting evidence from lesson observation was a standardised evidence form that was used to collect all types of inspection evidence, not just from lesson observation. Inspectors used it to collect evidence on teacher behaviours and student interactions in a high-inference qualitative format. It was used in combination with the evaluation criteria from the inspection handbook to help inspectors judge the quality of teaching. The 1995 evaluation criteria for the overall quality of teaching, not in individual lessons, can be found below. There is little difference in the published evaluation criteria from the 1999 and 2003 handbooks:

1995 framework
Secure knowledge and understanding of the subjects or areas taught
Setting of high expectations to challenge and deepen pupils' knowledge and understanding
Plan effectively
Employing teaching methods and organisational strategies which match curricular objectives and the needs of all pupils
Manage pupils well and achieve high standards of discipline
Use of time and resources
Assessing pupils work thoroughly and constructively to inform teaching (in the case of lesson observation through listening and responding to pupils)
Use homework effectively to reinforce and/or extend what is learned in school

[6] 'Inspecting schools: handbook for inspecting secondary schools', Ref: HMI 1360, Ofsted, 2003, p.12-13.

Research on the use of the standardised evidence form for lesson observation at the turn of the millennium identified that it was being used in a valid and reliable way by inspectors. For instance, an unpublished study conducted by Ofsted and the Dutch inspectorate identified that the reliability of Ofsted's standardised evidence form was equal to that of the detailed checklist (the initial instrument that evolved into the ICALT model) used by the Dutch inspectorate at that time.[7] Additionally, other research also showed that registered inspectors carrying out a paired observation of the same lesson generally came to similar outcomes about the quality of the lesson when using this type of instrument.[8]

Post-2005 framework

In comparison, the more proportionate inspection process introduced by the September 2005 framework led to a reduction in the time available for lesson observation during inspection.[9] Observation has remained the central method through which teaching and learning are assessed, but there was clearly greater coverage possible when large inspection teams inspected a school for a week. There is no longer any expectation for a certain number of observations to be completed by the inspection team. In today's shorter inspections, lesson observation is, instead, generally used to provide evidence for the inspection team's main lines of enquiry. For example, if girls' performance in mathematics was a key line of enquiry for the inspection, lesson observation would focus on this element.

In addition, as another result of less inspection resource being available, lesson observation has become less focused on subject-specific content. The focus has shifted towards generic attributes of teaching and learning across subjects. Furthermore, since 2015, inspectors no longer grade the quality of teaching and learning in individual lessons nor do they judge the quality of teaching of individual teachers. This change was made in response to the difficulties identified in making reliable judgements on quality through a single brief lesson observation of an individual teacher. Instead, inspectors now observe many lessons across the school to provide a reliable aggregate picture of teaching quality. This has changed the scope of observation, moving it away from individual practitioners and towards the school as a whole.

The importance of observation in helping inspectors to form a judgement on the quality of teaching and learning remains largely intact though. Inspectors are still expected to use a considerable amount of first-hand evidence to determine the

---

[7] W van de Grift, P Matthews, B Corporaal & M Collier, 'Do English and Dutch inspectors judge lessons in the same way?' Internal paper (not published), 2002.

[8] P Matthews, JR Holmes, P Vickers and B Corporaal, 'Aspects of the reliability and validity of school inspection judgements of teaching quality', Educational Research and Evaluation, 4:2, 1998. https://www.tandfonline.com/doi/abs/10.1076/edre.4.2.167.6959.

[9] The inspection frameworks and handbooks from 2005 to 2014 can be found on the National Archives: http://webarchive.nationalarchives.gov.uk/20141107100046/http://www.ofsted.gov.uk/resources/maintained-schools-inspection-documents-archive.

quality of teaching and learning. This includes observing pupils in lessons, talking to them about their work, scrutinising their work and assessing how well leaders are securing continual improvements in teaching. Inspectors are expected to triangulate direct observation from lessons with a range of other evidence so that they can evaluate the impact that teaching is having on pupils' progress.[10]

There has also been a greater focus on learning over time, for example through work scrutiny and discussions with pupils about what they do and do not remember about what they have been taught. At the same time, inspectors stand at the back of a classroom to observe a lot less than they used to. They engage with and ask pupils questions during the lesson to enhance the evidence around what has been learnt.

Inspectors' use of the standardised form for collecting evidence from lesson observations has been a consistent element of practice since 1995. This has changed very little since then. The widespread perception that inspectors apply a tick-box approach is therefore contradicted by the large amount of qualitative information inspectors record from lesson observation.

Ofsted has also retained a degree of consistency in the evaluation criteria for the quality of teaching in the latest inspection handbook. While changes have featured across handbooks since September 2005, there are some commonalities in focus between the pre-2005 evaluation criteria and that identified in the 2015 framework.[11]

2015 framework
The teachers' standards are being met[12] ■ Set high expectations which inspire, motivate and challenge pupils ■ Promote good progress and outcomes by pupils ■ Demonstrate good subject and curriculum knowledge ■ Plan and teach well structured lessons ■ Adapt teaching to respond to the strengths and needs of all pupils ■ Make accurate and productive use of assessment ■ Manage behaviour effectively to ensure a good and safe learning environment Teachers and other staff have consistently high expectations of what each pupil can achieve, including disadvantaged pupils and the most able

[10] Triangulation is the process we refer to where inspectors weigh different evidence sources against each other in coming to a judgement.
[11] 'School inspection handbook', Ofsted, 2018 www.gov.uk/government/publications/school-inspection-handbook-from-september-2015
[12] 'The teachers' standards', Department for Education, 2011; www.gov.uk/government/publications/teachers-standards.

Teachers and other staff have a secure understanding of the age group they are working with and have relevant subject knowledge that is detailed and communicated well to pupils

Assessment information is gathered from looking at what pupils already know, understand and can do

Assessment information is used to plan appropriate teaching and learning strategies, including to identify pupils who are falling behind in their learning or who need additional support, enabling pupils to make good progress and achieve well

Pupils understand how to improve as a result of useful feedback, written or oral, from teachers

Equality of opportunity and recognition of diversity are promoted through teaching and learning

English, mathematics and the skills necessary to function as an economically active member of British society are promoted through teaching and learning.

Six international models of lesson observation

The six models were presented at the seminar in the following order by the named individuals:

- The Classroom Assessment Scoring System (CLASS) – Bridget Hamre
- Framework for Teaching (FfT) – Charlotte Danielson
- The International Comparative Analysis of Learning and Teaching (ICALT) – Wim van de Grift
- The International System for Teacher Observation and Feedback (ISTOF) – Daniel Muijs
- The Mathematical Quality of Instruction (MQI) – Heather Hill
- Generic Dimensions of Teaching Quality – Eckhard Klieme.

CLASS

CLASS was developed by Robert Pianta, Karen LaParo and Bridget Hamre.[13] It was one of the protocols included in the 'Measures of Effective Teaching' (MET) study, where it was identified as a reliable observation instrument that was positively, albeit modestly, associated with student achievement gains.[14] Originally designed for early years contexts and later extended to the full age range for schooling, CLASS is an observational tool that provides a common lens and language focused on assessing the effectiveness of classroom interactions between teachers and students.

Initially developed for research, CLASS has been scaled for use in practice over the last decade. Research from over 2,000 classrooms using this model provides useful evidence about the nature of teacher–child interactions and the ways in which these interactions promote children's social and academic development. Four overarching conclusions have emerged from the research:

- effective teacher–child interactions are an active and crucial ingredient for children's social and academic development
- children are not consistently exposed to effective teacher–child interactions
- to maximise the impact for children, quality improvements need to focus explicitly on teacher–child interactions

[13] RC Pianta, KM LaParo, & BK Hamre, (2008) Classroom Assessment Scoring System Manual: Pre-K. Baltimore: Brookes.

[14] TJ Kane & DO Staiger, 'Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains', MET Project Research Paper, Bill & Melinda Gates Foundation, 2012. http://k12education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-2/

- carefully designed and implemented professional development support can improve the quality of teacher–child interactions.

In schools, a main component of the CLASS model is that it links teacher behaviours with student achievement as part of an observational teacher-assessment tool. It is also aligned with a set of professional development materials that show links to improvement. In this way, the model is about enhancing the overall relationship between teachers and students and their learning. It is not just about monitoring and evaluating teachers' performance.

CLASS measures three broad domains of interactions among teachers and children: emotional support, classroom organisation and instructional support. Each domain includes several dimensions that relate to what is directly observed and indicators for each of the dimensions that act as behavioural markers. For instance, the dimensions of classroom organisation are:

- behaviour management
- productivity
- instructional learning formats.

Indicators for behaviour management include the teacher:

- 'having clear rules and expectations that are consistently reinforced'
- 'being proactive in anticipating difficulties'
- 'reinforcing positive behaviors and redirecting unwanted behaviors'.

Collectively, the 11 dimensions in the model assess the extent to which teachers are effectively supporting children's development, both social and academic.[15]

The tool includes four cycles of approximately 15-minute observations of teachers and students by a certified CLASS observer. These observations are then rated using a manual of behaviours and responses. However, CLASS is not a simple checklist, but a high-inference model. Observers take extensive behavioural notes throughout and are trained to ensure that they make high-level inferences to convert their observations to a seven-point scale.

FfT

The FfT was developed by Charlotte Danielson in 2007 and revised in 2013. It is 'a research-based set of components of instruction grounded in a constructivist view of

[15] Observation models are usually organised by domains, dimensions and indicators. Domains reflect the high-level criteria related to the quality of teaching that the model is focused on measuring. Each domain tends to have multiple dimensions that define the observable behaviours that can actually be measured for each domain. Indicators are the specific features through which the dimensions are measured.

learning and teaching'.[16] FfT claims to be the most widely used definition of teaching in the US and is frequently used for teacher accountability purposes. Along with CLASS, it was one of the protocols used and validated in the MET study. FfT has also been independently validated by the Chicago Consortium of School Research.[17]

The FfT is made up of 22 dimensions and 76 smaller indicators clustered into four domains of teaching responsibility:

- planning and preparation
- classroom environment
- instruction
- professional responsibilities.

Each dimension defines a distinct aspect of a domain; two to five indicators describe a specific feature of a dimension. Levels of teaching performance (rubrics) have been developed to describe each dimension and provide a roadmap for improving teaching. An evaluation instrument has been designed for the FfT that allows observers to measure the level of performance on different behaviours within the rubric on a four-point scale: unsatisfactory, basic, proficient and distinguished.

Unlike most of the other models presented, FfT is not an observation tool but a set of teacher standards, informed by research, that are linked with pupils' learning. This means that observation is just one way of measuring the standards in this framework. Teacher and student questionnaires are other methods for measuring the content of the framework. Indeed, the teaching evaluation instrument designed specifically for the framework indicates that evidence should be gathered through multiple methods and not just direct classroom observations.

The framework has been used for many purposes: for instance, for mentoring, coaching, professional development and teacher evaluation processes at school and district level. A subject-focused version of the framework has also been created for mathematics and literacy, alongside the generic version. A high level of training is required to ensure that a common understanding of the high inference framework is developed across its user-base, which is critical to accuracy, teaching advancement and the impact on students' core learning. For instance, the teaching evaluation instrument designed for the framework is not a checklist focusing on easy-to-measure yet trivial aspects of practice. It requires training and judgement on the part of observers to identify and rate, for example, the quality of teacher questioning, particularly the ability to differentiate between low-quality and high-quality questioning.

[16] www.danielsongroup.org/framework

[17] R Garrett & MP Steinberg, 'Examining teacher effectiveness using classroom observation scores: evidence from the randomization of teachers to students', Educational Evaluation and Policy Analysis, 37:2, 2015. http://journals.sagepub.com/doi/abs/10.3102/0162373714537551

ICALT

ICALT is an observation instrument developed by Wim van de Grift and colleagues for use in the national inspection system in The Netherlands.

ICALT was initially underpinned by a theoretical framework that established a relationship between the basic characteristics of teaching and the academic achievements of pupils. This was reinforced by the available research literature that identified standards and indicators of good and effective teaching. The researchers tested the model across a number of school inspectorates including in England, Flanders, Lower Saxony, The Netherlands, North-Rhine Westphalia, Scotland, Ireland and the Czech Republic. These results highlighted a great deal of agreement between the inspectorates as to the basic elements of what constitutes good and responsible teaching.[18] ICALT has been shown to be positively and significantly related with pupils' involvement, attitudes and behaviour, and attainment.[19]

Unlike most of the other models presented at the seminar, ICALT features a number of low-inference indicators in its design. That is, some of the observable factors measured by the instrument are essentially factual counts of an activity completed by the teacher or student, for example noting how many times students gave the correct answer to a question in a lesson. Low-inference indicators are typically rated with minimal subjective judgement or interpretation by the observer. ICALT includes 32 high-inference indicators (for optimising inter-rater reliability) and 120 low-inference indicators that specify observable teaching behaviours. For instance, the teacher offering weak students additional learning and instruction time is an observable behaviour that ICALT categorises into both high and low-inference indicators. These behaviours are grouped into six domains:

- safe learning climate: the relationship between teacher and class
- classroom management: the overall order in the classroom
- clear instruction: the quality explanations of lesson topics and overall lesson structure and the connections among lesson parts
- activating teaching methods: various teaching strategies that motivate students to think about the topic
- learning strategies: teachers' efforts to teach students how to learn
- differentiation: whether teachers are sensitive to and flexible in attempting to meet individual students' learning problems and needs.

Observers rate the items on a four-point scale (1 = mostly weak; 2 = more often weak than strong; 3 = more often strong than weak; 4 = mostly strong). This

[18] W van de Grift, 'Measuring teaching quality in several European countries', School Effectiveness and School Improvement, 25:3, 2014. www.tandfonline.com/doi/abs/10.1080/09243453.2013.794845
[19] W van de Grift, 'Quality of teaching in four European countries: A review of the literature and an application of an assessment instrument', Educational Research, 49:2, 2007. www.tandfonline.com/doi/abs/10.1080/00131880701369651

indicates that ICALT remains a relatively high inference model, despite the inclusion of low-inference indicators, as there is a clear element of observer judgement required in determing the rating given for some items. Items referring to the six domains together describe the latent variable of teaching skill.[20]

ICALT can be described as an event-sampling observation instrument. An important consideration of this is that the standards and indicators prescribed must be observable in almost each lesson. The instrument is also developmentally focused. The differential 'item difficulty' within the rubric represents specific aspects for teacher development purposes.

ISTOF

ISTOF is a generic teacher-observation framework developed as an instrument to work across borders in international school effectiveness studies. ISTOF was, from its development in 2004, structured as an international, collaborative effort. It was intended to enable formative feedback on teaching as well as collecting research data. The development team consisted of members from 20 countries that volunteered to take part. It was organised into a number of committees led by a central committee under the leadership of Charles Teddlie, Bert Creemers, Leonidas Kyriakides, David Reynolds and Daniel Muijs.[21]

The instrument was not based on a particular teaching approach or philosophy. Instead, it was iteratively developed by the international country teams. It contains items that draw on a variety of perspectives, from direct instruction to metacognitive approaches and active learning. The instrument has been used in a number of countries that have generally shown it to have good reliability and validity, albeit that the factor structure is subject to variation across studies.[22] The iterative development process resulted in 11 domains being identified as part of effective teaching across participating countries:

- assessment and evaluation: the extent to which effective feedback is provided and assessment is aligned to goals and objectives
- clarity of instruction: the extent to which lessons are well structured and purposeful and teacher communication is of high quality

[20] Latent variables are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed.
[21] C Teddlie, B Creemers, L Kyriakides, D Muijs & F Yu (2006), 'The international system for Teacher Observation and Feedback: Evolution of an international study of teacher effectiveness constructs', Educational Research and Evaluation, 12:6, 2006. www.tandfonline.com/doi/abs/10.1080/13803610600874067
[22] D Muijs, D Reynolds, P Sammons, L Kyriakides, B Creemers & C Teddlie, 'Assessing individual lessons using a generic teacher observation instrument: how useful is the International System for Teacher Observation and Feedback (ISTOF)?', ZDM Mathematics Education, Online early article, 2018. https://link.springer.com/article/10.1007/s11858-018-0921-9

- classroom climate: the extent to which the teacher communicates high expectations and communicates with and involves and values all students
- classroom management: the extent to which the teacher maximises learning time and deals with disruptions
- differentiation and inclusion: the extent to which all students are involved in the lesson and the teacher takes student differences into account
- instructional skills: the extent to which the teacher can engage students, shows good questioning skills and uses varied methods and strategies
- planning of single lessons: the extent to which the teacher has effectively planned the observed lesson
- long-term planning: the extent to which the teacher can plan a sequence of lessons
- teacher knowledge: subject, pedagogy and pedagogical content knowledge
- teacher professionalism and reflectivity: the extent to which the teacher can reflect on her/his own practice and contribute to the schools' learning community and the teaching profession
- promoting active learning and developing metacognitive skills: the extent to which the teacher develops pupils' metacognitive skills, provides opportunities for active learning and fosters critical thinking skills.

Note that four of the overarching domains (planning of single lessons, long-term planning, teacher knowledge, and teacher professionalism and reflectivity) are not observable so the actual observation instrument contains seven domains.

Each domain has between two and four dimensions. Each dimension consists of two indicators. An example can be given for the domain 'differentiation and inclusion'. This consists of the dimensions 'The teacher creates an environment in which all students are involved' and 'The teacher takes full account of student differences'. The latter indicator consists of two indicators: 'The teacher makes a distinction in the scope of the assignments for different groups of students' and 'The teacher gives additional opportunities for practice to students who need them'. Each indicator in the instrument is rated on a five-point Likert scale (labelled 5 = strongly agree, 4 = agree somewhat, 3= neutral, 2 = disagree somewhat, 1 = strongly disagree). A 'not applicable' category is also available. This was included because not all indicators can necessarily be observed in all lessons. For example, the item 'The teacher makes a distinction in the scope of the assignments for different groups of students' is dependent on the teacher giving assignments in the first place, which may not be the case in all lessons.

Observations can be done either in person or through video. The high-inference nature of the items requires observers to be appropriately trained before using the instrument.

MQI

MQI is a content-specific observational rubric.[23] It was developed by Heather Hill and colleagues at the University of Michigan and Harvard University to measure several dimensions around the quality of mathematics instruction reliably. The model was developed and piloted between 2003 and 2012 through a theory of instruction, existing educational psychology literature and video analysis of teaching to identify the key ingredients of mathematical instruction. Studies have shown that MQI is a valid model: teacher ratings are significantly related to student outcomes, teacher value-added scores and teachers' mathematical knowledge for teaching.[24] It was also one of the protocols used in the MET study.

Unlike the other models presented at the seminar, MQI does not measure domains like classroom environment. Instead, the model is based on the perspective that mathematical work that occurs in the classroom is distinct from classroom climate, pedagogical style or using generic instructional strategies. The five measurable domains of MQI include:

- common core-aligned student practices (captures the ways in which students engage with mathematical content)
- working with students and mathematics (identifies whether teachers can hear and understand what students are saying mathematically and respond appropriately)
- richness of mathematics (measures the attention to the meaning of mathematical facts and the procedures and engagement with mathematical practices and language)
- errors and imprecision (identifies mathematical errors and distortion of content by the teacher)
- classroom work is connected to mathematics (captures whether classroom work has a mathematical point or whether instructional time is spent on activities that do not develop mathematical ideas).

MQI therefore captures the nature and quality of the mathematical content available to students as expressed through teacher–student, teacher–content and student–content interactions. For instance, the presence of mathematical explanations and practices is scored separately from student participation in mathematical explanations and practices. This ensures that MQI provides a balanced view of the numerous elements that exist in a mathematics lesson.

In practice, MQI is applied through video capture of mathematics lessons. These are generally broken down into seven-and-a-half-minute segments, scored by two

[23] https://cepr.harvard.edu/mqi
[24] CY Charalambous & E Litke, 'Studying instructional quality by using a content-specific lens: the case of the Mathematical Quality of Instruction framework', ZDM Mathematics Education, Online early article, 2018 https://link.springer.com/article/10.1007/s11858-018-0913-9

observers per segment. Observers provide overall teacher scores for each lesson, alongside scores across each domain. Observers are also trained and certified through an online system and supervised weekly to ensure greater consistency. Owing to the specificity of the model, observers need to have strong mathematical content knowledge.

Generic Dimensions of Teaching Quality

The Generic Dimensions of Teaching Quality, also known as the German Framework of Three Basic Dimensions, is an observation model originally developed by Eckhard Klieme and colleagues for the German enhancement to the TIMSS-Video study 1995. It has subsequently been used in a large number of research studies in Germany, Switzerland and Austria and has been adapted at an international scale in OECD-TALIS and PISA, and for school inspection in Germany, most prominently in Hamburg.

The three basic dimensions of teaching quality in the framework are classroom management, student support and cognitive activation. These dimensions have been developed from general theories of schooling and teaching as well as established theories and research traditions from educational psychology. This has subsequently been supported by the use of the framework in a number of research studies, confirming the reliability and the validity of the three-factor structure, although the predictive validity on cognitive and motivational student learning outcomes remains weak.[25]

Unlike other models aiming at measuring teaching quality, the Generic Dimensions of Teacher Quality is not a standardised instrument. Instead, it has become associated with many different sub-dimensions (such as disruptions and discipline problems, teacher–student relationship, and challenging tasks and questions). Different measurements are used to focus on different perspectives (for instance observer, student and teacher perspectives), although it is more frequently applied using high-inference observation protocols. There is also no standardised rubric or training manual. In addition, while it was originally developed from research on mathematics instruction, the conceptual framework has been used with respect to several subjects, school types and educational systems. The framework, therefore, provides clear categorisation of a complex phenomenon (teacher quality), which has led to the development of a number of flexible observation models generalised across numerous contexts and subject areas. One consequence of this is that, because the model is not content-dependent, it may lack comprehensiveness.

An advantage of the fact that it is a framework rather than an instrument is that users can construct different and changeable indicators for use in observation. For

[25] AK Praetorius, E Klieme, B Herbert & P Pinger, 'Generic dimensions of teaching quality: the German framework of Three Basic Dimensions', ZDM Mathematics Education, Online early article, 2018 https://link.springer.com/article/10.1007/s11858-018-0918-4

instance, the Hamburg inspectorate can vary items across years, which limits the extent to which schools can focus practice on specific measured indicators.

What the six lesson observation models tell us

The six models presented at the seminar give a picture of lesson observation that contrasts in some key aspects with the current Ofsted method. For instance, the models appear to differ in terms of focus on teaching styles. A few include indicators that clearly draw on a range of teaching perspectives, while others could be viewed as reflecting a particular teaching style. Similarly, all the models routinely collect teacher data, but in the case of ISTOF this has to date been used only for research and not for teacher development or evaluation purposes.

Aspects of design around the models also differ. For instance, MQI has a specific subject focus, whereas the other models mostly feature a generic approach to the observation of teaching. Similarly, FfT and the Generic Dimensions of Teaching Quality are frameworks for measuring teaching quality, rather than a standardised lesson observation model.

The experts attending the seminar were keen to point out that their observation models are designed around an intended purpose and that it is the purpose that drives the focus of observation. A lack of focus on the specific goals of the model could render it invalid, particularly if observers are expected to capture data just for the sake of it or if the model is used for multiple unrelated purposes. So, there was a clear link between the purpose and the items measured. The experts therefore advised that Ofsted cannot simply take one of these models 'off-the-shelf' and apply it within an inspection framework. Instead, Ofsted's approach to lesson observation must be specific and related to the purpose of inspection.

The fact that the six models all focus observation at the level of the individual teacher means that they would not be suitable for Ofsted's context. The core purpose of the observation models from the US, for instance, is linked with teacher-level accountability and the available data on student progress. They also often include guidelines and support materials that have been developed for teacher professional development post-observation. Ofsted, on the other hand, does not grade teachers. We use observation as part of an overall judgement of the quality of teaching and learning across the school. Therefore, lesson observation needs to be done with this whole-school context in mind.

The experts at the seminar supported the whole-school approach. They were clear and in agreement with Ofsted about the difficulty of reliably making inferences on the quality of a single observed lesson. In their models, multiple observations and data sources are needed to come to an even partially reliable judgement on individual teachers. For instance, most of the models are similar in terms of their relationship with pupil attainment measures, which are typically modest with

correlations of between .3 and .4.[26] This suggests that it is important to look at observation as one of a set of measures, rather than the sole measure of quality of education or teaching.

One shared aspect of the six models is that they are very systematic in their design. Each includes a clear conceptualisation and description of the key criteria (the domains) that observers use to measure the quality of the lesson and teacher performance. This was broken down into useful dimensions and indicators that ensured that a level of consistency could be expected between observers. This structure, at the very least, ensures that observers look across the same set of indicators governed by the instrument during an observation. Most of the models also featured a set of instructions (a rubric) for collecting and rating this data in a routine way, adding a further layer of consistency to their design.

It is worth noting that, apart from ICALT to an extent, the models are all high inference despite the use of more routine indicators. This is due to the fact that the indicators being used all require a level of subjective interpretation on the part of the observer for a reliable score to be provided. There was clear agreement among seminar participants that low-inference measures on their own lack validity in respect to teaching quality. Indeed, aspects of teaching that are significantly related to pupil outcomes cannot easily be captured by low-inference methods such as counts or presence/absence type measures. For these models, the expectation is that the behaviour observed also needs to be recorded qualitatively, alongside quantitative scores for each indicator.

With this in mind, the experts agreed that observers generally require a high standard of training in the use of high-inference observation instruments to obtain a sufficient degree of reliability. However, the training for a few of the models is not just a one-off in-depth session but something that is carried out on a more regular basis. In the case of MQI, for instance, weekly calibration sessions are used to ensure that observers are sufficiently skilled to carry out lesson observations. Despite this, none of the models were considered 100% reliable, with an interrater reliability of .7 – often considered by the experts to be a good result for the reliability of their models.[27]

The development of each model has commonly involved the design of an underpinning theory, supported through existing educational research literature, to

[26] In social science terms, it is rare for a correlation between two meaningful variables to reach a coefficent of .5. Observation measures are typically more highly correlated to attainment than most other measurable school/teacher level variables, and not much lower than key individual pupil measures. For instance, setting/streaming typically correlates with attainment in the range of -.1 and .1. This means that lesson observation tells us something about the quality of teaching that is important to capture as part of understanding school effectiveness.

[27] Inter-rater reliability is the degree of agreement among raters. The statistic for inter-rater reliability is often considered to be better than a straightforward percentage difference, as this takes into account chance variation. A statistic between .6 and .8 is generally considered to have a high level of reliability.

devise the model's purpose. The indicators determined from this process have then been tested more widely through live or video-recorded lesson observation. The amount of data collected during the piloting phases has provided a degree of assurance that the indicators being measured are relevant and gives each of the six models strong claims towards validity.

The overlap of similar domains across the six observation models also suggests there are aspects of measuring teacher quality that are prevalent in all the models. This concurrent validity would suggest that some indicators, therefore, may be relevant for the inspection context. For instance, aspects of classroom management, clarity of instruction and student behaviour and attitudes were routinely included in the models presented at the seminar.

An important distinction to make is between the four generic models (FfT, CLASS, ISTOF and ICALT) and the subject-specific models of MQI and the Generic Dimensions of Teacher Quality. The latter was originally developed for mathematics, but has since been used both for generic purposes and for some other subjects (notably home language). There was general agreement that teaching quality has both generic and subject-specific components. In practice, however, subject-specific instruments currently only exist for a limited number of core subjects, mostly mathematics and reading/language.

The experts were also clear that lesson observation criteria should focus on components that can be directly observed in the classroom. This means pupils' 'learning' is absent from the criteria of the six models. It was generally agreed among the international experts that learning is invisible and happens over a long period of time. It is not something that can be directly observed. At the very least, this is something they felt could not be measured in a valid way through observation alone, hence its exclusion in these models and why it remains just one of many data points for assessing the quality of teaching in Ofsted inspections. Instead, it was suggested that other methods alongside observation or high quality, structured assessment would be necessary to effectively capture the learning made by pupils.

The experts discussed two interesting but less critical points during the seminar. Some models tended to use video recordings of lessons as the raw material for observation. This raised some debate on the relative merits and weaknesses of live observation against recordings. On the one hand, they argued that video could improve accuracy because a recording could be watched multiple times. On the other, they argued that video could lose vital contextual information about the classroom when filmed from a single or even from multiple perspective(s). For instance, unless filmed from multiple perspectives, a recording may only show the back of students' heads, making it difficult to make inferences on student engagement. There is also a cost implication with using video and participant agreement in the English context may also prove problematic. Finally, there was no agreement among the experts on the ideal observation length or number of observations required to ensure reliability. This was largely due to this being linked to and dependent on the context and intended purpose of the observation tool.

Next steps

The in-depth discussion on these models means that we need to reflect further before deciding how we can use this information to develop the validity and reliability of Ofsted's current lesson observation model. There are a number of alternatives that seem to be available, but we will need to consider how these best fit into a lesson observation model that is fit for purpose in supporting inspector judgements at the school level. During the summer and autumn terms 2018, we will be carrying out further research to test a number of these alternative model designs. The outcomes of this will feed into the 2019 education inspection framework.

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. 180022